

PROJECT PERIODIC REPORT

Grant Agreement number: 288956

Project acronym: NADINE

Project title: New tools and Algorithms for Directed Network analysis

Funding Scheme: Small or medium-scale focused research project (STREP)

Periodic report: 1st 2nd

Period covered: from 1.11.2013 to 30.04.2015

Name, title and organisation of the scientific representative of the project's coordinator¹:

Dr. Dima Shepelyansky

Directeur de recherche au CNRS

Lab de Phys. Theorique, Universite Paul Sabatier, 31062 Toulouse, France

Tel: +331 5 61556068, Fax: +33 5 61556065, Secr.: +33 5 61557572

E-mail: dima@irsamc.ups-tlse.fr; URL: www.quantware.ups-tlse.fr/dima

Project website address: www.quantware.ups-tlse.fr/FETNADINE/

¹

Usually the contact person of the coordinator as specified in Art. 8.1. of the grant agreement

NADINE DELIVERABLE D4.2.

It is based on milestones M7(WP4.1-WP5.2), M8(WP4.4), M13(WP4.3) with deliverable publications:

[2] P1.2 L.Ermann and D.L. Shepelyansky "**Ecological analysis of world trade**", Phys. Lett. A v.377, p.250 (2013) (arXiv:1201.3584[q-fin.GN], 2012) [M8-WP4.4 – reported in period 1]

[33] P1.13 V.Kandiah and D.L.Shevelyansky, "**Google matrix analysis of C.elegans neural network**", Phys. Lett. A v.378, p.1932 (2014) (arXiv:1311.2013[physics.soc-ph]) [M13-WP4.3]

[34] P1.14 K.M.Frahm and D.L.Shevelyansky, "**Poisson statistics of PageRank probabilities of Twitter and Wikipedia networks**", Eur. Phys. J. B v.87, p. 93 (2014) (arXiv:1402.5839[physics.soc-ph]) [M13-WP4.3]

[35] P1.15 L.Ermann, K.M.Frahm and D.L.Shevelyansky, "**Google matrix analysis of directed networks**", submitted to Rev. Mod. Phys. (2014) (arXiv:1409.0428[physics.soc-ph]) [M8-WP4.4, M13-WP4.3]

[36] P1.16 Young-Ho Eom and Hong-Hyun Jo, "**Generalized friendship paradox in complex networks: the case of scientific collaboration**", Scientific Reports v.4, p.4603 (2014) [M13-WP4.3]

[37] P1.17 Hong-Hyun Jo and Young-Ho Eom, "**Generalized friendship paradox in networks with tunable degree-attribute correlation**", Phys. Rev. E v.90, p.022809 (2014) [M13-WP4.3]

[38] P1.18 V.Kandiah, B.Georgeot and O.Giraud, "**More ordering and communities in complex networks describing the game of go**", Eur. Phys. J. B v.87, p.246 (2014) [M13-WP4.3]

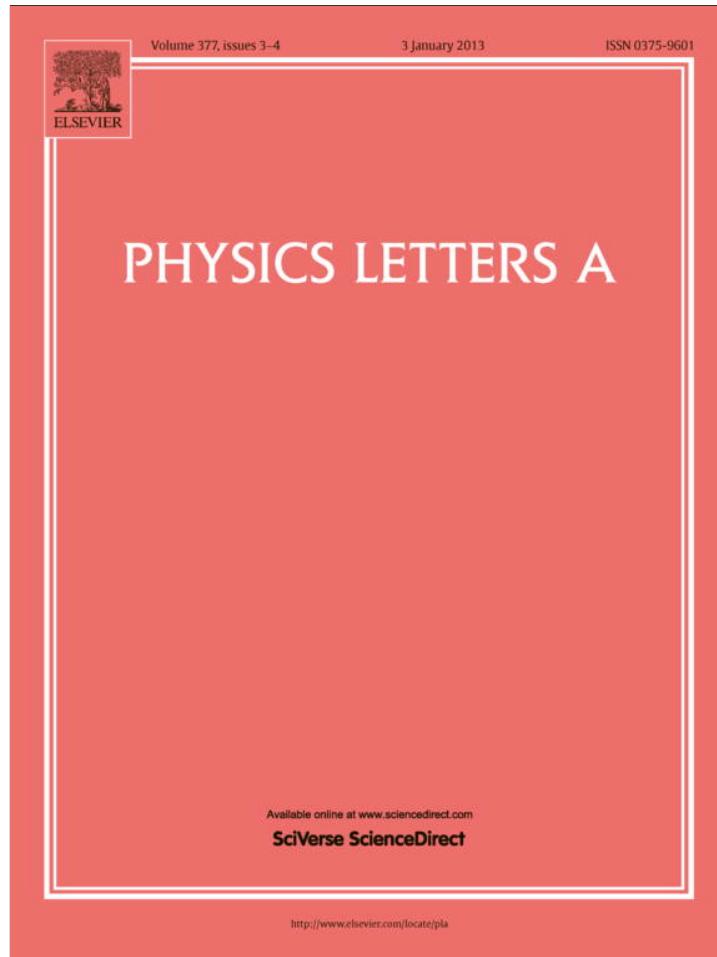
[39] P1.19 L.Ermann and D.L.Shevelyansky, "**Google matrix analysis of the multiproduct world trade network**", Eur.Phys. J. B v.88, p.84 (2015) (arXiv:1502.00584[cond-mat.dis-nn]) [M8-WP4.4]

[42] P1.22 V.Kandiah, H.Escaith and D.L.Shevelyansky, "**Google matrix of the world network of economic activities**", submitted to Eur. Phys. J. B April (2015) (arXiv:1504.XXXX[q-fin.ST]) [M8-WP4.4]

[43] P1.23 D.L.Shevelyansky and other Wikipedia authors, "**Top 100 historical figures of Wikipedia**", Wikipedia article (2014) [M13-WP4.3]

- [44] P2.7 P. van der Hoorn and N. Litvak, "**Convergence of rank based degree-degree correlations in random directed networks**", to appear in Moscow Journal of Combinatorics (2015) (arXiv:1407.7662[math.PR], 2014) [M13-WP4.3]
- [45] P2.8 P. van der Hoorn and N. Litvak, "**Phase transitions for scaling of structural correlations in directed networks**", (arXiv:1504.01535[physics.soc-ph], 2015 [M13-WP4.3]
- [46] P2.9 M. Ten Thij, T. Ouboter, D. Worm, N. Litvak, J.L. van den Berg and S. Bhulai, **Modelling of trends in Twitter using retweet graph dynamics**, Proceedings 11th International Workshop Algorithms and Models for the Web Graph, WAW 2014, 17-18 Dec 2014, Beijing, China. pp. 132-147; Lecture Notes in Computer Science 2014 (8882), Springer (2014), (arXiv:1502.00166[cs.SI], 2015) [M13-WP4.3]
- [62] P4.9 Robert Meusel, Sebastiano Vigna, Oliver Lehmburg, and Christian Bizer, "**Graph structure in the web - Revisited, or a trick of the heavy-tail**", WWW'14 Companion, pp.427-432, International World Wide Web Conferences Steering Committee, 2014; a revised version is to appear in the Journal of Web Science (2015) [M13-WP4.3]
- [63] P4.10 Djamal Belazzougui, Paolo Boldi, Giuseppe Ottaviano, Rossano Venturini, and Sebastiano Vigna, "**Cache-oblivious peeling of random hypergraphs**", 2014 Data Compression Conference (DCC 2014), IEEE pp.352-361. (2014) [M7-WP4.1]
- [64] P4.11 Paolo Boldi, Irene Crimaldi, and Corrado Monti, "**A network model characterized by a latent attribute structure with competition**", submitted CoRR (2014), (arXiv:1407.7729[cs.SI], 2014 [M7-WP4.1]
- [65] P4.12 Roi Blanco, Paolo Boldi, and Andrea Marino, "**Entity-linking via graph-distance minimization**", Proceedings 3rd Workshop on GRAPH Inspection and Traversal Engineering, GRAPHITE 2014, Grenoble, France, 5th April 2014., pp.30-43 (2014) [M7-WP4.1]
- [69] P4.16 Young Ho Eom, Pablo Aragon, David Laniado, Andreas Kaltenbrunner, Sebastiano Vigna, and Dima L. Shepelyansky, "**Interactions of cultures and top people of Wikipedia from ranking of 24 language editions**", PLoS ONE v.10(3), p.e0114825 (2015) (arXiv:1405.7183[cs.SI], 2014) [M13-WP4.3-WP5.2]
- [70] P4.17 Sebastiano Vigna, "**A weighted correlation index for rankings with ties**", Proceedings of the 24th international conference on World Wide Web, ACM (2015) (arXiv:1404.3325[cs.SI], 2014) [M13-WP4.3-WP5.2]
- [73] P4.20 Michele Trevisio, Luca Maria Aiello, Paolo Boldi and Roi Blanco, "**Local Ranking Problem on the BrowseGraph**", accepted for publication in SIGIR (2015) [M13-WP4.3-WP5.2]

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Ecological analysis of world trade



L. Ermann ^{b,a}, D.L. Shepelyansky ^{a,*}

^a Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, F-31062 Toulouse, France

^b Departamento de Física Teórica, GlyA, Comisión Nacional de Energía Atómica, Buenos Aires, Argentina

ARTICLE INFO

Article history:

Received 22 March 2012

Received in revised form 28 August 2012

Accepted 31 October 2012

Available online 21 November 2012

Communicated by A.R. Bishop

Keywords:

Complex networks

Nestedness

World trade

ABSTRACT

Ecological systems have a high complexity combined with stability and rich biodiversity. The analysis of their properties uses a concept of mutualistic networks and provides a detailed understanding of their features being linked to a high nestedness of these networks. Using the United Nations COMTRADE database we show that a similar ecological analysis gives a valuable description of the world trade: countries and trade products are analogous to plants and pollinators, and the whole trade network is characterized by a high nestedness typical for ecological networks. Our approach provides new mutualistic features of the world trade.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Ecological systems are characterized by high complexity and biodiversity [1] linked to nonlinear dynamics and chaos emerging in the process of their evolution [2,3]. The interactions between species form a complex network whose properties can be analyzed by the modern methods of scale-free networks [4–7]. An important feature of ecological networks is that they are highly structured, being very different from randomly interacting species [7,8]. Recently it has been shown that the mutualistic networks between plants and their pollinators [8–12] are characterized by high nestedness [13–16] which minimizes competition and increases biodiversity. It is argued [14] that such type of networks appear in various social contexts such as garment industry [15] and banking [17,18]. Here we apply a nestedness analysis to the world trade network using the United Nations COMTRADE database [19] for the years 1962–2009. Our analysis shows that countries and trade products have relations similar to those of plants and pollinators and that the world trade network is characterized by a high nestedness typical of ecosystems [14]. This provides new mutualistic characteristics for the world trade.

2. Results

The mutualistic World Trade Network (WTN) is constructed on the basis of the UN COMTRADE database [19] from the matrix

of trade transactions $M_{c',c}^p$ expressed in USD for a given product (commodity) p from country c to country c' in a given year (from 1962 to 2009). For product classification we use 3-digit Standard International Trade Classification (SITC) Rev. 1 with the number of products $N_p = 182$. All these products are described in [19] in the commodity code document SITC Rev. 1. The number of countries varies between $N_c = 164$ in 1962 and $N_c = 227$ in 2009. The import and export trade matrices are defined as $M_{p,c}^{(i)} = \sum_{c'=1}^{N_c} M_{c',c}^p$ and $M_{p,c}^{(e)} = \sum_{c'=1}^{N_c} M_{c',c}^p$ respectively. We use the dimensionless matrix elements $m^{(i)} = M^{(i)}/M_{max}$ and $m^{(e)} = M^{(e)}/M_{max}$ where for a given year $M_{max} = \max\{\max[M_{p,c}^{(i)}], \max[M_{p,c}^{(e)}]\}$. The distribution of matrix elements $m^{(i)}, m^{(e)}$ in the plane of indexes p and c , ordered by the total amount of import/export in a decreasing order, is shown in Fig. 1 for years 1968 and 2008 (years 1978, 1988, 1998 are shown in Fig. S-1 of Supporting Information (SI)). These figures show that globally the distributions of $m^{(i)}, m^{(e)}$ remain stable in time especially in a view of 100 times growth of the total trade volume during the period 1962–2009. The fluctuations of $m^{(e)}$ are visibly larger compared to $m^{(i)}$ case since certain products, e.g. petroleum, are exported by only a few countries while it is imported by almost all countries.

To use the methods of ecological analysis we construct the mutualistic network matrix for import $Q^{(i)}$ and export $Q^{(e)}$ whose matrix elements take binary value 1 or 0 if corresponding elements $m^{(i)}$ and $m^{(e)}$ are respectively larger or smaller than a certain trade threshold value μ . The fraction φ of nonzero matrix elements varies smoothly in the range $10^{-6} \leq \mu \leq 10^{-2}$ (see Fig. S-2 of SI) and the further analysis is not really sensitive to the actual μ value inside this broad range. Indeed, the variation of μ in

* Corresponding author.

E-mail address: ermann@tandar.cnea.gov.ar (L. Ermann).

URLs: <http://www.tandar.cnea.gov.ar/~ermann> (L. Ermann), <http://www. quantware.ups-tlse.fr/dima> (D.L. Shepelyansky).

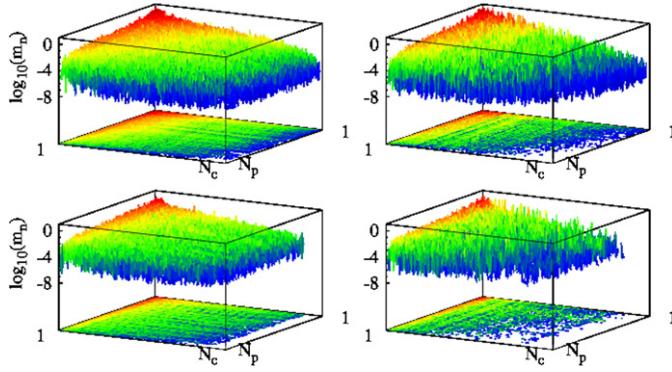


Fig. 1. Normalized import/export WTN matrix elements $m^{(i)}$ and $m^{(e)}$ shown on left/right panels for years 1968 (bottom) and 2008 (top). Each panel represents the dimensionless trade matrix elements $m^{(i)} = M^{(i)}/M_{\max}$ and $m^{(e)} = M^{(e)}/M_{\max}$ on a three-dimensional (3D) plot as a function of indexes of countries and products. Here products/countries ($p = 1, \dots, N_p$ and $c = 1, \dots, N_c$) are ordered in a decreasing order of product/country total import or export in a given year. The color is proportional to the amplitude of the matrix element changing from red (for amplitude maximum) to blue (for zero amplitude). Each panel shows the 3D distribution and its projection on 2D plane of countries–products in which the amplitude of matrix elements is shown by the same color as in 3D. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this Letter.)

the range $10^{-5} \leq \mu \leq 10^{-3}$ by two orders of magnitude produces a rather restricted variation of φ only by a factor two.

It is important to note that in contrast to ecological systems [14] the world trade is described by a directed network and hence we characterize the system by two mutualistic matrices $Q^{(i)}$ and $Q^{(e)}$ corresponding to import and export. Using the standard nestedness BINMATNEST algorithm [20] we determine the nestedness parameter η of the WTN and the related nestedness temperature $T = 100(1 - \eta)$. The algorithm reorders lines and columns of a mutualistic matrix concentrating nonzero elements as much as possible in the top-left corner and thus providing information about the role of immigration and extinction in an ecological system. A high level of nestedness and ordering can be reached only for systems with low T . It is argued that the nested architecture of real mutualistic networks increases their biodiversity.

The nestedness matrices generated by the BINMATNEST algorithm [20] are shown in Fig. 2 for ecology networks ARR1 ($N_{pl} = 84$, $N_{anim} = 101$, $\varphi = 0.043$, $T = 2.4$) and WES ($N_{pl} = 207$, $N_{anim} = 110$, $\varphi = 0.049$, $T = 3.2$) from [12,21]. Using the same algorithm we generate the nestedness matrices of WTN using the mutualistic matrices for import $Q^{(i)}$ and export $Q^{(e)}$ for the WTN in years 1968 and 2008 using a fixed typical threshold $\mu = 10^{-3}$ (see Fig. 2; the distributions for other μ values have a similar form and are shown in Fig. S-3 of SI). As for ecological systems, for the WTN data we also obtain rather small nestedness temperature ($T \approx 6/8$ for import/export in 1968 and $T \approx 4/8$ in 2008 respectively). These values are by a factor 9/4 of times smaller than the corresponding T values for import/export from random generated networks with the corresponding values of φ .

The detailed data for T in all years are shown in Fig. 3 and the comparison with the data for random networks is given in Figs. S-4–S-6 in SI. The data of Fig. 3 show that the value of T changes by about 30–40% with variation of μ by a factor 1000. We think that this is relatively small variation of T compared to enormous variation of μ that confirms the stability and relevance of ecological analysis and nestedness ordering. The nestedness temperature T remains rather stable in time: in average there is 40% drop of T from 1962 to 2000 and 20% growth from 2000 to 2009. We attribute the growth in last decade to the globalization of trade. Even if the nestedness temperature T may be sensitive to

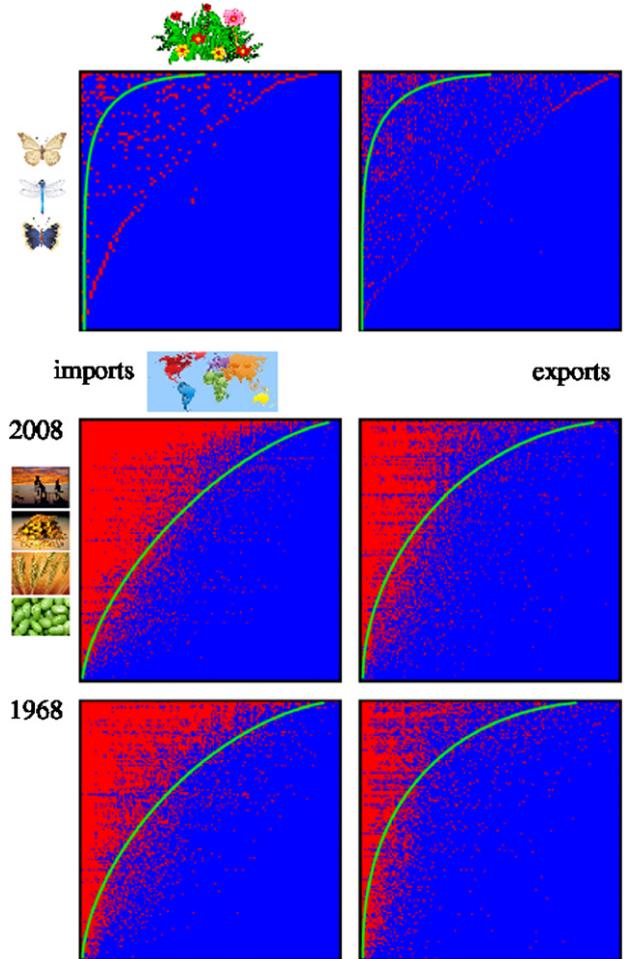


Fig. 2. Nestedness matrices for the plant–animal mutualistic networks on top panels, and for the WTN of countries–products on middle and bottom panels. Top-left and top-right panels represent data of ARR1 and WES networks from [12,21]. The WTN matrices are computed with the threshold $\mu = 10^{-3}$ and corresponding $\varphi \approx 0.2$ for years 1968 (bottom) and 2008 (middle) for import (left panels) and export (right panels). Red and blue represent unit and zero elements respectively; only lines and columns with nonzero elements are shown. The order of plants–animals, countries–products is given by the nestedness algorithm [20], the perfect nestedness is shown by green curves for the corresponding values of φ . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this Letter.)

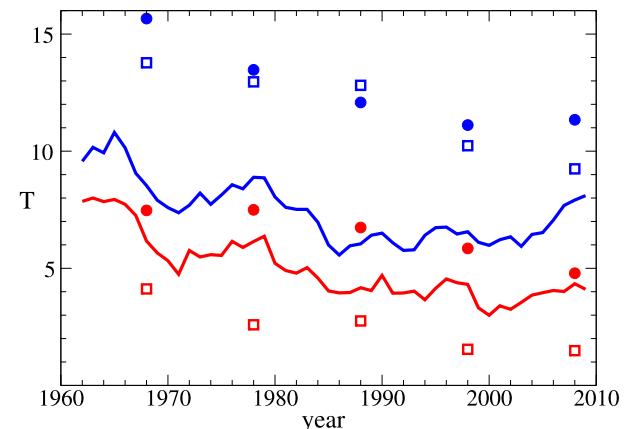


Fig. 3. Nestedness temperature T as a function of years for the WTN for $\mu = 10^{-3}$ (curves), 10^{-4} (circles), 10^{-6} (squares); import and export data are shown in red and blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this Letter.)

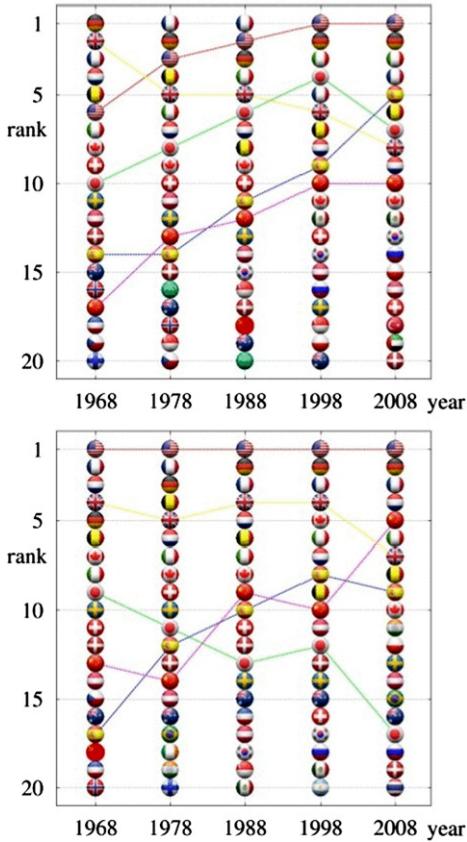


Fig. 4. Top 20 EcoloRank countries as a function of years for the WTN import/export on top/bottom panels. The ranking is given by the nestedness algorithm [20] for the trade threshold $\mu = 10^{-3}$; each country is represented by its corresponding flag. As an example, dashed lines show time evolution of the following countries: USA, UK, Japan, China, Spain. (For interpretation of the references to color in this figure, the reader is referred to the web version of this Letter.)

variation of φ the data of Figs. S-2 and S-6 show that in the main range of $10^{-5} \leq \mu \leq 10^{-3}$ the variation of φ and T remains rather small. The comparison with the randomly generated networks also shows that they have significantly larger T values compared to the values found for the WTN (see also discussion of Figs. S-4–S-6 in SI).

The small value of nestedness temperature obtained for the WTN confirms the validity of the ecological analysis of WTN structure: trade products play the role of pollinators which produce exchange between world countries, which play the role of plants. Like in ecology the WTN evolves to the state with very low nestedness temperature that satisfies the ecological concept of system stability appearing as a result of high network nestedness [14].

The nestedness algorithm [20] creates effective ecological ranking (EcoloRanking) of all UN countries. The evolution of 20 top ranks throughout the years is shown in Fig. 4 for import and export. This ranking is quite different from the more commonly applied ranking of countries by their total import/export monetary trade volume [22] (see corresponding data in Fig. 5) or recently proposed democratic ranking of WTN based on the Google matrix analysis [23]. Indeed, in 2008 China is at the top rank for total export volume but it is only at 5th position in EcoloRanking (see Figs. 4, 5 and Table 1 in SI). In a similar way Japan moves down from 4th to 17th position while the USA raises up from 3rd to 1st rank.

The same nestedness algorithm generates not only the ranking of countries but also the ranking of trade products for import and export which is presented in Fig. 6. For comparison we also

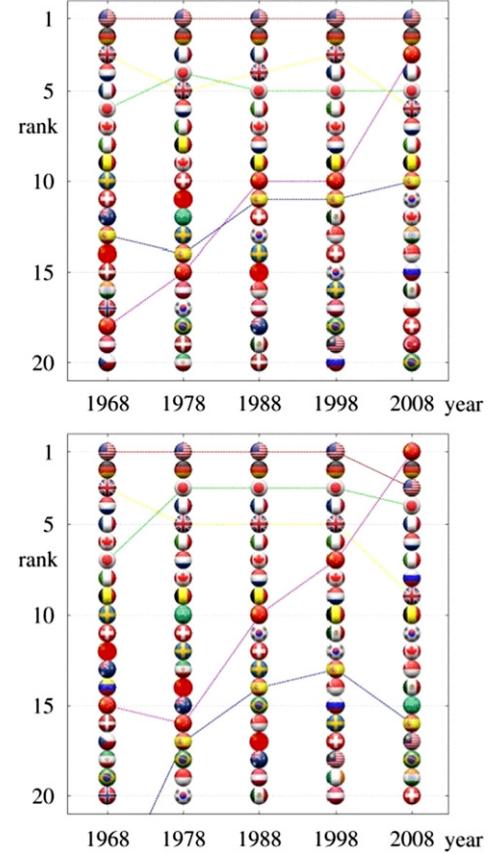


Fig. 5. Top 20 countries as a function of years ranked by the total monetary trade volume of the WTN in import/export on top/bottom panels respectively; each country is represented by its corresponding flag. Dashed lines show time evolution of the same countries as in Fig. 4.

show there the standard ranking of products by their trade volume. In Fig. 6 the color of symbol marks the 1st SITC digit described in [19] and in Table 2 in SI.

3. Discussion

The origin of such a difference between EcoloRanking and trade volume ranking of countries is related to the main idea of mutualistic ranking in ecological systems: the nestedness ordering stresses the importance of mutualistic pollinators (products for WTN) which generate links and exchange between plants (countries for WTN). In this way generic products, which participate in the trade between many countries, become of primary importance even if their trade volume is not at the top lines of import or export. In fact such mutualistic products glue the skeleton of the world trade while the nestedness concept allows to rank them in order of their importance. The time evolution of this EcoloRanking of products of WTN is shown in Fig. 6 for import/export in comparison with the product ranking by the monetary trade volume (since the trade matrix is diagonal in product index the ranking of products in the latter case is the same for import/export). The top and middle panels have dominate colors corresponding to machinery (SITC 7; blue) and mineral fuels (3; black) with a moderate contribution of chemicals (5; yellow) and manufactured articles (8; cyan) and a small fraction of goods classified by material (6; green). Even if the global structure of product ranking by trade volume has certain similarities with import EcoloRanking there are also important new elements. Indeed, in 2008 the mutualistic significance of petroleum products (SITC 332), *machindus* (machines for special industries 718) and

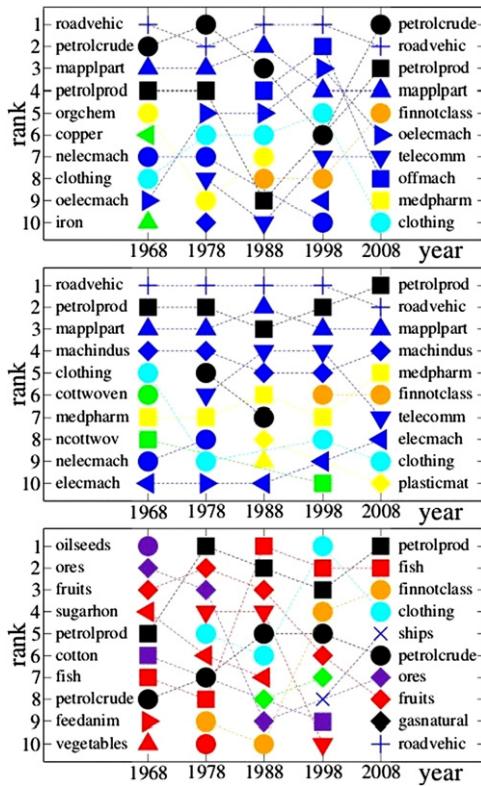


Fig. 6. Top 10 ranks of trade products as a function of years for the WTN. Top panel: ranking of products by monetary trade volume; middle/bottom panels: ranking is given by the nestedness algorithm [20] for import/export with the trade threshold $\mu = 10^{-3}$. Each product is shown by its own symbol with short name written for years 1968, 2008; symbol colors mark 1st SITC digit; SITC codes of products and their names are given in Table 2 of SI. (For interpretation of the references to color in this figure, the reader is referred to the web version of this Letter.)

medpharm (medical-pharmaceutic products 541) is much higher compared to their volume ranking, while petroleum crude (331) and office machines (714) have smaller mutualistic significance compared to their volume ranking.

The new element of EcoloRanking is that it differentiates between import and export products while for trade volume they are ranked in the same way. Indeed, the dominant colors for export (Fig. 6, bottom panel) correspond to food (SITC 0; red) with contribution of black (present in import) and crude materials (2; violet), followed by cyan (present in import) and more pronounced presence of *finnotclass* (commodities/transactions not classified 9; brown). EcoloRanking of export shows a clear decrease tendency of dominance of SITC 0 and SITC 2 with time and increase of importance of SITC 3, 7. It is interesting to note that petroleum products SITC 332 is very vulnerable in volume ranking due to significant variations of petroleum prices but in EcoloRanking this product keeps the stable top positions in all years showing its mutualistic structural importance for the world trade. EcoloRanking of export shows also importance of fish (SITC 031), clothing (SITC 841) and fruits (SITC 051) which are placed on higher positions compared to their volume ranking. At the same time *roadvehic* (SITC 732), which are at top volume ranking, have relatively low ranking in export since only a few countries dominate the production of road vehicles.

It is interesting to note that in Fig. 6 petroleum crude is at the top of trade volume ranking e.g. in 2008 (top panel) but it is absent in import EcoloRanking (middle panel) and it is only on 6th position in export EcoloRanking (bottom panel). A similar feature is visible for years 1968, 1978. On a first glance this looks surprising but in fact for mutualistic EcoloRanking it is important that

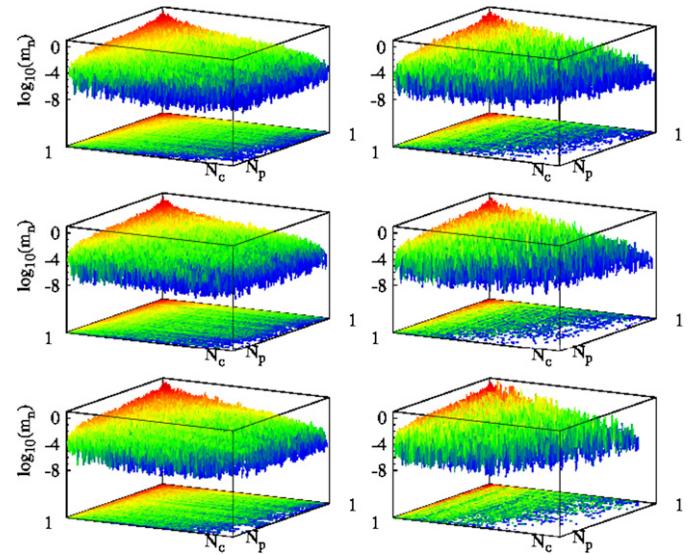


Fig. S-1. Same type of WTN matrix data as in Fig. 1 shown for years 1978, 1988, 1998 in panels from bottom to top respectively.

a given product is imported from top EcoloRank countries: this is definitely not the case for petroleum crude which practically is not produced inside top 10 import EcoloRank countries (the only exception is the USA, which however also does not export much). Due to that reason this product has low mutualistic significance.

The mutualistic concept of product importance is at the origin of significant difference of EcoloRanking of countries compared to the usual trade volume ranking (see Figs. 4, 5). Indeed, in the latter case China and Japan are at the dominant positions but their trade is concentrated in specific products which mutualistic role is relatively low. In contrast the USA, Germany and France keep top three EcoloRank positions during almost 40 years clearly demonstrating their mutualistic power and importance for the world trade.

In conclusion, our results show the universal features of eco-logic ranking of complex networks with promising future applications to trade, finance and other areas.

Acknowledgements

We thank Arlene Adriano and Matthias Reister (UN COMTRADE) for provided help and friendly access to the database [19]. This work is done in the frame of the EC FET Open project “New tools and algorithms for directed network analysis” (NADINE No. 288956).

Appendix A. Supporting information

Here we present the Supporting Information (SI) for the main part of the Letter, it includes Figs. S-1–S-6, Table 1, Table 2.

In Fig. S-1, in a complement to Fig. 1, we show the normalized WTN matrix for import $m^{(i)}$ and export $m^{(e)}$ at additional years 1978, 1988, 1998. As in Fig. 1 all products and countries are ordered in a decreasing order of product ($p = 1, \dots, N - p$) and country ($c = 1, \dots, N_c$) import (left panels) and export (right panels) in a given year. These data show that the global distribution remains stable in time: indeed, the global monetary trade volume was increased by a factor 100 from year 1962 to 2008 (see e.g. Fig. 5 in [20]) but the shape of the distribution remained essentially the same.

The dependence of the fraction φ of nonzero elements of the mutualistic matrices of import $Q^{(i)}$ and export $Q^{(e)}$ on the cutoff threshold μ is shown in Fig. S-2. In the range of $10^{-6} \leq \mu \leq 10^{-2}$ there is a smooth relatively weak variation of φ with μ .

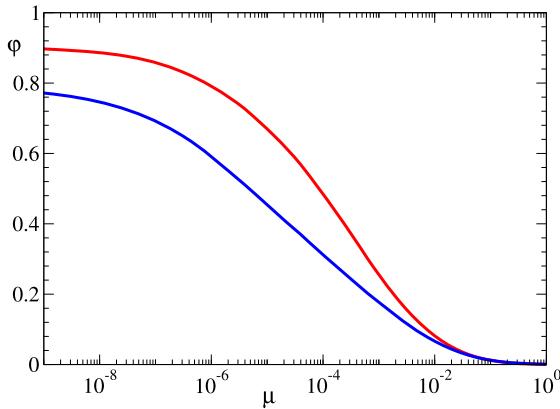


Fig. S-2. The fraction φ of nonzero matrix elements for the mutualistic network matrices of import $Q^{(i)}$ and export $Q^{(e)}$ as a function of the cutoff trade threshold μ for the normalized WTN matrices $m^{(i)}$ and $m^{(e)}$ for the year 2008; the red curve shows the case of import while the blue curve shows the case of export network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this Letter.)

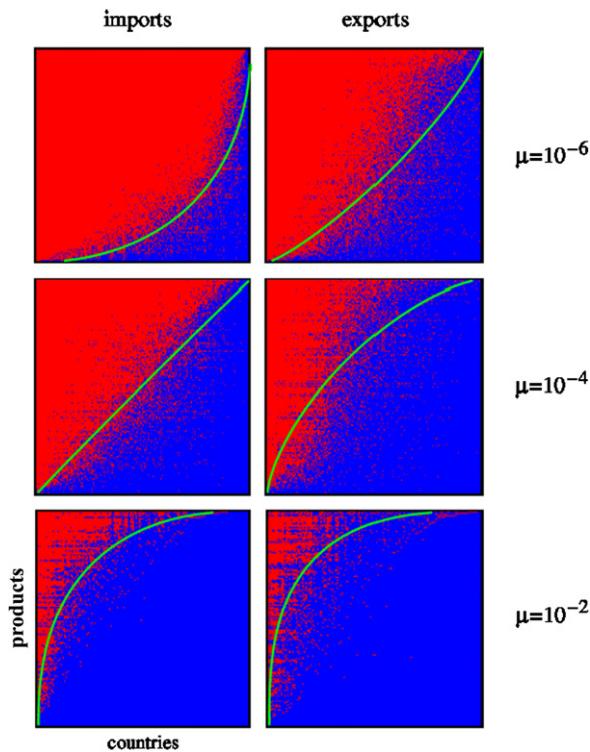


Fig. S-3. Same as in Fig. 2: nestedness matrix for the WTN data in 2008 shown for the threshold values $\mu = 10^{-6}, 10^{-4}, 10^{-2}$ (from top to bottom); the perfect nestedness is shown by green curves for the corresponding values of φ taken from Fig. S-2. (For interpretation of the reference to color in this figure legend, the reader is referred to the web version of this Letter.)

In Fig. S-3, in addition to Fig. 2, we show the nestedness matrices of WTN at various values of the cutoff threshold μ . The data at various μ values show that in all cases the nestedness algorithm [17] correctly generates a matrix with nestedness structure.

The variation of the nestedness temperature T with time is shown in Fig. 3 at several values of the trade threshold μ . These data show that in average the value of T for export is higher than for import. We attribute this to stronger fluctuations of matrix elements of $m^{(e)}$ compared to those of $m^{(i)}$ that is well visible in Figs. 1, S-1. As it is pointed in the main part, we attribute this

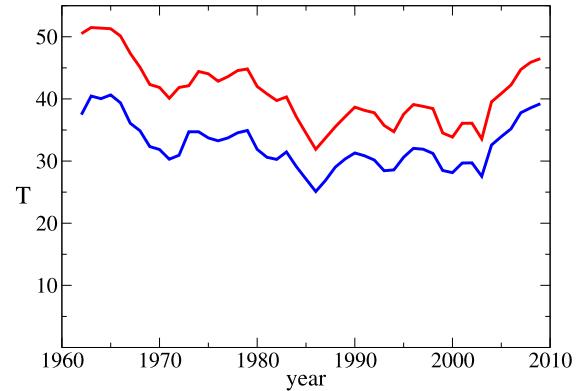


Fig. S-4. Nestedness temperature T for the model given by random generated networks; here T is computed with 500 random realizations of network for each year using N_p , N_c and φ of the corresponding WTN data in this year at $\mu = 10^{-3}$; import/export data are shown by red/blue curves respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this Letter.)

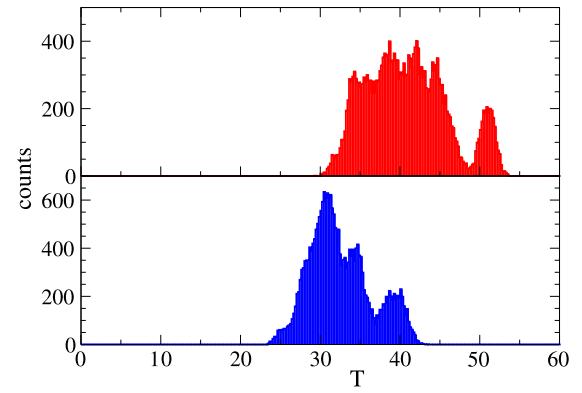


Fig. S-5. Histogram of temperatures for 500 random generated networks per year (from 1962 to 2009). Top (bottom) panel represents import (export) data; here the parameter values of N_p , N_c and φ are as for the corresponding WTN years at $\mu = 10^{-3}$.

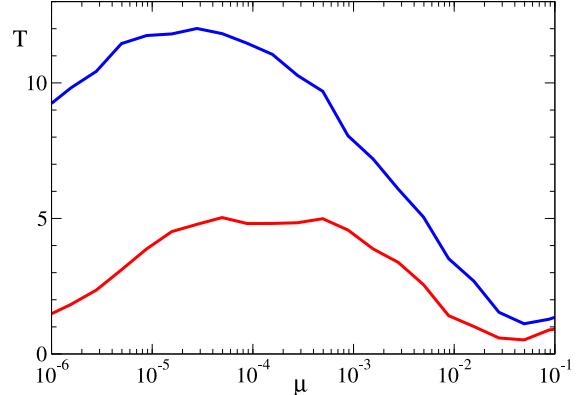


Fig. S-6. Nestedness temperature in the WTN for the year 2008 as a function of threshold μ ; import/export networks are shown by red/blue curves respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this Letter.)

to the fact that e.g. only a few countries export petroleum crude while the great majority of countries import this product.

In Fig. S-4 we show the nestedness temperature dependence on time for the case of random generated networks which have the same fraction of nonzero matrix elements φ as the WTN in the given year and $\mu = 10^{-3}$. These data, compared with those of Fig. 3, really demonstrate that the real WTN has values of T

Table 1Top 20 ranks of countries for import and export with ranking by the monetary trade volume and by the nestedness algorithm at two threshold values μ (year 2008).

Rank	Import			Export		
	Money	$\mu = 10^{-3}$	$\mu = 10^{-2}$	Money	$\mu = 10^{-3}$	$\mu = 10^{-2}$
1	USA	USA	USA	China	USA	USA
2	Germany	Germany	Germany	Germany	Germany	Germany
3	China	Italy	France	USA	France	China
4	France	France	UK	Japan	Netherlands	France
5	Japan	Spain	Italy	France	China	Italy
6	UK	Belgium	Netherlands	Netherlands	Italy	Netherlands
7	Netherlands	Japan	Belgium	Italy	UK	Belgium
8	Italy	UK	Japan	Russian Federation	Belgium	UK
9	Belgium	Netherlands	China	UK	Spain	Japan
10	Canada	China	Spain	Belgium	Canada	Spain
11	Spain	Canada	Canada	Canada	India	Canada
12	Republic of Korea	Mexico	Russian Federation	Republic of Korea	Poland	Switzerland
13	Russian Federation	Republic of Korea	Republic of Korea	Mexico	Sweden	India
14	Mexico	Russian Federation	Switzerland	Saudi Arabia	Austria	Republic of Korea
15	Singapore	Poland	Austria	Singapore	Brazil	Poland
16	India	Austria	Poland	Spain	Australia	Turkey
17	Poland	Switzerland	Sweden	Malaysia	Japan	Czech Republic
18	Switzerland	Turkey	Mexico	Brazil	Russian Federation	Austria
19	Turkey	United Arab Emirates	India	India	Denmark	Thailand
20	Brazil	Denmark	Singapore	Switzerland	Thailand	Denmark

Table 2

Product names for SITC Rev. 1 3-digit code used in Fig. 6.

Symbol	Code	Abbreviation	Name
●	001	animals	Live animals
■	031	fish	Fish, fresh and simply preserved
◆	051	fruits	Fruit, fresh, and nuts excl. oil nuts
▲	054	vegetables	Vegetables, roots and tubers, fresh or dried
▼	061	sugarhon	Sugar and honey
▶	071	coffee	Coffee
▶	081	feedanim	Feed. stuff for animals excl. unmilled cereals
●	221	oilseeds	Oil seeds, oil nuts and oil kernels
■	263	cotton	Cotton
◆	283	ores	Ores and concentrates of non-ferrous base metals
●	331	petrolcrude	Petroleum, crude and partly refined
■	332	petrolprod	Petroleum products
◆	341	gas	Gas, natural and manufactured
●	512	orgchem	Organic chemicals
■	541	medpharm	Medicinal and pharmaceutical products
◆	581	plasticmat	Plastic materials, regenerated cellulose and resins
▲	599	chemmat	Chemical materials and products, n.e.s.
●	652	cottwoven	Cotton fabrics, woven ex. narrow or spec. fabrics
■	653	ncottwov	Textile fabrics, woven ex. narrow, spec., not cotton
◆	667	pearlsprec	Pearls and precious and semi precious stones
▲	674	iron	Universals, plates and sheets of iron or steel
●	682	copper	Copper
●	711	nelecmach	Power generating machinery, other than electric
■	714	offmach	Office machines
◆	718	machindus	Machines for special industries
▲	719	mapplpart	Machinery and appliances non-electrical parts
●	722	elecmach	Electric power machinery and switchgear
■	724	telecomm	Telecommunications apparatus
◆	729	oelecmach	Other electrical machinery and apparatus
▶	732	roadvehicles	Road motor vehicles
●	735	ships	Ships and boats
●	841	clothing	Clothing except fur clothing
●	931	finnotclass	Special transactions not class. accord. to kind

by a factor 5 (export) to 10 (import) smaller comparing to the random networks. This confirms the nestedness structure of WTN being similar to the case of ecology networks discussed in [12]. It is interesting to note that for random generated networks the values of T for import are larger than for export while to the WTN we have the opposite relation. The histogram of distribution of T for random generated networks for all years 1962–2009 is shown

in Fig. S-5. Even minimal values of T remain several times larger than the WTN values of T .

In Fig. S-6 we show the dependence of T on the trade threshold μ for the WTN data in year 2008. We see that there is only about 10–20% of variation of T for the range $10^{-5} \leq \mu \leq 10^{-3}$. Even for a much larger range $10^{-6} \leq \mu \leq 10^{-2}$ the variation of T remains smooth and remains in the bounds of 100%. This confirms

the stability of nestedness temperature in respect to broad range variations of μ . We present the majority of our data for $\mu = 10^{-3}$ which is approximately located in the flat range of T variation in year 2008. The data of Table 1 for EcoloRanking of countries at two different values of μ in year 2008 confirm the stability of this nestedness ordering. At the same time larger values of μ stress the importance of countries with a large trade volume, e.g. the position of China in export goes up from rank 5 at $\mu = 10^{-3}$ to rank 3 at $\mu = 10^{-2}$.

In Table 1 we present trade volume ranking and EcoloRanking of top 20 countries for import/export of WTN in year 2008.

In Table 2 we give the notations and symbols for Fig. 6 with corresponding SITC Rev. 1 codes and names. The list of all SITC Rev. 1 codes is available at [16] (see file <http://unstats.un.org/unsd/tradekb/Attachment193.aspx>). The colors of symbols in Fig. 4 mark the first digit of SITC Rev. 1 code: 0 – red (Food and live animals); 1 – does not appear in Fig. 4 (Beverages and tobacco); 2 – violet (Crude materials, inedible, except fuels); 3 – black (Mineral fuels, lubricants and related materials); 4 – does not appear in Fig. 4 (Animal and vegetable oils and fats); 5 – yellow (Chemicals); 6 – green (Manufactured goods classified chiefly by material); 7 – blue (Machinery and transport equipment); 8 – cyan (Miscellaneous manufactured articles); 9 – brown (Commod. and transacts. not class. accord. to kind).

References

- [1] R.M. May, Stability and Complexity in Model Ecosystems, Princeton Univ. Press, New Jersey, USA, 2001.
- [2] R.M. May, Nature 261 (1976) 459.
- [3] E. Ott, Chaos in Dynamical Systems, Cambridge Univ. Press, Cambridge, UK, 2002.
- [4] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of Networks, Oxford Univ. Press, Oxford, UK, 2003.
- [5] G. Caldarelli, Scale-Free Networks, Oxford Univ. Press, Oxford, UK, 2007.
- [6] G. Caldarelli, A. Vespignani (Eds.), Large Structure and Dynamics of Complex Networks, World Sci. Publ., Singapore, 2007.
- [7] M. Pascual, J.A. Dunne (Eds.), Ecological Networks: Linking Structure to Dynamics in Food Webs, Oxford Univ. Press, Oxford, UK, 2006.
- [8] J. Bascompte, P. Jordano, C.J. Melian, J.M. Olesen, Proc. Natl. Acad. Sci. USA 100 (2003) 9383.
- [9] D.P. Vázquez, M.A. Aizen, Ecology 85 (2004) 1251.
- [10] J. Memmott, N.M. Waser, M.V. Price, Proc. R. Soc. Lond. B 271 (2004) 2605.
- [11] J.M. Olesen, J. Bascompte, Y.L. Dupont, P. Jordano, Proc. Natl. Acad. Sci. USA 104 (2007) 19891.
- [12] E.L. Rezende, J.E. Lavabre, P.R. Guimarães, P. Jordano, J. Bascompte, Nature 448 (2007) 925.
- [13] E. Burgos, H. Ceva, R.P.J. Perazzo, M. Devoto, D. Medan, M. Zimmermann, A.M. Delbue, J. Theor. Biol. 249 (2007) 307.
- [14] U. Bastolla, M.A. Fortuna, A. Pascual-Garcia, A. Ferrera, B. Luque, J. Bascompte, Nature 458 (2009) 1018.
- [15] S. Saaverda, D.B. Stouffer, B. Uzzi, J. Bascompte, Nature 478 (2011) 233.
- [16] E. Burgos, H. Ceva, L. Hernández, R.P.J. Perazzo, M. Devoto, D. Medan, Phys. Rev. E 78 (2008) 046113;
- E. Burgos, H. Ceva, L. Hernández, R.P.J. Perazzo, Comput. Phys. Commun. 180 (2009) 532.
- [17] R.M. May, S.A. Levin, G. Sugihara, Nature 451 (2008) 893.
- [18] A.G. Haldane, R.M. May, Nature 469 (2011) 351.
- [19] United Nations Commodity Trade Statistics Database, <http://comtrade.un.org/db/>.
- [20] M.A. Rodríguez-Gironés, L. Santamaría, J. Biogeogr. 33 (2006) 924.
- [21] <http://ieg.ebd.csic.es/JordiBascompte/Resources.html>.
- [22] Central Intelligence Agency, The CIA Wold Factbook 2010, Skyhorse Publ. Inc., 2009.
- [23] L. Ermann, D.L. Shepelyansky, Acta Phys. Pol. A 120 (2011) A-158, <http://www.quantware.ups-tlse.fr/QWLIB/tradcheirank/>.

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/authorsrights>



Google matrix analysis of *C.elegans* neural network



V. Kandiah, D.L. Shepelyansky*

Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, F-31062 Toulouse, France

ARTICLE INFO

Article history:

Received 8 January 2014

Received in revised form 17 April 2014

Accepted 19 April 2014

Available online 30 April 2014

Communicated by C.R. Doering

ABSTRACT

We study the structural properties of the neural network of the *C.elegans* (worm) from a directed graph point of view. The Google matrix analysis is used to characterize the neuron connectivity structure and node classifications are discussed and compared with physiological properties of the cells. Our results are obtained by a proper definition of neural directed network and subsequent eigenvector analysis which recovers some results of previous studies. Our analysis highlights particular sets of important neurons constituting the core of the neural system. The applications of PageRank, CheiRank and ImpactRank to characterization of interdependency of neurons are discussed.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The human brain neural network has an enormous complexity containing about 10^{11} neurons and 10^{14} synapses linking various neurons [1]. Such a complex network can only be compared with the World Wide Web (WWW) which indexed size is estimated to be of about 10^{10} pages [2]. This comparison gives an idea that the methods of computer science, developed for WWW analysis, can be suitable for the investigations of neural networks. Among these methods the PageRank algorithm of the Google matrix of WWW [3] clearly demonstrated its efficiency being at the heart of Google search engine [4]. Thus we can expect that the Google matrix analysis can find useful applications for the neural networks. This approach has been tested in [5] on a reduced brain model of mammalian thalamocortical systems studied in [6]. However, it is more interesting to perform the Google matrix analysis for real neural networks. In this Letter we apply this analysis to characterize the properties of neural network of *C.elegans* (worm). The full connectivity of this directed network is known and documented at [7]. The number of linked neurons (nodes) is $N = 279$ with the number of synaptic connections and gap junctions (links) between them being $N_\ell = 2990$. This network is significantly smaller compared to the one studied in [5] but now we are working not with a model network but with the real worm network. Also, we use several new rank-based methods of network analysis comparing to those used in [5].

Recently, there is a growing interest to the complex network approach for investigation of brain neural networks [8–12]. Generally these networks are directional but it is difficult to determine directionality of links by physical and physiological measurements. Thus, at present, the worm network is practically the only example of neural network where the directionality of all links is established [7]. The analysis of certain properties this directed network has been reported recently in [11,12], however, the approach based on the Google matrix has not been used yet.

In the last years there is a clear trend to apply various advanced methods of network science to understand in a deeper way the connectivity properties of brain. Thus the properties of network centrality were used to characterize the human brain functional graphs [13]. A study of the whole connectivity matrix of the mouse brain has been reported recently [14]. Thus we think that our study will allow to highlight the features of worm network using recent advancements of computer science and push forward such methods for investigation of more complex brain networks.

2. Google matrix construction

The Google matrix G of *C.elegans* is constructed using the connectivity matrix elements $S_{ij} = S_{syn,ij} + S_{gap,ij}$, where S_{syn} is an asymmetric matrix of synaptic links whose elements are 1 if neuron j connects to neuron i through a chemical synaptic connection and 0 otherwise. The matrix part S_{gap} is a symmetric matrix describing gap junctions between pairs of cells, $S_{gap,ij} = S_{gap,ji} = 1$ if neurons i and j are connected through a gap junction and 0 otherwise. Following the standard rule [3,4], the matrix elements S_{ij} are renormalized ($S_{ij} = S_{ij} / \sum_i S_{ij}$) for each column with non-zero elements; the columns with all zero elements are replaced by

* Corresponding author.

E-mail addresses: kandiah@irsamc.ups-tlse.fr (V. Kandiah), dima@irsamc.ups-tlse.fr (D.L. Shepelyansky).

URL: <http://www.quantware.ups-tlse.fr/dima> (D.L. Shepelyansky).

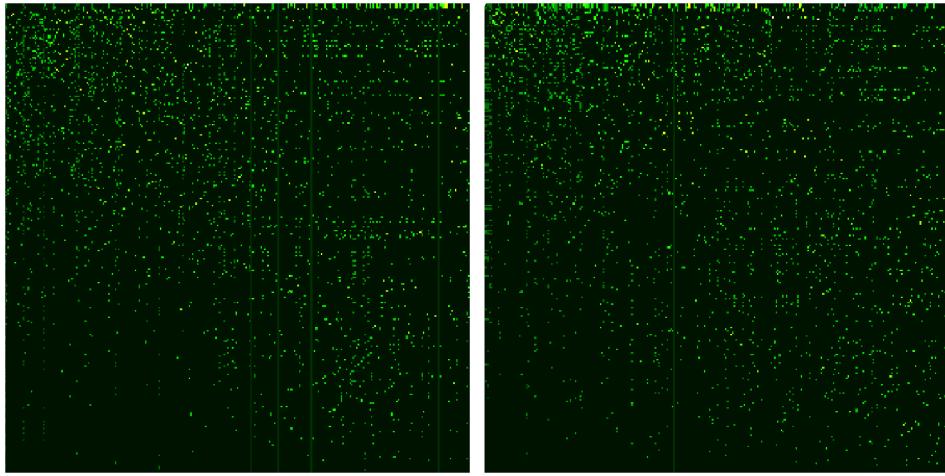


Fig. 1. (Color on-line.) Google matrix G (left) and G^* (right) for the neural network of *C.elegans* for $N = 279$ connected neurons. Matrix elements $G_{KK'}$ are shown in the basis of PageRank index K (and K') and elements $G_{K^*, K^{*\prime}}^*$ are shown in the basis of CheiRank index K^* (and $K^{*\prime}$) at $\alpha = 0.85$. Here, x and y axes show $1 \leq K, K' \leq N$ and $1 \leq K^*, K^{*\prime} \leq N$; the elements G_{11}, G^*_{11} are placed at the top left corner; color is proportional to the square root of matrix elements which are changing from black at minimum value $(1 - \alpha)/N$ to light yellow at maximum.

columns with all elements $1/N$. Thus the sum of elements in each column is equal to unity and the Google matrix takes the form

$$G_{ij} = \alpha S_{ij} + (1 - \alpha)/N. \quad (1)$$

Here α is the damping factor introduced in [3]. In the context of the WWW, the last term of the equation describes a probability for a random surfer to jump on any node of the network [4]. Below we use the usual value $\alpha = 0.85$ [4]. All matrix elements $S_{\text{syn},ij}, S_{\text{gap},ij}, S_{ij}$ are given at [15].

The eigenspectrum λ_i and right eigenvectors $\psi_i(j)$ of G satisfy the equation

$$\sum_{j'} G_{jj'} \psi_i(j') = \lambda_i \psi_i(j). \quad (2)$$

The eigenvector at $\lambda = 1$ is known as the PageRank vector. According to the Perron–Frobenius theorem [4] its elements $P(j) \sim \psi_1(j)$ are positive and their sum is normalized to unity. Thus $P(j)$ gives a probability to find a random surfer on a node j . All nodes can be ordered in a decreasing order of probability $P(K_j)$ with highest probability at top values of PageRank index $K_j = 1, 2, \dots$. For large matrices $P(j)$ can be found numerically by the iteration method [4] but for *C.elegans* case it can be obtained by a direct matrix diagonalization. We note that it is well established and verified for various complex networks that the PageRank distribution is stable in respect to variation of damping factor α in a range $0.5 \leq \alpha < 0.95$ [4]. We also checked that it is the case for our network and thus we used the usual value $\alpha = 0.85$.

It is also useful to consider the Google matrix obtained from the network with inverted directions of links (see e.g. [16–18]). The matrix G^* for this network with inverted direction of links is constructed following the same definition (1). The PageRank vector of this matrix G^* is called the CheiRank vector with probability $P^*(K_j^*)$ and CheiRank index K^* . According to the known results [3,4] the top nodes of PageRank are the most popular pages, while the top nodes of CheiRank are the most communicative nodes [17,18].

The structure of the matrix elements of G , presented in the PageRank ordering of nodes, and G^* , presented in the CheiRank ordering of nodes, is shown in Fig. 1. The number of nonzero elements N_G with indexes less than K is determined for various values of $K = 10, 100$. We find the values of ratio $N_G/K \approx 1.2, 10$ at $K = 10, 100$. These values correspond approximately to those of WWW networks of Universities of Cambridge and Oxford being

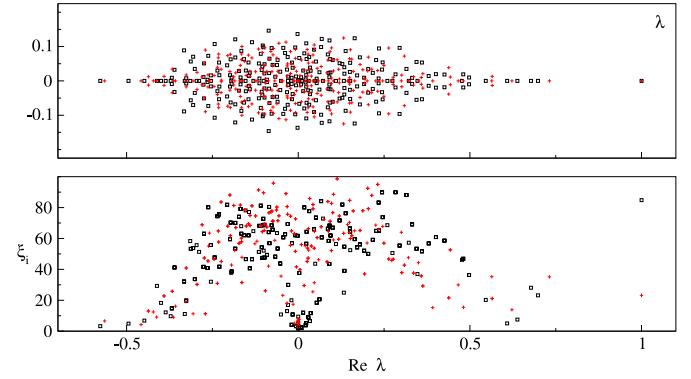


Fig. 2. (Color on-line.) Top panel: spectrum of eigenvalues λ for the Google matrices G and G^* at $\alpha = 0.85$ (black and red symbols). Bottom panel: IPR ξ of eigenvectors as a function of corresponding $\text{Re } \lambda$ (same colors).

significantly smaller than the values of Twitter network characterized by a strong connectivity between top PageRank nodes with $N_g/K \approx 100$ for $K = 100$ (see Fig. 2 in [20]). We note that the average number of links per neuron is $\eta = N_\ell/N = 10.71$ being approximately the same as for WWW of Universities of Cambridge and Oxford in 2006 [18].

The global matrix structure is asymmetric. This leads to a complex spectrum of eigenvalues of G and G^* as shown in top panel of Fig. 2. The imaginary part of eigenvalues is distributed in a range $-0.2 < \text{Im } \lambda < 0.2$ which is more narrow than for the networks of Wikipedia and UK universities [19]. This is related to a significant number of symmetric links. On the other side the networks of Le Monde or Python have comparable width for $\text{Im } \lambda$ [19]. We find that the second by modulus eigenvalues are $\lambda_2 = 0.8214$ for G and $\lambda_2 = 0.8608$ for G^* . Thus the network relaxation time $\tau = 1/|\ln \lambda_2|$ is approximately 5, 6.7 iterations of G, G^* .

The properties of eigenstates ψ_i can be characterized by the Inverse Participation Ratio (IPR) $\xi_i = (\sum_j |\psi_i(j)|^2)^2 / \sum_j |\psi_i(j)|^4$, which is broadly used in analysis of electron conductivity in disordered systems (see e.g. [19,20]). This quantity effectively determines the number of nodes on which is located an eigenstate ψ_i . We see that some eigenstates have rather large $\xi \approx N/3$ while others have ξ located only on about ten nodes. We will return to the discussion of properties of eigenstates later.

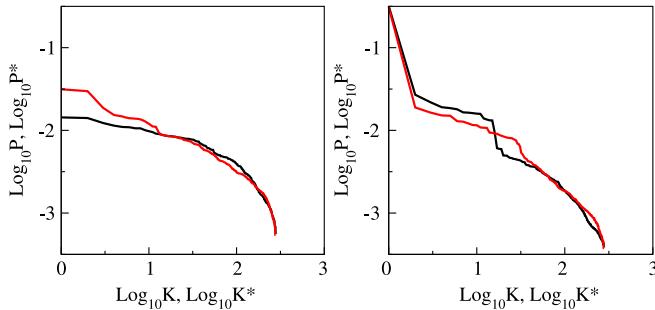


Fig. 3. (Color on-line.) Left panel: dependence of PageRank (CheiRank) probability $P(K)$ ($P^*(K^*)$) on its index K (K^*) shown by black (red) curve. Right panel: dependence of ImpactRank probability P (P^*) on its index K (K^*), obtained via propagator of G (G^*) at $\alpha = 0.85$ and $\gamma = 0.7$ for the initial probability located on neuron RMGL (see text).

3. CheiRank versus PageRank

The dependence of probabilities of PageRank and CheiRank vectors on their indexes K and K^* is shown in Fig. 3. A formal fit for a power law dependence $P \propto 1/K^\nu$, $P^* \propto 1/K^{*\nu}$ in the range $1 \leq K, K^* \leq 200$ gives $\nu = 0.33 \pm 0.03$ for PageRank and $\nu = 0.50 \pm 0.03$ for CheiRank. Of course, the number of nodes is small compared to the WWW or Wikipedia networks but on average we can say that a power law provides a satisfactory description of data. We note that the values of ν are notably smaller than the usual exponent value $\nu \approx 0.9$ (in K), 0.6 (in K^*) found for the WWW or Wikipedia networks (see e.g. [4,17]). Also, in our neural network we find that the exponent in K is smaller than in K^* while usually one finds the opposite situation. At the same time due to a small size of the network we do not claim that the exact value of ν is so important. It is better to say that its values give an indication of tendency. We think that for large size brain network this exponent can be determined with a better precision.

We also find that IPR $\xi \approx 85$ for P and $\xi \approx 23$ for P^* so that PageRank is distributed over a larger number of neurons. It is possible that such an inversion is related to a significant importance of outgoing links in neural systems: in a sense such links transfer orders, while ingoing links bring instructions to a given neuron from other neurons. We note that somewhat similar situation appears for networks of Business Process Management (BPM) where *Principals* of a company are located at the top CheiRank position while the top PageRank positions belong to company *Contacts* [21].

We note that our network is a directional network and as a result we have a significant asymmetry between ingoing and outgoing links. As a result the ranking nodes of PageRank and CheiRank have different probabilities and thus the top nodes have different functions. This fact is well known for directed networks (see e.g. [4,18,20,21]).

The correlations between PageRank and CheiRank vectors is convenient to characterize by the correlator $\kappa = N \sum_i P(i)P^*(i) - 1 = 0.125$. For *C.elegans* network the value of correlator is relatively small compared to those found for Wikipedia ($\kappa \approx 4$) and WWW of UK universities ($\kappa \sim 3$) [18]. In a sense for *C.elegans* neural network the situation if more similar to the networks of Linux Kernel ($\kappa \approx 0$) [16] and BMP ($\kappa = 0.164$) [21]. Thus, the *C.elegans* network has practically no correlations between ingoing and outgoing links. It is argued in [16,18] that such a network structure allows to perform a control of information flow in a more efficient way. Namely, it allows to reduce the propagation of errors in software codes. It seems that the neural networks also adopt such a structure.

Each neuron i belongs to two ranks K_i and K_i^* and it is convenient to represent the distribution of neurons on the two-dimensional plane (2D) of PageRank–CheiRank indexes (K, K^*) shown in Fig. 4. The plot confirms that there are little correlations be-

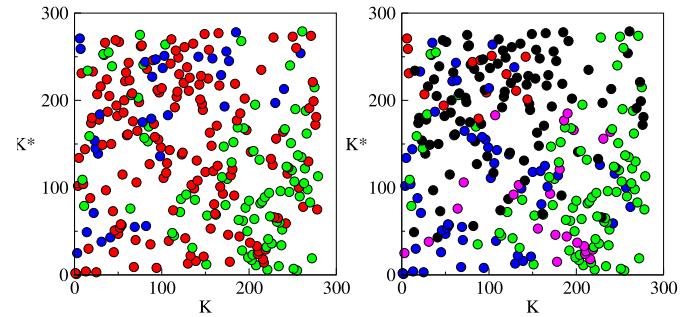


Fig. 4. (Color on-line.) PageRank–CheiRank plane (K, K^*) showing distribution of neurons according to their ranking. Left panel: soma region coloration – head (red), middle (green), tail (blue). Right panel: neuron type coloration – sensory (red), motor (green), interneuron (blue), polymodal (purple) and unknown (black). The classifications and colors are given according to WormAtlas [7].

Table 1

Top twenty neurons of PageRank (PR), CheiRank (CR); ImpactRank of G (IMPR) and G^* (IMCR) at initial state RMGL at $\gamma = 0.7$; following [7], the colors mark: interneurons (blue *bu*), motor neurons (green *gn*), sensory neurons (red *rd*), polymodal neurons (purple *pu*).

	PR	CR	IMPR	IMCR
1	AVAR (<i>bu</i>)	AVAL (<i>bu</i>)	RMGL (<i>bu</i>)	RMGL (<i>bu</i>)
2	AVAL (<i>bu</i>)	AVAR (<i>bu</i>)	URXL (<i>bu</i>)	AVAL (<i>bu</i>)
3	PVCR (<i>bu</i>)	AVBR (<i>bu</i>)	ADEL (<i>rd</i>)	ASHL (<i>rd</i>)
4	RIH (<i>bu</i>)	AVBL (<i>bu</i>)	AIAL (<i>bu</i>)	AVBR (<i>bu</i>)
5	AIAL (<i>bu</i>)	DD02 (<i>gn</i>)	IL2L (<i>rd</i>)	URXL (<i>bu</i>)
6	PHAL (<i>rd</i>)	VD02 (<i>gn</i>)	ADLL (<i>rd</i>)	AVEL (<i>bu</i>)
7	PHAR (<i>rd</i>)	DD01 (<i>gn</i>)	PVQL (<i>bu</i>)	RIBL (<i>bu</i>)
8	ADEL (<i>rd</i>)	RIBL (<i>bu</i>)	ALML (<i>rd</i>)	RMDR (<i>pu</i>)
9	HSNR (<i>gn</i>)	RIBR (<i>bu</i>)	ASKL (<i>rd</i>)	RMDL (<i>pu</i>)
10	RMGR (<i>bu</i>)	VD04 (<i>gn</i>)	CEPDL (<i>rd</i>)	RMDVL (<i>pu</i>)
11	VC03 (<i>gn</i>)	VD03 (<i>gn</i>)	ASHL (<i>rd</i>)	AVAR (<i>bu</i>)
12	AIAR (<i>bu</i>)	VD01 (<i>gn</i>)	AWBL (<i>rd</i>)	SIBVR (<i>bu</i>)
13	AVBL (<i>bu</i>)	AVER (<i>bu</i>)	SAADR (<i>bu</i>)	AIBR (<i>bu</i>)
14	PVPL (<i>bu</i>)	RMEV (<i>gn</i>)	RMHR (<i>gn</i>)	ADAL (<i>bu</i>)
15	AVM (<i>rd</i>)	RMDVR (<i>pu</i>)	RMHL (<i>gn</i>)	RMHL (<i>gn</i>)
16	AVKL (<i>bu</i>)	AVEL (<i>bu</i>)	RIH (<i>bu</i>)	AVBL (<i>bu</i>)
17	HSNL (<i>gn</i>)	VD05 (<i>gn</i>)	OLQVL (<i>pu</i>)	SIBVL (<i>bu</i>)
18	RMGL (<i>bu</i>)	SMDDR (<i>pu</i>)	AIML (<i>bu</i>)	ASKL (<i>rd</i>)
19	AVHR (<i>bu</i>)	DD03 (<i>gn</i>)	HSNL (<i>gn</i>)	RID (<i>bu</i>)
20	AVFL (<i>bu</i>)	VA02 (<i>gn</i>)	SDQR (<i>bu</i>)	SMBVL (<i>pu</i>)

tween both ranks since the points are scattered over the whole plane. Neurons ranked at top K positions of PageRank have their soma located mainly in both extremities of the worm (head and tail) showing that neurons in those regions have important connections coming from many other neurons which control head and tail movements. This tendency is even more visible for neurons at top K^* positions of CheiRank but with a preference for head and middle regions. In general, neurons, that have their soma in the middle region of the worm, are quite highly ranked in CheiRank but not in PageRank. The neurons located at the head region have top positions in CheiRank and also PageRank, while the middle region has some top CheiRank indexes but rather large indexes of PageRank (Fig. 4 left panel). The neuron type coloration (Fig. 4 right panel) also reveals that sensory neurons are at top PageRank positions but at rather large CheiRank indexes, whereas in general motor neurons are in the opposite situation.

The top 20 neurons of PageRank and CheiRank vectors are given in the first two columns of Table 1. We note that both rankings favor important signal relaying neurons such as *AVA* and *AVB* that integrate signals from crucial nodes and in turn pilot other crucial nodes. Neurons *AVAL*, *AVAR*, *AVBL*, *AVBR* and *AVER*, *AVFL* are considered to belong to the rich club analyzed in [12]. The right panel

Table 2

Top ten neurons of the eigenvectors of G (left panel) and G^* (right panel) corresponding to the 10th largest eigenvalues $|\lambda|$; IPR are respectively $\xi \approx 5$ and $\xi \approx 4$.

	$\lambda_{10} = -0.49446$	$ \psi_i $		$\lambda_{10} = -0.45784$	$ \psi_i^* $
1	AIAR	0.11986	1	AVAL	0.10651
2	AIAL	0.11159	2	AVAR	0.079403
3	ASIL	0.096475	3	AVBR	0.036779
4	ASIR	0.096236	4	VD05	0.025086
5	AWAR	0.024228	5	VA09	0.02438
6	ASHR	0.022241	6	VD06	0.020977
7	RMGR	0.018502	7	VA08	0.020242
8	AIMR	0.018387	8	AVBL	0.019225
9	ADLL	0.01837	9	DD02	0.018684
10	PVQL	0.017547	10	PDB	0.016485

Table 3

Same as in Table 2 for 48th largest eigenvalue modulus $|\lambda|$; IPR are respectively $\xi \approx 54$ and $\xi \approx 47$.

	$\lambda_{48} = -0.30615 - 0.07037i$	$ \psi_i $		$\lambda_{48} = 0.26353 - 0.095716i$	$ \psi_i^* $
1	RIH	0.017854	1	RMEV	0.026461
2	BDUR	0.017737	2	RIBR	0.013343
3	OLLR	0.016701	3	OLQDR	0.013145
4	CEPDR	0.016463	4	IL1DL	0.012932
5	RMGR	0.016357	5	IL1DR	0.012911
6	AIAL	0.016072	6	RIAR	0.012896
7	ASHR	0.015585	7	RICR	0.012728
8	VC04	0.015265	8	OLQDL	0.012586
9	ASKR	0.014	9	RIGR	0.012256
10	IL2R	0.013978	10	SMDDR	0.011958

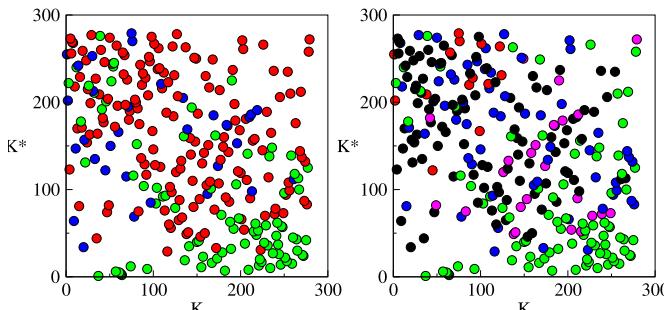


Fig. 5. (Color on-line.) Distribution of neurons in the plane (K, K^*) of equal opportunity ranks (see text); colors are the same as in Fig. 4.

in Fig. 3 and second two columns of Table 1 represent ImpactRank which is discussed below.

We can also use 2DRank index K_2 , discussed in [17], which counts nodes in order of their appearance on ribs of squares in (K, K^*) plane with the square size growing from $K = 1$ to $K = N$. The top neurons in K_2 are AVAL, AVAR, AVBL, AVBR, PVCR. Thus at the top K_2 values we find dominance of interneurons. More detailed listings are available at [15].

It may be also useful to consider renormalized equal opportunity rank recently discussed in [22]. In this approach PageRank probability of node i is replaced by $P(i)/d(i)$ where $d(i)$ is in-degree of node i . For the Google matrix this recipe should be replaced by $P(i) \rightarrow P(i)/\sum_j G_{ij}$ and respectively for CheiRank by $P^*(i) \rightarrow P^*(i)/\sum_j G_{ij}^*$. The corresponding rank indexes K, K^* rank the neurons in the decreasing order of these renormalized probabilities. The distribution of nodes in the plane (K, K^*) is shown in Fig. 5. In this ranking the top K nodes correspond to important sensory neurons rather than information relaying centers, whereas the top nodes of K^* are composed mainly by motor neurons. Thus such an approach allows to highlight additional features of *C. elegans* network being complementary to PageRank and CheiRank properties discussed above. Tables for neuron renormalized ranking are available at [15].

4. ImpactRank

In certain cases it is useful to determine an influence or impact of a given neuron on other neurons. A recent proposal of ImpactRank, described in [20], is based on the probability distribution of a vector $v_f = (1 - \gamma)(1 - \gamma G)^{-1}v_0$, $v_f^* = (1 - \gamma)(1 - \gamma G^*)^{-1}v_0$, where v_0 is initially populated neuron. The vector v_f can be viewed as a Green function propagator. The computation of v_f is obtained numerically by a summation of geometrical expansion series which are convergent within approximately first 200 terms at $\gamma \sim 0.7$ (see also [20]). The distributions of probabilities of ImpactRank $P(i) = v_f(i)$, $P^*(i) = v_f^*(i)$ versus the corresponding ImpactRank indexes K, K^* are shown in Fig. 3 (right panel) for the initial state neuron RMGL. The corresponding top 20 ImpactRank neurons influenced by RMGL are given in columns IMPR, IMCR of Table 1. The analysis of neurons linked to RMGL shows that indeed, ImpactRank correctly selects neurons influenced by RMGL. The neurons in the top list of $P(i)$ are those pointed by outgoing links of RMGL while those in the top list of $P^*(i)$ are those that have ingoing links to RMGL. Such a method can be easily applied to other initial neuron states of interest showing a contamination propagation over the neural network starting from initial neuron RMGL.

5. Properties of eigenstates

The Google matrix analysis of the Wikipedia hyperlink network [19] demonstrated that the eigenstates with large values of $|\lambda|$ select well defined communities. Thus we can expect that other eigenstates of matrices G and G^* with $|\lambda| < 1$ correspond to certain physiological functions of worm neural network. It is convenient to order index of eigenstates in a decreasing order of $|\lambda_i|$ with $\lambda_1 = 1$.

The top ten neurons in eigenfunction amplitude for four specific eigenstates of G and G^* are given in Table 2, Table 3. In Table 2 we have eigenstates with low value of IPR so that the corresponding wavefunctions are localized essentially on only about 4 neurons being AIAR, AIAL, ASIL, ASIR and AVAL, AVBR for λ_{10} of G

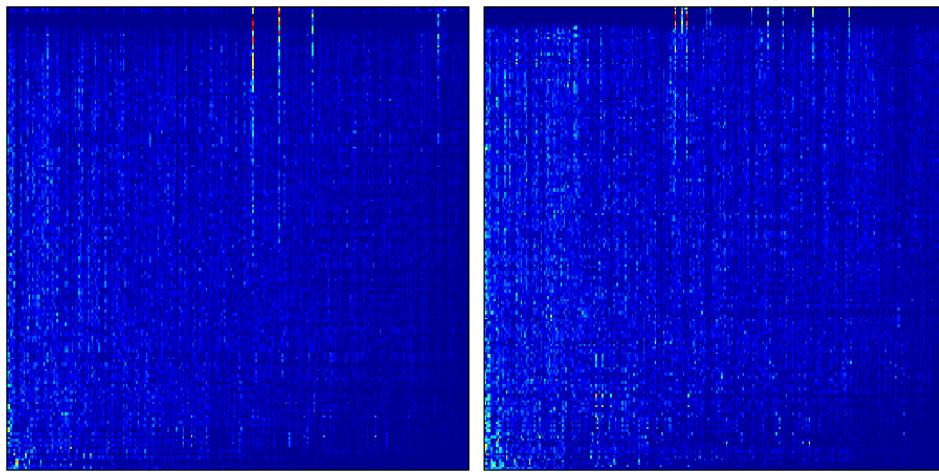


Fig. 6. (Color on-line.) Dependence of amplitude of eigenstates $|\psi_i(K)|$ of G and $|\psi_i^*(K^*)|$ of G^* on PageRank index K (left panel) and CheiRank index K^* (right panel); here x -axis shows values of K and K^* , while y -axis shows index i of eigenstates being ordered in a decreasing order of $|\lambda_i|$ (see text). The whole index range $1 \leq K, K^* \leq 279$ is shown with PageRank (CheiRank) vector being at the bottom line of each panel. The color is proportional to $|\psi_i(j)|$ changing from minimum blue value to maximum value in red.

and G^* respectively. In Table 3 the values of IPR are rather large and these eigenstates are spread over many neurons.

To determine if some eigenvectors are localized on a certain group of neurons, we plot in Fig. 6 the amplitude of each eigenstate horizontally in the basis of neurons ordered by indexes of K and K^* of PageRank and CheiRank vectors. The eigenstates of G matrix show four distinct vertical stripes at $K = 149$, $K = 165$, $K = 185$, $K = 261$ which correspond respectively to neurons *PVDR*, *IL2DR*, *IL2DL*, *PLNR*. The same plot for G^* matrix shows a larger number of stripes which have less pronounced amplitudes. These stripes of G^* are located on the following neurons $K^* = 116$ (*RIPR*), $K^* = 123$ (*RIPR*), $K^* = 120$ (*AS07*), $K^* = 122$ (*AS10*), $K^* = 135$ (*DB06*), $K^* = 137$ (*DB05*), $K^* = 215$ (*DA07*), $K^* = 162$ (*VA10*), $K^* = 172$ (*SIADL*), $K^* = 181$ (*SIAVL*), $K^* = 199$ (*SIAVR*), $K^* = 221$ (*SIADR*).

We think that an identification of eigenstates with certain physiological functions of worm can be an interesting task which however requires further more detailed studies in collaboration with physiologists. The tables of top 20 nodes of eigenstates with 50 largest $|\lambda_i|$ values are available at [15].

6. Discussion

In this Letter, we analyzed the neural network of *C.elegans* using Google matrix approach to directed networks which proved its efficiency for the WWW. We classify worm neurons using PageRank and CheiRank probabilities corresponding to the principal vectors of the Google matrix with direct and inverted links. Thus neurons in the head region take top positions in PageRank, CheiRank and combined 2DRank. Also, interneurons occupy top ranking positions. We show that influences and interdependency between neurons can be studied using the ImpactRank propagator. We argue that the eigenvectors with large modulus of eigenvalues of the Google matrix may select specific physiological functions. This conjecture still need to be investigated in more detailed studies. Our research shows that the Google matrix analysis represents a useful and powerful method of neural network analysis.

Acknowledgements

We thank Emma K. Towlson and Petra E. Vértes for useful discussions and for providing us the links between neurons available from *C.elegans* neural network data set at [7]. This work is supported in part by EC FET Open project “New tools and algorithms for directed network analysis” (NADINE No. 288956).

References

- [1] R.W. Williams, K. Herrup, *Annu. Rev. Neurosci.* 11 (1988) 423.
- [2] <http://www.worldwidewebsize.com/>.
- [3] S. Brin, L. Page, *Comput. Netw. ISDN Syst.* 30 (1998) 107.
- [4] A.M. Langville, C.D. Meyer, *Google's PageRank and Beyond*, Princeton University Press, 2006.
- [5] D.L. Shepelyansky, O.V. Zhirov, *Phys. Lett. A* 374 (2010) 3206.
- [6] E.M. Izhikevich, G.M. Edelman, *Proc. Natl. Acad. Sci.* 105 (2008) 3593.
- [7] Z.F. Altun, L.A. Herndon, C. Crocker, R. Lints, D.H. Hall (Eds.), *WormAtlas*, <http://www.wormatlas.org>, 2012.
- [8] A. Arenas, A. Fernández, S. Gómez, in: *Bio-Inspired Computing and Communication*, in: *Lect. Notes Comput. Sci.*, vol. 5151, 2008, p. 9.
- [9] E. Bullmore, O. Sporns, *Nat. Rev. Neurosci.* 10 (2009) 312.
- [10] S. Varier, M. Kaiser, *PLoS Comput. Biol.* 7 (1) (2011) e1001044.
- [11] L.R. Varshney, B.L. Chen, E. Paniagua, D.H. Hall, D.B. Chklovskii, *PLoS Comput. Biol.* 7 (2011) e1001066.
- [12] E.K. Towlson, P.E. Vértes, S.E. Ahnert, W.R. Schafer, E.T. Bullmore, *J. Neurosci.* 33 (15) (2013) 6380.
- [13] X.-N. Zuo, R. Ehmke, M. Mennes, D. Imperati, F.X. Castellanos, O. Sporns, M.P. Milham, *Cereb. Cortex* 22 (2012) 1862.
- [14] S.W. Oh, et al., *Nature* 508 (2014) 207.
- [15] V. Kandiah, D.L. Shepelyansky, <http://www.quantware.ups-tlse.fr/QWLIB/wormgooglematrix/>, 2013.
- [16] A.D. Chepeliantskii, [arXiv:1003.5455 \[cs.SE\]](https://arxiv.org/abs/1003.5455), 2010.
- [17] A.O. Zhirov, O.V. Zhirov, D.L. Shepelyansky, *Eur. Phys. J. B* 77 (2010) 523.
- [18] L. Ermann, A.D. Chepeliantskii, D.L. Shepelyansky, *J. Phys. A, Math. Theor.* 45 (2012) 275101.
- [19] L. Ermann, K.M. Frahm, D.L. Shepelyansky, *Eur. Phys. J. B* 86 (2013) 193.
- [20] K.M. Frahm, Y.-H. Eom, D.L. Shepelyansky, [arXiv:1310.5624 \[physics.soc-ph\]](https://arxiv.org/abs/1310.5624), 2013.
- [21] M. Abel, D.L. Shepelyansky, *Eur. Phys. J. B* 84 (2011) 493.
- [22] D. Bankey, G. Ivan, V. Grolmusz, *PLoS ONE* 8 (1) (2013) e54204.

Poisson statistics of PageRank probabilities of Twitter and Wikipedia networks

Klaus M. Frahm and Dima L. Shepelyansky^a

Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France

Received 24 February 2014 / Received in final form 18 March 2014

Published online 16 April 2014 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2014

Abstract. We use the methods of quantum chaos and Random Matrix Theory for analysis of statistical fluctuations of PageRank probabilities in directed networks. In this approach the effective energy levels are given by a logarithm of PageRank probability at a given node. After the standard energy level unfolding procedure we establish that the nearest spacing distribution of PageRank probabilities is described by the Poisson law typical for integrable quantum systems. Our studies are done for the Twitter network and three networks of Wikipedia editions in English, French and German. We argue that due to absence of level repulsion the PageRank order of nearby nodes can be easily interchanged. The obtained Poisson law implies that the nearby PageRank probabilities fluctuate as random independent variables.

1 Introduction

The PageRank vector $P(K)$ of the Google matrix G_{ij} had been proposed by Brin and Page for ranking of nodes of the World Wide Web (WWW) [1]. At present the PageRank algorithm became a fundamental element of various search engines including Google search [2]. This ranking works reliably also for other networks like the Physical Review citation network [3,4], Wikipedia [5–7] and other networks including even the world trade network [8]. Thus it is important to understand the statistical properties of the PageRank vector.

To study the properties of PageRank probabilities we use the standard approach [1,2] following the notation used in reference [6]. The directed network is constructed in a usual way: a directed link is formed from a node j to a node i when j quotes i and an element A_{ij} of the adjacency matrix is taken to be unity when there is such a link and zero in absence of link. Then the matrix S_{ij} of Markov transitions is constructed by normalizing elements of each column to unity ($\sum_i S_{ij} = 1$) and replacing columns with only zero elements (*dangling nodes*) by $1/N$, with N being the matrix size. Then the Google matrix of the network takes the form [1,2]:

$$G_{ij} = \alpha S_{ij} + (1 - \alpha)/N. \quad (1)$$

The damping parameter α in the WWW context describes the probability $(1 - \alpha)$ to jump to any node for a random surfer. For WWW the Google search engine uses $\alpha \approx 0.85$ [2]. The matrix G belongs to the class of Perron-Frobenius operators [2], its largest eigenvalue is $\lambda = 1$ and other eigenvalues have $|\lambda| \leq \alpha$. The right eigenvector

at $\lambda = 1$, which is called the PageRank, has real non-negative elements $P(i)$ and gives a probability $P(i)$ to find a random surfer at site i . Thus we can rank all nodes in a decreasing order of PageRank probability $P(K(i))$ so that the PageRank index $K(i)$ counts all N nodes i according to their ranking, placing the most popular nodes at the top values $K = 1, 2, 3 \dots$ In numerical simulations the vector $P(K_i)$ can be obtained by the power iteration method [2]. The Arnoldi method allows to compute efficiently a significant number of eigenvalues and eigenvectors corresponding to large values of $|\lambda|$ (see e.g. [9–11]).

From a physical viewpoint we can make a conjecture that the PageRank probabilities are described by a steady-state quantum Gibbs distribution [12] over certain quantum levels with energies E_i . In the frame of this conjecture the PageRank probabilities on nodes i are given by:

$$P(i) = \exp(-E_i/T)/Z, \quad Z = \sum_i \exp(-E_i/T) \quad (2)$$

and inversely the effective energies E_i are given by:

$$E_i = -T \ln P(i) - T \ln Z. \quad (3)$$

Here Z is the statistical sum and T is a certain effective temperature. In some sense the above conjecture assumes that the operator matrix G can be represented as a sum of two operators G_H and G_{NH} where G_H describes a hermitian system while G_{NH} represents a non-Hermitian operator which creates a system thermalization at a certain effective temperature T with the quantum Gibbs distribution over energy levels E_i of operator G_H . The last term in (3) is independent of i and gives a global energy shift which is not important. We note that PageRank probabilities describe a stationary state of G and its probability

^a e-mail: dima@irsamc.ups-tlse.fr

can be always presented in the form (3). Thus our method can be used for any directed network. However, implicitly it is assumed that the relaxation dynamics is a complex process and that a considered network has many nodes and many complex links between nodes.

The statistical properties of fluctuations of levels have been extensively studied in the fields of Random Matrix Theory (RMT) [13] and quantum chaos [14]. The most direct characteristics is the probability distribution $p(s)$ of level spacings s statistics. Here $s = (E_{i+1} - E_i)/\Delta E$ is a spacing between nearest levels measured in the units of average local energy spacing ΔE . Thus the probability distribution $p(s)$ is obtained via the unfolding procedure which takes into account the variation of energy level density with energy E [14]. We note that the value of T in (3) does not influence the statistics $p(s)$ due to spectrum unfolding and definition of s in units of local level spacing.

In the field of quantum chaos it is well established that $p(s)$ is a powerful tool to characterize the spectral properties of quantum systems. For quantum systems, which have a chaotic dynamics in the classical limit (e.g. Sinai or Bunimovich billiards [15]), it is known that in generic cases the statistics $p(s)$ is the same as for the RMT, invented by Wigner to describe the spectra of complex nuclei [13,16,17]. This statement is known as the Bohigas-Giannoni-Schmit conjecture [16]. In such cases the distribution is well described by the so-called Wigner surmise $p(s) = (\pi s/2) \exp(-\pi s^2/4)$ [14,17]. For integrable quantum systems (e.g. circular or elliptic billiards) one finds a Poisson distribution $p(s) = \exp(-s)$ corresponding to the fluctuations of random independent variables. Such a Poisson distribution is drastically different from the RMT results characterized by the level repulsion at small s values.

The strong feature of $p(s)$ statistics is that it describes the universal statistical fluctuations. Thus its use for description of PageRank fluctuations is very relevant, it provides a new statistical information about PageRank properties. We describe the results obtained within such an approach in next sections.

2 Statistical properties of PageRank probabilities

For our studies we use the network of entire Twitter 2009 studied in [11] with number of nodes $N = 41\,652\,230$ and number of links $N_\ell = 1\,468\,365\,182$; network of English Wikipedia (Aug 2009; noted below as Wikipedia) articles from [5] with $N = 3\,282\,257$, $N_\ell = 71\,012\,307$; German Wikipedia (dated November 2013, noted below as Wikipedia-DE) with $N = 1\,532\,977$, $N_\ell = 36\,781\,077$ and French Wikipedia (dated November 2013; noted below as Wikipedia-FR) with $N = 1\,352\,825$, $N_\ell = 34\,431\,943$. For the last two cases we use the network data collected by Vigna [18].

For a given network the PageRank is computed as usually by the power or iteration method for a typical value of the damping factor $\alpha = 0.85$. The probabilities P_i

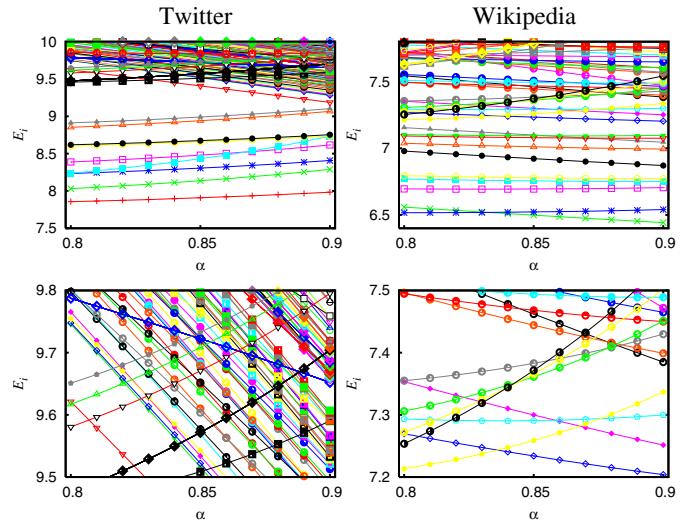


Fig. 1. Dependence of top PageRank levels $E_i = -\ln(P_i)$ on the damping factor α for Twitter (left panel) and Wikipedia (right panel). Data points on curves with the same color symbol correspond to the same node i . The lower panels are obtained by a zoom in an energy range from the top panels. About 150 (for Twitter) or 50 (for Wikipedia) lowest levels are shown in top panels.

are computed with a relative precision better than 10^{-12} . For each node i its PageRank value P_i is associated to a pseudo-energy E_i by the relation $E_i = -\ln(P_i)$. Obviously the energy spectrum is ordered if the index is given in the rank index K , i.e. $E_{K+1} \geq E_K$. Therefore the number n of levels below a given pseudo-energy E is given by $n = K$ if $E_K < E < E_{K+1}$ (we also use index i for E_i).

The evolution of energy levels E_i with the variation of the damping factor α is shown in Figure 1 for Twitter and Wikipedia networks. The results show many level crossings which are typical of Poisson statistics. We note that here each level has its own index so that it is rather easy to see if there is a real or avoided level crossing. In this respect the situation is simpler compared to energy levels in quantum systems.

In the following we fix the damping factor to the standard value $\alpha = 0.85$. To obtain the unfolded spectrum with an average uniform level spacing of unity (see e.g. [14]) one has to replace the function E_i by a smooth function. As shown in Figure 2, one can very well approximate E_K by a polynomial $Q(x)$ of modest degree in the variable $x = \ln(K)$. In this procedure it is better to exclude the first ten nodes with $K \leq 10$ which do not affect the global statistics. For a fit range $10 < K \leq 10^4$ a polynomial of degree 2 is already sufficient. However, for larger intervals, e.g. $10 < K \leq 10^7$ for Twitter or $10 < K \leq 10^6$ for Wikipedia it is better to increase the polynomial degree up to 20. Once the polynomial fit is known one obtains the unfolded energy eigenvalues S_i by solving the equation $E_i = Q(\ln(S_i))$ using the Newton method. For each energy the obtained value of $S_i \approx i$ is rather close to $K = i$ index with an average spacing of unity. In certain cases this equation does not provide a solution for energies

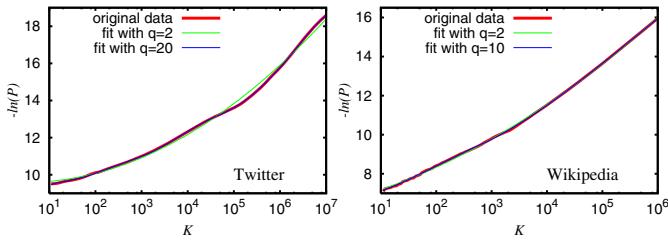


Fig. 2. The thick red curve shows $-\ln(P) = E$ versus K for the PageRank probability P of Twitter (Wikipedia) in the left (right) panel. The thin green curve corresponds to the fit $-\ln(P) = Q(\ln(K))$ where $Q(x)$ is a polynomial of degree $q = 2$. The thin blue curve corresponds to the fit with a polynomial of degree $q = 20$ ($q = 10$). The fits are obtained for the range $10 < K \leq 10^7$ ($10 < K \leq 10^6$) with weights $\sim 1/K$ attributed to each data point. Here and in next figures $\alpha = 0.85$.

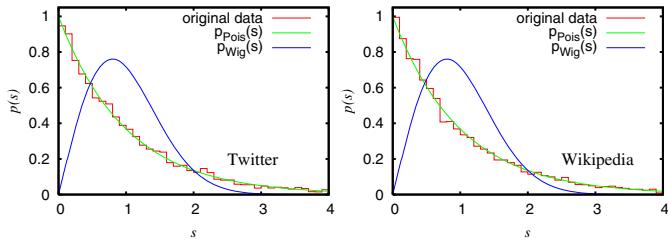


Fig. 3. Histogram of unfolded level spacing statistics using pseudo-energies $E_i = -\ln(P_i)$ of Twitter (Wikipedia) shown in the left (right) panel. The unfolding is done with the fit shown in Figure 2 using a polynomial of degree 2 and a fit range $10 < K \leq 10^4$. The Poisson distribution $p_{\text{Pois}}(s) = \exp(-s)$ and the Wigner surmise $p_{\text{Wig}}(s) = \frac{\pi}{2} s \exp(-\frac{\pi}{4} s^2)$ are also shown for comparison.

close to the boundary of the fit range. In these cases the unfolded spectrum is slightly reduced with respect to the initial fit range.

In Figure 3 only a polynomial of degree 2 is used since the fit range $10 < K \leq 10^4$ is rather small and the histogram fluctuations, compared with the Poisson distribution, are still quite considerable due to the limited number of $N_s \sim 10^4$ data points. The obtained data show a good agreement of results with the Poisson statistics.

In Figure 4 we show the integrated probability to find a level spacing larger than s :

$$I_p(s) = \int_s^\infty d\tilde{s} p(\tilde{s}). \quad (4)$$

This quantity is numerically more stable since no histogram is required. One simply orders the spacings $s_i = S_{i+1} - S_i$ and draws the ratio $1 - i/N_s$ versus s_i where i is the ordering index of the spacings and N_s is the number of spacings in the numerical data.

The data shown in Figure 4 clearly demonstrate that $I_p(s)$ follows the Poisson expression $I_p(s) = \exp(-s)$ for a quite large range of level spacings. Of course, for the largest values of s there are deviations which are either due to the lack of statistics (especially for modest values

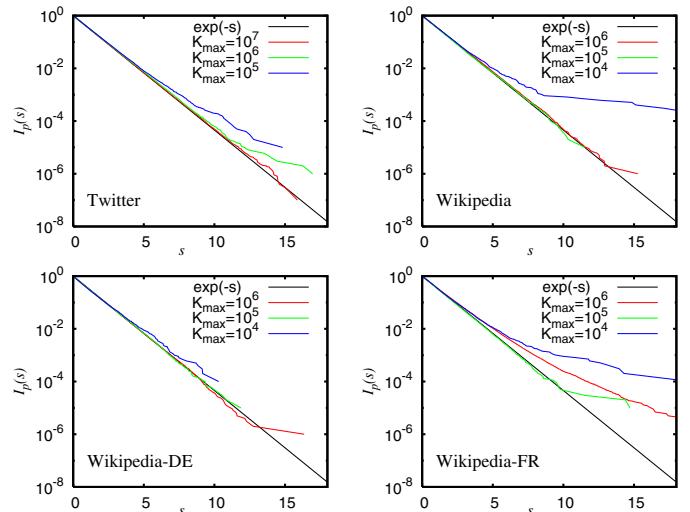


Fig. 4. The color curves show the integrated probability $I_p(s) = \int_s^\infty d\tilde{s} p(\tilde{s})$, given in semi-logarithmic representation, for the PageRank probabilities for networks of Twitter, Wikipedia, Wikipedia-DE and Wikipedia-FR. The unfolding is done as in Figure 2 using a fit polynomial of degree 20 and a fit range $10 < K \leq K_{\text{max}}$ with three different values of K_{max} given in the panels. The black line corresponds to $I_p(s) = \exp(-s)$ obtained for the case of Poisson distributed levels.

of K_{max}) or due to the fact that the number of levels is close to the total network size.

We also note that for large values of $K \geq 10^6$ there are N_d degenerate nodes with identical $P(i)$ values with at least one more another node or a few nodes. Such an effect has been pointed in reference [11]. These artificial degeneracies provide an additional delta function contribution $w_0 \delta(s)$ in the Poisson statistics $p(s)$ where w_0 is the probability to find such a degeneracy. There are about $N_d \approx 10^2$ ($N_d \approx 10^5$) degeneracies for Twitter nodes for $K < 10^6$ ($K < 10^7$) which gives $w_0 \approx 10^{-4}$ ($w_0 \approx 10^{-2}$). In a histogram of bin-width $\Delta s = 0.1$ this gives a relative change of the height of the first bin at $s = 0$ of $10 w_0 \approx 10^{-3}$ ($\approx 10^{-1}$) and unless we use too large K value the statistical contribution of such degenerate nodes is indeed very small.

We note that if we use all nodes of Twitter up to $K < 4.2 \times 10^7$ we have $N_d \approx 1.1 \times 10^7$ with $w_0 \approx 0.26$ which is indeed considerable. In this particular case also the distribution of close degeneracies ($0 < s \ll 1$) is quite different from the (rescaled) Poisson distribution $(1 - w_0) \exp[-(1 - w_0) s]$ for the non-degenerate levels. Apparently a particular network structure, which is responsible for the degeneracies, also enhances the number of close degeneracies. We attribute the appearance of such degeneracies to weak interconnections between nodes at the tail of PageRank probability where the fluctuations are not stabilized being sensible to the finite network size.

Our data show that the Poisson statistics gives a good description of fluctuations of PageRank probabilities. It may be interesting to determine what are the nodes which have very large spacings s from nearest levels on both sides. It is natural to expect that those nodes will be rather

Table 1. List of nodes with unfolded neighbor level spacings $s_i = S_i - S_{i-1} > 4$ for Wikipedia network.

K	$S_i - S_{i-1}$	$S_{i+1} - S_i$	Title
996	8.43535	6.57294	Henry VIII of England
2966	4.07317	4.09474	The Age
3398	4.21163	4.65018	Debt
3982	4.30229	4.01818	GREEN
6098	4.42446	4.78164	Vomiting
6632	4.22776	4.38045	Mary I of Scotland
9388	4.42904	4.94249	Simulation

stable in respect to modifications of network or damping factor variations. Such nodes for Wikipedia network are shown in Table 1 for $s > 4$ and $K < 10^4$. Such a selection captures two important figures of English history but the reasons for appearing of other nodes still need to be clarified. We think that a further study of nodes with large statistical deviations of spacing values can provide a new interesting information about robust nodes of a given network. Even if such events are due to random fluctuations still it is interesting to analyze the properties of such extreme events. The validity of the Poisson statistics means that the ranking order can be easily interchanged between nodes with nearby values of PageRank index K .

We also analyzed the statistics of PageRank probabilities for a random triangular matrix model (triangular RPBM) introduced in reference [19]. We find here the Poisson statistics. We also consider CheiRank probability vector of Wikipedia (it is given by the PageRank probability for the Wikipedia network with inverted direction of links) [5] and also find here the Poisson distribution.

3 Discussion

We use the methods of quantum chaos to study the statistical fluctuations of PageRank probabilities in four networks of Twitter, Wikipedia English, German and French. We associated the effective pseudo-energy levels E_i to PageRank probabilities via the relation $E_i = -\ln P_i$ and use the unfolding level density procedure to have homogeneous spacings between levels. This procedure is commonly used in the field of quantum chaos (see e.g. [14,17]). Our studies show that the level spacing statistics is well described by the Poisson distribution $p(s) = \exp(-s)$. Thus there is any sign of level repulsion typical of the quantum chaotic billiards [16] and RMT [13]. Such a result can be considered as a natural one for nodes with large values of PageRank index K where nodes can be assumed as independent. However, the Poisson distribution remains valid even for relatively low values $K \leq 10^4$ where a significant number of links exist between the users of Twitter as discussed in reference [11]. Thus even a large number of links between top nodes does not lead to their interdependence so that nearby PageRank probabilities behave themselves as random independent variables. In all examples of large directed networks considered we found the Poisson statistics. We can make a conjecture that this is

a generic situation. However, it may happen that some networks can have a repulsion of nodes and, who knows, may the Wigner-Dyson statistics.

We should note that the relation $E_i = -\ln P_i$, used in our studies to have a correspondence with level spacing statistics, is not really so important since after that we apply the unfolding procedure. Due to this our method simply gives us the fluctuations of nearby PageRank probabilities in a correctly weighted dimensionless representation where the validity of Poisson distribution becomes directly visible. We think that the investigation of nodes with large spacings with nearby nodes in K can provide a new useful information for network analysis.

This research is supported in part by the EC FET Open project “New tools and algorithms for directed network analysis” (NADINE No. 288956). We thank Sebastiano Vigna for providing us the network data for German and French Wikipedia, collected in the frame of NADINE project; these data sets can be obtained from the web page of Vigna [18].

References

1. S. Brin, L. Page, Comput. Networks and ISDN Systems **30**, 107 (1998)
2. A.M. Langville, C.D. Meyer, *Google’s PageRank and Beyond: The Science of Search Engine Rankings* (Princeton University Press, Princeton, 2006)
3. S. Redner, Phys. Today **58**, 49 (2005)
4. F. Radicchi, S. Fortunato, B. Markines, A. Vespignani, Phys. Rev. E **80**, 056103 (2009)
5. A.O. Zhirov, O.V. Zhirov, D.L. Shepelyansky, Eur. Phys. J. B **77**, 523 (2010)
6. Y.-H. Eom, K.M. Frahm, A. Benczur, D.L. Shepelyansky, Eur. Phys. J. B **86**, 492 (2013)
7. Y.-H. Eom, D.L. Shepelyansky, PLoS ONE **8**, e74554 (2013)
8. L. Ermann, D.L. Shepelyansky, Acta Phys. Pol. A **120**, A158 (2011)
9. G.W. Stewart, *Matrix Algorithms Volume II: Eigensystems* (SIAM, 2001)
10. K.M. Frahm, D.L. Shepelyansky, Eur. Phys. J. B **76**, 57 (2010)
11. K.M. Frahm, D.L. Shepelyansky, Eur. Phys. J. B **85**, 355 (2012)
12. L.D. Landau, E.M. Lifshitz, *Statistical Mechanics* (Nauka, Moskva, 1976) (in Russian), Vol. 5
13. M.L. Mehta, *Random Matrices* (Elsevier-Academic Press, Amsterdam, 2004)
14. F. Haake, *Quantum Signatures of Chaos* (Springer, Berlin, 2010)
15. L. Bunimovich, Scholarpedia **2**, 1813 (2007)
16. O. Bohigas, M.-J. Giannoni, C. Schmit, Phys. Rev. Lett. **52**, 1 (1984)
17. H.-J. Stöckmann, Scholarpedia **5**, 10243 (2010)
18. S. Vigna, <http://vigna.di.unimi.it/>
19. K.M. Frahm, Y.-H. Eom, D.L. Shepelyansky, arXiv:1310.5624 [physics.soc-ph] (2013)

Google matrix analysis of directed networks

Leonardo Ermann

Departamento de Física Teórica, GlyA, Comisión Nacional de Energía Atómica, Buenos Aires, Argentina

Klaus M. Frahm and Dima L. Shepelyansky

Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France

(Dated: July 25, 2014)

In past ten years, modern societies developed enormous communication and social networks. Their classification and information retrieval processing become a formidable task for the society. Due to the rapid growth of World Wide Web, social and communication networks, new mathematical methods have been invented to characterize the properties of these networks on a more detailed and precise level. Various search engines are essentially using such methods. It is highly important to develop new tools to classify and rank enormous amount of network information in a way adapted to internal network structures and characteristics. This review describes the Google matrix analysis of directed complex networks demonstrating its efficiency on various examples including World Wide Web, Wikipedia, software architecture, world trade, social and citation networks, brain neural networks, DNA sequences and Ulam networks. The analytical and numerical matrix methods used in this analysis originate from the fields of Markov chains, quantum chaos and Random Matrix theory.

Keywords: Markov chains, World Wide Web, search engines, complex networks, PageRank, 2DRank, CheiRank

“The Library exists *ab aeterno*.”
Jorge Luis Borges *The Library of Babel*

Contents

I. Introduction	2	B. Universal emergence of PageRank	17
II. Scale-free properties of directed networks	3	C. Two-dimensional ranking for University networks	19
III. Construction of Google matrix and its properties	3	IX. Wikipedia networks	19
A. Construction rules	3	A. Two-dimensional ranking of Wikipedia articles	19
B. Markov chains and Perron-Frobenius operators	5	B. Spectral properties of Wikipedia network	20
C. Invariant subspaces	5	C. Communities and eigenstates of Google matrix	21
D. Arnoldi method for numerical diagonalization	6	D. Top people of Wikipedia	22
E. General properties of eigenvalues and eigenstates	7	E. Multilingual Wikipedia editions	22
IV. CheiRank versus PageRank	7	F. Networks and entanglement of cultures	25
A. Probability decay of PageRank and CheiRank	7	X. Google matrix of social networks	26
B. Correlator between PageRank and CheiRank	7	A. Twitter network	27
C. PageRank-CheiRank plane	8	B. Poisson statistics of PageRank probabilities	28
D. 2DRank	8	XI. Google matrix analysis of world trade	29
E. Historical notes on spectral ranking	9	A. Democratic ranking of countries	29
V. Complex spectrum and fractal Weyl law	9	B. Ranking of countries by trade in products	30
VI. Ulam networks	10	C. Ranking time evolution and crises	31
A. Ulam method for dynamical maps	10	D. Ecological ranking of world trade	31
B. Chirikov standard map	10	E. Remarks on world trade and banking networks	35
C. Dynamical maps with strange attractors	12	XII. Networks with nilpotent adjacency matrix	36
D. Fractal Weyl law for Perron-Frobenius operators	12	A. General properties	36
E. Intermittency maps	13	B. PageRank of integers	36
F. Chirikov typical map	13	C. Citation network of Physical Review	37
VII. Linux Kernel networks	14	XIII. Random matrix models of Markov chains	40
A. Ranking of software architecture	14	A. Albert-Barabási model of directed networks	40
B. Fractal dimension of Linux Kernel Networks	15	B. Random matrix models of directed networks	40
VIII. WWW networks of UK universities	16	C. Anderson delocalization of PageRank?	41
A. Cambridge and Oxford University networks	16	XIV. Other examples of directed networks	43
XV. Discussion		A. Brain neural networks	43
		B. Google matrix of DNA sequences	45
		C. Gene regulation networks	48
		D. Networks of game go	48
		E. Opinion formation on directed networks	49
			51

XVI. Acknowledgments	52
References	52

I. INTRODUCTION

On a scale of ten years, modern societies developed enormous communication and social networks. The World Wide Web (WWW) alone has about 50 billion indexed web pages, so that their classification and information retrieval processing become a formidable task which the society has to face every day. Various search engines have been developed by private companies such as Google, Yahoo! and others which are extensively used by Internet users. In addition, social networks (Facebook, LiveJournal, Twitter, etc) gained enormous popularity in the last few years. Active use of social networks spreads beyond their initial purposes making them important for political or social events.

To handle such enormous databases, fundamental mathematical tools and algorithms related to centrality measures and network matrix properties are actively being developed. Indeed, the PageRank algorithm, which was initially at the basis of the development of the Google search engine (Brin and Page , 1998; Langville and Meyer, 2006), is directly linked to the mathematical properties of Markov chains (Markov , 1906) and Perron-Frobenius operators (Brin and Stuck, 2002; Langville and Meyer, 2006). Due to its mathematical foundation, this algorithm determines a ranking order of nodes that can be applied to various types of directed networks. However, the recent enormous development of WWW and communication networks requires the creation of new tools and algorithms to characterize the properties of these networks on a more detailed and precise level. For example, such networks contain weakly coupled or secret communities which may correspond to very small values of the PageRank and are hard to detect. It is therefore highly important to have new methods to classify and rank enormous amount of network information in a way adapted to internal network structures and characteristics.

This review describes matrix tools and algorithms which facilitate classification and information retrieval from large networks recently created by human activity. The Google matrix formed by links of the network has typically a huge size. Thus, the analysis of its spectral properties including complex eigenvalues and eigenvectors represents a challenge for analytical and numerical methods. It is rather surprising, but the class of such matrices, belonging to the class of Markov chains and Perron-Frobenius operators, was practically not investigated in physics. Indeed, usually the physical problems belong to the class of Hermitian or unitary matrices. Their properties had been actively studied in the frame of Random Matrix Theory (RMT) (Akemann *et al.*, 2011; Guhr *et al.*, 1998; Mehta, 2004) and quantum chaos (Haake, 2010). The analytical and numerical tools

developed in these research fields allowed to understand many universal and peculiar features of such matrices in the limit of large matrix size corresponding to many-body quantum systems (Guhr *et al.*, 1998), quantum computers (Shepelyansky , 2001) and a semiclassical limit of large quantum numbers in the regime of quantum chaos (Haake, 2010). In contrast to the Hermitian problem, the Google matrices of directed networks have complex eigenvalues. The only physical systems where similar matrices had been studied analytically and numerically correspond to models of quantum chaotic scattering whose spectrum is known to have such unusual properties as the fractal Weyl law (Gaspard, 2014; Nonnenmacher and Zworski , 2007; Shepelyansky , 2008; Sjöstrand , 1990; Zworski , 1999).

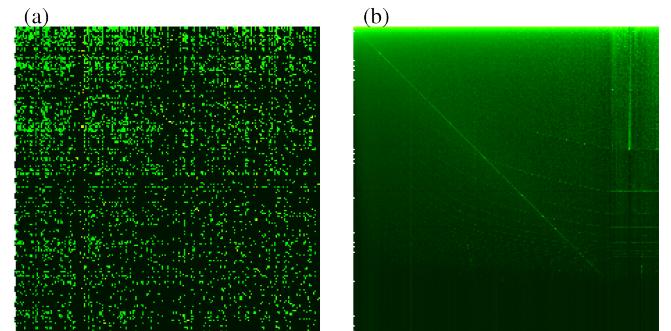


FIG. 1 (Color online) Google matrix of the network Wikipedia English articles for Aug 2009 in the basis of PageRank index K (and K'). Matrix $G_{KK'}$ corresponds to x (and y) axis with $1 \leq K, K' \leq 200$ on panel (a), and with $1 \leq K, K' \leq N$ on panel (b); all nodes are ordered by PageRank index K of matrix G and thus we have two matrix indexes K, K' for matrix elements in this basis. Panel (a) shows the first 200×200 matrix elements of G matrix (see Sec. III). Panel (b) shows density of all matrix elements coarse-grained on 500×500 cells where its elements, $G_{K,K'}$, are written in the PageRank basis $K(i)$ with indexes $i \rightarrow K(i)$ (in x -axis) and $j \rightarrow K'(j)$ (in a usual matrix representation with $K = K' = 1$ on the top-left corner). Color shows the density of matrix elements changing from black for minimum value $((1 - \alpha)/N)$ to white for maximum value via green (gray) and yellow (light gray); here the damping factor is $\alpha = 0.85$ After (Ermann *et al.*, 2012a).

In this review we present extensive analysis of a variety of Google matrices emerging from real networks in various sciences including WWW of UK universities, Wikipedia, Physical Review citation network, Linux Kernel network, world trade network from the UN COM-TRADE database, brain neural networks, networks of DNA sequences and many others. As an example, the Google matrix of Wikipedia network of English articles (2009) is shown in Fig. 1. We demonstrate that the analysis of the spectrum and eigenstates of a Google matrix of a given network provides a detailed understanding about the information flow and ranking. We also show that such type of matrices naturally appear for Ulam networks of dynamical maps (Frahm and Shepelyansky , 2012b; She-

pelyansky and Zhirov , 2010a) in the framework of the Ulam method (Ulam, 1960).

At present, Wikipedia, a free online encyclopaedia, stores more and more information becoming the largest database of human knowledge. In this respect it is similar to the Library of Babel, described by Jorge Luis Borges (Borges, 1962). The understanding of hidden relations between various areas of knowledge on the basis of Wikipedia can be improved with the help of Google matrix analysis of directed hyperlink network of Wikipedia articles as described in this review.

The RMT and quantum chaos tools, combined with the efficient numerical methods for large matrix diagonalization like the Arnoldi method (Stewart, 2001), allow to analyze the spectral properties of such large matrices as an entire Twitter network of 41 millions users (Frahm and Shepelyansky , 2012b). In 1998 Brin and Page pointed out that “despite the importance of large-scale search engines on the web, very little academic research has been done on them” (Brin and Page , 1998). We hope that this review provides solid mathematical basis of matrix methods of efficient analysis of directed networks emerging in various sciences. The described methods will find broad interdisciplinary applications in mathematics, physics and computer science with the cross-fertilization of different research fields.

An interested reader can find a general information about complex networks (see also Sec. II) in well established papers, reviews and books (Watts and Strogatz , 1998), (Albert and Barabási , 2002; Caldarelli, 2003; Newman , 2003), (Castellano *et al.*, 2009; Dorogovtsev *et al.*, 2008), (Dorogovtsev, 2010; Fortunato , 2010; Newman, 2010). Descriptions of Markov chains and Perron-Frobenius operators are given in (Brin and Page , 1998; Langville and Meyer, 2006) while properties of Random Matrix Theory (RMT) and quantum chaos are described in (Akemann *et al.*, 2011; Guhr *et al.*, 1998; Haake, 2010; Mehta, 2004).

The data sets of the main part of networks considered here are available at (FETNADINE database, 2014) from Quantware group.

II. SCALE-FREE PROPERTIES OF DIRECTED NETWORKS

The distributions of the number of ingoing or outgoing links per node for directed networks with N nodes and N_ℓ links are well known as indegree and outdegree distributions in the community of computer science (Caldarelli, 2003; Donato *et al.*, 2004; Pandurangan *et al.*, 2005). A network is described by an adjacency matrix A_{ij} of size $N \times N$ with $A_{ij} = 1$ when there is a link from a node j to a node i in the network, i. e. “ j points to i ”, and $A_{ij} = 0$ otherwise. Real networks are often characterized by power law distributions for the number of ingoing and outgoing links per node $w_{\text{in,out}}(k) \propto 1/k^{\mu_{\text{in,out}}}$ with typical exponents $\mu_{\text{in}} \approx 2.1$ and $\mu_{\text{out}} \approx 2.7$ for the WWW.

For example, for the Wikipedia network of Fig. 1 one finds $\mu_{\text{in}} = 2.09 \pm 0.04$, $\mu_{\text{out}} = 2.76 \pm 0.06$ as shown in Fig. 2 (Zhirov *et al.*, 2010).

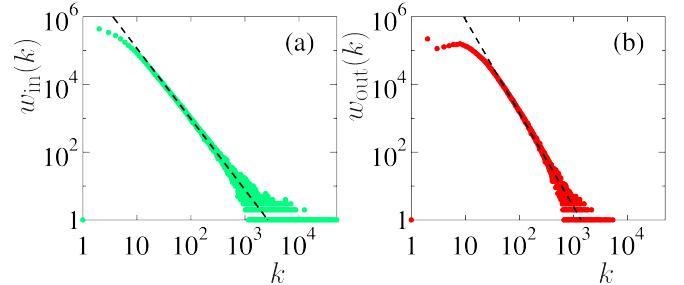


FIG. 2 (Color online) Distribution $w_{\text{in,out}}(k)$ of number of ingoing (a) and outgoing (b) links k for $N = 3282257$ Wikipedia English articles (Aug 2009) of Fig. 1 with total number of links $N_\ell = 71012307$. The straight dashed fit line shows the slope with $\mu_{\text{in}} = 2.09 \pm 0.04$ (a) and $\mu_{\text{out}} = 2.76 \pm 0.06$ (b). After (Zhirov *et al.*, 2010).

Statistical preferential attachment models were initially developed for undirected networks (Albert and Barabási , 2000). Their generalization to directed networks (Giraud *et al.*, 2009) generates a power law distribution for ingoing links with $\mu_{\text{in}} \approx 2$ but the distribution of outgoing links is more close to an exponential decay. We will see below that these models are not able to reproduce the spectral properties of G in real networks.

The most recent studies of WWW, crawled by the Common Crawl Foundation in 2012 (Meusel *et al.*, 2014) for $N \approx 3.5 \times 10^9$ nodes and $N_\ell \approx 1.29 \times 10^{11}$ links, provide the exponents $\mu_{\text{in}} \approx 2.24$, $\mu_{\text{out}} \approx 2.77$, even if the authors stress that these distributions describe probabilities at the tails which capture only about one percent of nodes. Thus, at present the existing statistical models of networks capture only in an approximate manner the real situation in large networks.

III. CONSTRUCTION OF GOOGLE MATRIX AND ITS PROPERTIES

A. Construction rules

The matrix S_{ij} of Markov transitions (Markov , 1906) is constructed from the adjacency matrix $A_{ij} \rightarrow S_{ij}$ by normalizing elements of each column so that their sum is equal to unity ($\sum_i S_{ij} = 1$) and replacing columns with only zero elements (*dangling nodes*) by $1/N$. Such matrices with columns sum normalized to unity and $S_{ij} \geq 0$ belong to the class of Perron-Frobenius operators with a possibly degenerate unit eigenvalue $\lambda = 1$ and other eigenvalues obeying $|\lambda| \leq 1$ (see Sec. III.B). Then the Google matrix of the network is introduced as: (Brin and Page , 1998)

$$G_{ij} = \alpha S_{ij} + (1 - \alpha)/N . \quad (1)$$

The damping factor α in the WWW context describes the probability $(1 - \alpha)$ to jump to any node for a random surfer. For WWW the Google search engine uses $\alpha \approx 0.85$ (Langville and Meyer, 2006). For $0 \leq \alpha \leq 1$ the matrix G also belongs to the class of Perron-Frobenius operators as S and with its columns sum normalized. However, for $\alpha < 1$ its largest eigenvalue $\lambda = 1$ is not degenerate and the other eigenvalues lie inside a smaller circle of radius α , i.e. $|\lambda| \leq \alpha$ (Brin and Stuck, 2002; Langville and Meyer, 2006).

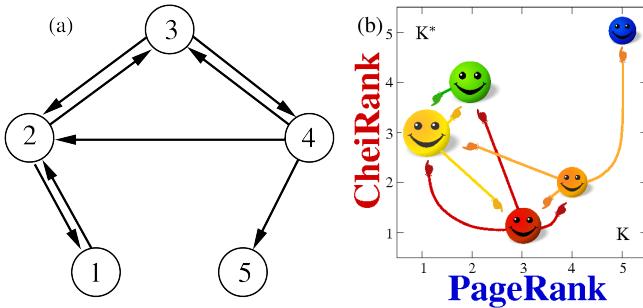


FIG. 3 (Color online) (a) Example of simple network with directed links between 5 nodes. (b) Distribution of 5 nodes from (a) on the PageRank-CheiRank plane (K, K^*), where the size of node is proportional to PageRank probability $P(K)$ and color of node is proportional to CheiRank probability $P^*(K^*)$, with maximum at red/gray and minimum at blue/black; the location of nodes of panel (a) on (K_i, K_i^*) plane is: (2, 4), (1, 3), (3, 1), (4, 2), (5, 5) for original nodes $i = 1, 2, 3, 4, 5$ respectively; PageRank and CheiRank vectors are computed from the Google matrices G and G^* shown in Fig. 4 at a damping factor $\alpha = 0.85$.

The right eigenvector at $\lambda = 1$, which is called the PageRank, has real nonnegative elements $P(i)$ and gives the probability $P(i)$ to find a random surfer at site i . The PageRank can be efficiently determined by the power iteration method which consists of repeatedly multiplying G to an iteration vector which is initially chosen as a given random or uniform initial vector. Developing the initial vector in a basis of eigenvectors of G one finds that the other eigenvector coefficients decay as $\sim \lambda^n$ and only the PageRank component, with $\lambda = 1$, survives in the limit $n \rightarrow \infty$. The finite gap $1 - \alpha \approx 0.15$ between the largest eigenvalue and other eigenvalues ensures, after several tens of iterations, the fast exponential convergence of the method also called the “PageRank algorithm”. A multiplication of G to a vector requires only $O(N_\ell)$ multiplications due to the links and the additional contributions due to dangling nodes and damping factor can be efficiently performed with $O(N)$ operations. Since often the average number of links per node is of the order of a few tens for WWW and many other networks one has effectively N_ℓ and N of the same order of magnitude. At $\alpha = 1$ the matrix G coincides with the matrix S and we will see below in Sec. VIII that for this case the largest eigenvalue $\lambda = 1$ is usually highly degenerate due to many invariant subspaces which define many in-

dependent Perron-Frobenius operators with at least one eigenvalue $\lambda = 1$ for each of them.

Once the PageRank is found, e.g. at $\alpha = 0.85$, all nodes can be sorted by decreasing probabilities $P(i)$. The node rank is then given by the index $K(i)$ which reflects the relevance of the node i . The top PageRank nodes, with largest probabilities, are located at small values of $K(i) = 1, 2, \dots$.

It is known that the PageRank probability is proportional to the number of ingoing links (Langville and Meyer, 2006; Litvak *et al.*, 2008), characterizing how popular or known a given node is. Assuming that the PageRank probability decays algebraically as $P_i \sim 1/K_i^\beta$ we obtain that the number of nodes N_P with PageRank probability P scales as $N_P \sim 1/P^{\mu_{\text{in}}}$ with $\mu_{\text{in}} = 1 + 1/\beta$ so that $\beta \approx 0.9$ for $\mu_{\text{in}} \approx 2.1$ being in agreement with the numerical data for WWW (Donato *et al.*, 2004; Meusel *et al.*, 2014; Pandurangan *et al.*, 2005) and Wikipedia network (Zhirov *et al.*, 2010).

In addition to a given directed network with adjacency matrix A it is useful to analyze an inverse network where links are inverted and whose adjacency matrix A^* is the transpose of A , i.e. $A_{ij}^* = A_{ji}$. The matrices S^* and the Google matrix G^* of the inverse network are then constructed in the same way from A^* as described above and according to the relation (1) using the same value of α as for the G matrix. The right eigenvector of G^* at eigenvalue $\lambda = 1$ is called CheiRank giving a complementary rank index $K^*(i)$ of network nodes (Chepelianskii, 2010; Ermann *et al.*, 2012a; Zhirov *et al.*, 2010). The CheiRank probability $P^*(K^*)$ is proportional to the number of outgoing links highlighting node communicativity (see e.g. (Ermann *et al.*, 2012a; Zhirov *et al.*, 2010)). In analogy with the PageRank we obtain that $P^* \sim 1/K^{*\beta}$ with $\beta = 1/(\mu_{\text{out}} - 1) \approx 0.6$ for typical $\mu_{\text{out}} \approx 2.7$. The statistical properties of distribution of nodes on the PageRank-CheiRank plane are described in (Ermann *et al.*, 2012a) for various directed networks. We will discuss them below.

$$\begin{aligned}
 (a) \quad & A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} & (b) \quad & A^* = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \\
 (c) \quad & S = \begin{pmatrix} 0 & 1/2 & 1/3 & 0 & 1/5 \\ 1 & 0 & 1/3 & 1/3 & 1/5 \\ 0 & 1/2 & 0 & 1/3 & 1/5 \\ 0 & 0 & 1/3 & 0 & 1/5 \\ 0 & 0 & 0 & 1/3 & 1/5 \end{pmatrix} & (d) \quad & S^* = \begin{pmatrix} 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/2 & 1/3 & 0 & 1 & 0 \\ 0 & 1/3 & 1/2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \\
 (e) \quad & G = \begin{pmatrix} 0.03 & 0.455 & 0.313 & 0.03 & 0.2 \\ 0.88 & 0.03 & 0.313 & 0.313 & 0.2 \\ 0.03 & 0.455 & 0.03 & 0.313 & 0.2 \\ 0.03 & 0.03 & 0.313 & 0.03 & 0.2 \\ 0.03 & 0.03 & 0.03 & 0.313 & 0.2 \end{pmatrix} & (f) \quad & G^* = \begin{pmatrix} 0.03 & 0.313 & 0.03 & 0.03 & 0.03 \\ 0.455 & 0.03 & 0.455 & 0.03 & 0.03 \\ 0.455 & 0.313 & 0.03 & 0.88 & 0.03 \\ 0.03 & 0.313 & 0.455 & 0.03 & 0.88 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.03 \end{pmatrix}
 \end{aligned}$$

FIG. 4 (a) Adjacency matrix A of network of Fig. 3(a) with indexes used there, (b) adjacency matrix A^* for the network with inverted links; matrices S (c) and S^* (d) corresponding to the matrices A , A^* ; the Google matrices G (e) and G^* (f) corresponding to matrices S and S^* for $\alpha = 0.85$ (only 3 digits of matrix elements are shown).

For an illustration we consider an example of a simple network of five nodes shown in Fig. 3(a). The corresponding adjacency matrices A , A^* are shown in Fig. 4 for the indexes given in Fig. 3(a). The matrices of Markov transitions S , S^* and Google matrices are computed as described above and from Eq. (1). The distribution of nodes on (K, K^*) plane is shown in Fig. 3(b). After permutations the matrix G can be rewritten in the basis of PageRank index K as it is done in Fig. 1.

B. Markov chains and Perron-Frobenius operators

Matrices with real non-negative elements and column sums normalized to unity belong to the class of Markov chains (Markov , 1906) and Perron-Frobenius operators (Brin and Stuck, 2002), which have been used in a mathematical analysis of dynamical systems. A numerical analysis of finite size approximants of such operators is closely linked with the Ulam method (Ulam, 1960) which naturally generates such matrices for dynamical maps (Erman and Shepelyansky , 2010a,b; Shepelyansky and Zhirov , 2010a). The Ulam method generates Ulam networks whose properties are discussed in Sec.VI.

Matrices G of this type have at least (one) unit eigenvalue $\lambda = 1$ since the vector $e^T = (1, \dots, 1)$ is obviously a left eigenvector for this eigenvalue. Furthermore one verifies easily that for any vector v the inequality $\|Gv\|_1 \leq \|v\|_1$ holds where the norm is the standard 1-norm. From this inequality one obtains immediately that all eigenvalues λ of G lie in a circle of radius unity: $|\lambda| \leq 1$. For the Google matrix G as given in (1) one can furthermore show for $\alpha < 1$ that the unity eigenvalue is not degenerate and the other eigenvalues obey even $|\lambda| \leq \alpha$ (Langville and Meyer, 2006).

It should be pointed out that due to the asymmetry of links on directed networks such matrices have in general a complex eigenvalue spectrum and sometimes they are not even diagonalizable, i.e. there may also be generalized eigenvectors associated to non-trivial Jordan blocks. Matrices of this type rarely appear in physical problems which are usually characterized by Hermitian or unitary matrices with real eigenvalues or located on the unitary circle. The universal spectral properties of such hermitian or unitary matrices are well described by RMT (Akemann *et al.*, 2011; Guhr *et al.*, 1998; Haake, 2010). In contrast to this non-trivial complex spectra appear in physical systems only in problems of quantum chaotic scattering and systems with absorption. In such cases it may happen that the number of states N_γ , with finite values $0 < \lambda_{\min} \leq |\lambda| \leq 1$ ($\gamma = -2 \ln |\lambda|$), can grow algebraically $N_\gamma \propto N^\nu$ with increasing matrix size N , with an exponent $\nu < 1$ corresponding to a fractal Weyl law proposed first in mathematics (Sjöstrand , 1990). Therefore most of eigenvalues drop to $\lambda = 0$ with $N \rightarrow \infty$. We discuss this unusual property in Sec.V.

C. Invariant subspaces

For typical networks the set of nodes can be decomposed in invariant *subspace nodes* and fully connected *core space nodes* leading to a block structure of the matrix S in (1) which can be represented as (Frahm *et al.*, 2011):

$$S = \begin{pmatrix} S_{ss} & S_{sc} \\ 0 & S_{cc} \end{pmatrix}. \quad (2)$$

The core space block S_{cc} contains the links between core space nodes and the coupling block S_{sc} may contain links from certain core space nodes to certain invariant subspace nodes. By construction there are no links from nodes of invariant subspaces to the nodes of core space. Thus the subspace-subspace block S_{ss} is actually composed of many diagonal blocks for many invariant subspaces whose number can generally be rather large. Each of these blocks corresponds to a column sum normalized matrix with positive elements of the same type as G and has therefore at least one unit eigenvalue. This leads to a high degeneracy N_1 of the eigenvalue $\lambda = 1$ of S , for example $N_1 \sim 10^3$ as for the case of UK universities (see Sec. VIII).

In order to obtain the invariant subspaces, we determine iteratively for each node the set of nodes that can be reached by a chain of non-zero matrix elements of S . If this set contains all nodes (or at least a macroscopic fraction) of the network, the initial node belongs to the *core space* V_c . Otherwise, the limit set defines a subspace which is invariant with respect to applications of the matrix S . At a second step all subspaces with common members are merged resulting in a sequence of disjoint subspaces V_j of dimension d_j and which are invariant by applications of S . This scheme, which can be efficiently implemented in a computer program, provides a subdivision over N_c core space nodes (70-80% of N for UK university networks) and $N_s = N - N_c$ subspace nodes belonging to at least one of the invariant subspaces V_j . This procedure generates the block triangular structure (2). One may note that since a dangling node is connected by construction to all other nodes it belongs obviously to the core space as well as all nodes which are linked (directly or indirectly) to a dangling node. As a consequence the invariant subspaces do not contain dangling nodes nor nodes linked to dangling nodes.

The detailed algorithm for an efficient computation of the invariant subspaces is described in (Frahm *et al.*, 2011). As a result the total number of all subspace nodes N_s , the number of independent subspaces N_d , the maximal subspace dimension d_{\max} etc. can be determined. The statistical properties for the distribution of subspace dimensions are discussed in Sec. VIII for UK universities and Wikipedia networks. Furthermore it is possible to determine numerically with a very low effort the eigenvalues of S associated to each subspace by separate diagonalization of the corresponding diagonal blocks in the matrix S_{ss} . For this, either exact diagonalization or, in

rare cases of quite large subspaces, the Arnoldi method (see the next subsection) can be used.

After the subspace eigenvalues are determined one can use the Arnoldi method to the projected core space matrix block S_{cc} to determine the leading core space eigenvalues. In this way one obtains accurate eigenvalues because the Arnoldi method does not need to compute the numerically very problematic highly degenerate unit eigenvalues of S since the latter are already obtained from the separate and cheap subspace diagonalization. Actually the alternative and naive application of the Arnoldi method on the full matrix S , without computing the subspaces first, does not provide the correct number N_1 of degenerate unit eigenvalues and also the obtained clustered eigenvalues, close to unity, are not very accurate. Similar problems hold for the full matrix G (with damping factor $\alpha < 1$) since here only the first eigenvector, the PageRank, can be determined accurately but there are still many degenerate (or clustered) eigenvalues at (or close to) $\lambda = \alpha$.

Since the columns sums of S_{cc} are less than unity, due to non-zero matrix elements in the block S_{sc} , the leading core space eigenvalue of S_{cc} is also below unity $|\lambda_1^{(\text{core})}| < 1$ even though in certain cases the gap to unity may be very small (see Sec. VIII).

We consider concrete examples of such decompositions in Sec. VIII and show in this review spectra with subspace and core space eigenvalues of matrices S for several network examples. The mathematical results for properties of the matrix S are discussed in (Serra-Capizzano , 2005).

D. Arnoldi method for numerical diagonalization

The most adapted numerical method to determine the largest eigenvalues of large sparse matrices is the Arnoldi method (Arnoldi , 1951; Frahm and Shepelyansky , 2010; Golub and Greif , 2006; Stewart, 2001). Indeed, usually the matrix S in Eq. (1) is very sparse with only a few tens of links per node $\zeta = N_\ell/N \sim 10$. Thus, a multiplication of a vector by G or S is numerically cheap. The Arnoldi method is similar in spirit to the Lanczos method, but is adapted to non-Hermitian or non-symmetric matrices. Its main idea is to determine recursively an orthonormal set of vectors $\xi_0, \dots, \xi_{n_A-1}$, which define a *Krylov space*, by orthogonalizing $S\xi_k$ on the previous vectors ξ_0, \dots, ξ_k by the Gram-Schmidt procedure to obtain ξ_{k+1} and where ξ_0 is some normalized initial vector. The dimension n_A of the Krylov space (in the following called the *Arnoldi-dimension*) should be “modest” but not too small. During the Gram-Schmidt procedure one obtains furthermore the explicit expression: $S\xi_k = \sum_{j=0}^{k+1} h_{jk} \xi_j$ with matrix elements h_{jk} , of the Arnoldi representation matrix of S on the Krylov space, given by the scalar products or inverse normalization constants calculated during the orthogonalization. In order to obtain a closed representation matrix one needs to replace the last coupling

element $h_{n_A, n_A-1} \rightarrow 0$ which introduces a mathematical approximation. The eigenvalues of the $n_A \times n_A$ matrix h are called the *Ritz eigenvalues* and represent often very accurate approximations of the exact eigenvalues of S , at least for a considerable fraction of the Ritz eigenvalues with largest modulus.

In certain particular cases, when ξ_0 belongs to an S invariant subspace of small dimension d , the element $h_{d, d-1}$ vanishes automatically (if $d \leq n_A$ and assuming that numerical rounding errors are not important) and the Arnoldi iteration stops at $k = d$ and provides d exact eigenvalues of S for the invariant subspace. One can mention that there are more sophisticated variants of the Arnoldi method (Stewart, 2001) where one applies (implicit) modifications on the initial vector ξ_0 in order to force this vector to be in some small dimensional invariant subspace which results in such a vanishing coupling matrix element. These variants known as (implicitly) restarted Arnoldi methods allow to concentrate on certain regions on the complex plane to determine a few but very accurate eigenvalues in these regions. However, for the cases of Google matrices, where one is typically interested in the largest eigenvalues close to the unit circle, only the basic variant described above was used but choosing larger values of n_A as would have been possible with the restarted variants. The initial vector was typically chosen to be random or as the vector with unit entries.

Concerning the numerical resources the Arnoldi method requires ζN double precision registers to store the non-zero matrix elements of S , $n_A N$ registers to store the vectors ξ_k and $\text{const.} \times n_A^2$ registers to store h (and various copies of h). The computational time scales as $\zeta n_A N_d$ for the computation of $S\xi_k$, with $N_d n_A^2$ for the Gram-Schmidt orthogonalization procedure (which is typically dominant) and with $\text{const.} \times n_A^3$ for the diagonalization of h .

The details of the Arnoldi method are described in Refs. given above. This method has problems with degenerate or strongly clustered eigenvalues and therefore for typical examples of Google matrices it is applied to the core space block S_{cc} where the effects of the invariant subspaces, being responsible for most of the degeneracies, are exactly taken out according to the discussion of the previous subsection. In typical examples it is possible to find about $n_A \approx 640$ eigenvalues with largest $|\lambda|$ for the entire Twitter network with $N \approx 4.1 \times 10^7$ (see Sec. X) and about $n_A \approx 6000$ eigenvalues for Wikipedia networks with $N \approx 3.2 \times 10^6$ (see Sec. IX). For the two university networks of Cambridge and Oxford 2006 with $N \approx 2 \times 10^5$ it is possible to compute $n_A \approx 20000$ eigenvalues (see Sec. VIII). For the case of the Citation network of Physical Review (see Sec. XII) with $N \approx 4.6 \times 10^5$ it is even possible and necessary to use high precision computations (with up to 768 binary digits) to determine accurately the Arnoldi matrix h with $n_A \approx 2000$ (Frahm *et al.*, 2014b).

E. General properties of eigenvalues and eigenstates

According to the Perron-Frobenius theorem all eigenvalues λ_i of G are distributed inside the unitary circle $|\lambda| \leq 1$. It can be shown that at $\alpha < 1$ there is only one eigenvalue $\lambda_0 = 1$ and all other $|\lambda_i| \leq \alpha$ having a simple dependence on α : $\lambda_i \rightarrow \alpha \lambda_i$ (see e.g. (Langville and Meyer, 2006)). The right eigenvectors $\psi_i(j)$ are defined by the equation

$$\sum_{j'} G_{jj'} \psi_i(j') = \lambda_i \psi_i(j) . \quad (3)$$

Only the PageRank vector is affected by α while other eigenstates are independent of α due to their orthogonality to the left unit eigenvector at $\lambda = 1$. Left eigenvectors are orthonormal to right eigenvectors (Langville and Meyer, 2006).

It is useful to characterize the eigenvectors by their Inverse Participation Ratio (IPR) $\xi_i = (\sum_j |\psi_i(j)|^2)^2 / \sum_j |\psi_i(j)|^4$ which gives an effective number of nodes populated by an eigenvector ψ_i . This characteristic is broadly used for description of localized or delocalized eigenstates of electrons in a disordered potential with Anderson transition (see e.g. (Evers and Mirlin , 2008; Guhr *et al.*, 1998)). We discuss the specific properties of eigenvectors in next Secs.

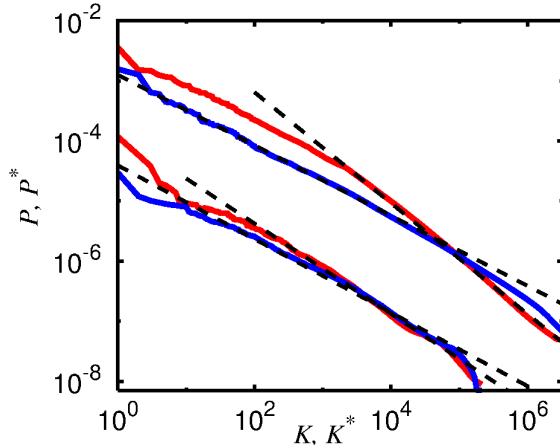


FIG. 5 (Color online) Dependence of probabilities of PageRank P (red/gray curve) and CheiRank P^* (blue/black curve) vectors on the corresponding rank indexes K and K^* for networks of Wikipedia Aug 2009 (top curves) and University of Cambridge (bottom curves, moved down by a factor 100). The straight dashed lines show the power law fits for PageRank and CheiRank with the slopes $\beta = 0.92, 0.58$ respectively, corresponding to $\beta = 1/(\mu_{\text{in,out}} - 1)$ for Wikipedia (see Fig. 2), and $\beta = 0.75, 0.61$ for Cambridge. After (Zhirov *et al.*, 2010) and (Frahm *et al.*, 2011).

IV. CHEIRANK VERSUS PAGERANK

It is established that ranking of network nodes based on PageRank order works reliably not only for WWW but also for other directed networks. As an example it is possible to quote the citation network of Physical Review (Radicchi *et al.*, 2009; Redner , 1998, 2005), Wikipedia network (Aragón *et al.*, 2012; Eom and Shepelyansky , 2013a; Skiena and Ward, 2014; Zhirov *et al.*, 2010) and even the network of world commercial trade (Ermann and Shepelyansky , 2011b). Here we describe the main properties of PageRank and CheiRank probabilities using a few real networks. More detailed presentation for concrete networks follows in next Secs.

A. Probability decay of PageRank and CheiRank

Wikipedia is a useful example of a scale-free network. An article quotes other Wikipedia articles that generates a network of directed links. For Wikipedia of English articles dated by Aug 2009 we have $N = 3282257$, $N_\ell = 71012307$ ((Zhirov *et al.*, 2010)). The dependencies of PageRank $P(K)$ and CheiRank $P^*(K^*)$ probabilities on indexes K and K^* are shown in Fig. 5. In a large range the decay can be satisfactory described by an algebraic law with an exponent β . The obtained β values are in a reasonable agreement with the expected relation $\beta = 1/(\mu_{\text{in,out}} - 1)$ with the exponents of distribution of links given above. However, the decay is algebraic only on a tail, showing certain nonlinear variations well visible for $P^*(K^*)$ at large values of P^* .

Similar data for network of University of Cambridge (2006) with $N = 212710$, $N_\ell = 2015265$ (Frahm *et al.*, 2011) are shown in the same Fig. 5. Here, the exponents β have different values with approximately the same statistical accuracy of β .

Thus we come to the same conclusion as (Meusel *et al.*, 2014): the probability decay of PageRank and CheiRank is only approximately algebraic, the relation between exponents β and μ also works only approximately.

B. Correlator between PageRank and CheiRank

Each network node i has both PageRank $K(i)$ and CheiRank $K(i)^*$ indexes so that it is interesting to know what is a correlation between the corresponding vectors of PageRank and CheiRank. It is convenient to characterized this by a correlator introduced in (Chepelianskii, 2010)

$$\kappa = N \sum_{i=1}^N P(K(i)) P^*(K^*(i)) - 1 . \quad (4)$$

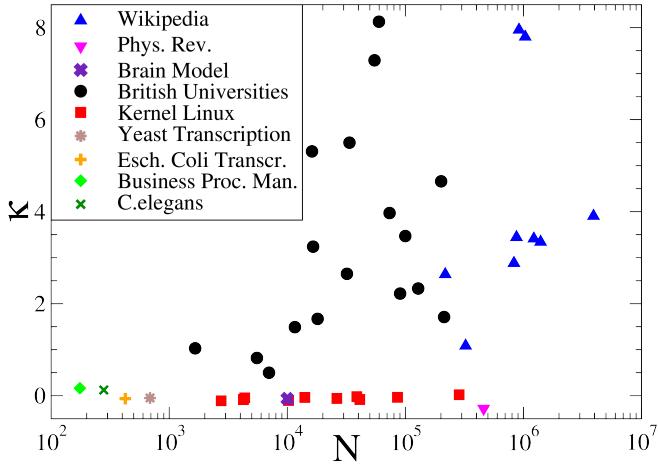


FIG. 6 (Color online) Correlator κ as a function of the number of nodes N for different networks: Wikipedia networks, Phys Rev network, 17 UK universities, 10 versions of Kernel Linux Kernel PCN, Escherichia Coli and Yeast Transcription Gene networks, Brain Model Network, C.elegans neural network and Business Process Management Network. After (Ermann *et al.*, 2012a) with additional data from (Abel and Shepelyansky , 2011), (Eom and Shepelyansky , 2013a), (Kandiah and Shepelyansky , 2014a), (Frahm *et al.*, 2014b).

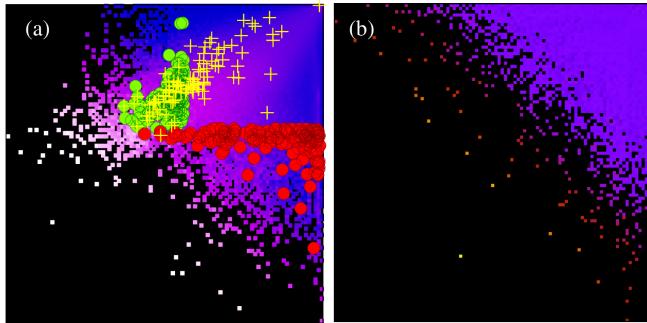


FIG. 7 (Color online) Density distribution of network nodes $W(K, K^*) = dN_i/dKdK^*$ shown on the plane of PageRank and CheiRank indexes in logscale ($\log_N K, \log_N K^*$) for all $1 \leq K, K^* \leq N$, density is computed over equidistant grid in plane ($\log_N K, \log_N K^*$) with 100×100 cells; color shows average value of W in each cell, the normalization condition is $\sum_{K, K^*} W(K, K^*) = 1$. Density $W(K, K^*)$ is shown by color with blue (dark gray) for minimum in (a),(b) and white (a) and yellow (white) (b) for maximum (black for zero). Panel (a): data for Wikipedia Aug (2009), $N = 3282257$, green/red (light gray/dark gray) points show top 100 persons from PageRank/CheiRank, yellow (white) pluses show top 100 persons from (Hart, 1992); after (Zhirov *et al.*, 2010). Panel (b): Density distribution $W(K, K^*) = dN_i/dKdK^*$ for Linux Kernel V2.4 network with $N = 85757$, after (Ermann *et al.*, 2012a).

Even if all the networks from Fig. 6 have similar algebraic decay of PageRank probability with K and similar $\beta \sim 1$ exponents we see that the correlations between

PageRank and CheiRank vectors are drastically different in these networks. Thus the networks of UK universities and 9 different language editions of Wikipedia have the correlator $\kappa \sim 1 - 8$ while all other networks have $\kappa \sim 0$. This means that there are significant differences hidden in the network architecture which are no visible from PageRank analysis. We will discuss the possible origins of such a difference for the above networks in next Secs.

C. PageRank-CheiRank plane

A more detailed characterization of correlations between PageRank and CheiRank vectors can be obtained from a distribution of network nodes on the two-dimensional plane (2D) of indexes (K, K^*) . Two examples for Wikipedia and Linux networks are shown in Fig. 7. A qualitative difference between two networks is obvious. For Wikipedia we have a maximum of density along the line $\ln K^* \approx 5 + (\ln K)/3$ that results from a strong correlation between PageRank and CheiRank with $\kappa = 4.08$. In contrast to that for the Linux network V2.4 we have a homogeneous density distribution of nodes along lines $\ln K^* = \ln K + \text{const}$ corresponding to uncorrelated probabilities $P(K)$ and $P^*(K^*)$ and even slightly negative value of $\kappa = -0.034$. We note that if for Wikipedia we generate nodes with independent probabilities distributions P and P^* , obtained from this network at the corresponding value of N , then we obtain a homogeneous node distribution in (K, K^*) plane (in $(\log K, \log K^*)$ plane it takes a triangular form, see Fig.4 at (Zhirov *et al.*, 2010)).

In Fig. 7(a) we also show the distribution of top 100 persons from PageRank and CheiRank compared with the top 100 persons from (Hart, 1992). There is a significant overlap between PageRank and Hart ranking of persons while CheiRank generates mainly another listing of people. We discuss the Wikipedia ranking of historical figures in Sec. IX.

D. 2DRank

PageRank and CheiRank indexes $K_i K_i^*$ order all network nodes according to a monotonous decrease of corresponding probabilities $P(K_i)$ and $P^*(K_i^*)$. While top K nodes are most popular or known in the network, top K^* nodes are most communicative nodes with many outgoing links. It is useful to consider an additional ranking K_2 , called 2DRank, which combines properties of both ranks K and K^* (Zhirov *et al.*, 2010).

The ranking list $K_2(i)$ is constructed by increasing $K \rightarrow K + 1$ and increasing 2DRank index $K_2(i)$ by one if a new entry is present in the list of first $K^* < K$ entries of CheiRank, then the one unit step is done in K^* and K_2 is increased by one if the new entry is present in the list of first $K < K^*$ entries of CheiRank. More

formally, 2DRank $K_2(i)$ gives the ordering of the sequence of sites, that appear inside the squares $[1, 1; K = k, K^* = k; \dots]$ when one runs progressively from $k = 1$ to N . In fact, at each step $k \rightarrow k + 1$ there are tree possibilities: (i) no new sites on two edges of square, (ii) only one site is on these two edges and it is added in the listing of $K_2(i)$ and (iii) two sites are on the edges and both are added in the listing $K_2(i)$, first with $K > K^*$ and second with $K < K^*$. For (iii) the choice of order of addition in the list $K_2(i)$ affects only some pairs of neighboring sites and does not change the main structure of ordering. An illustration example of 2DRank algorithm is given in Fig.7 at (Zhirov *et al.*, 2010). For Wikipedia 2DRanking of persons is discussed in Sec. IX.

E. Historical notes on spectral ranking

Starting from the work of Markov (Markov, 1906) many scientists contributed to the development of spectral ranking of Markov chains. Research of Perron (1907) and Frobenius (1912) led to the Perron-Frobenius theorem for square matrices with positive entries (see e.g. (Brin and Stuck, 2002)). Important steps have been done by researchers in psychology, sociology and mathematics including J.R.Seeley (1949), T.-H.Wei (1952), L.Katz (1953), C.H.Hubbell (1965). The detailed historical description of spectral ranking research is reviewed by (Franceschet, 2011) and (Vigna, 2013). In the WWW context, the Google matrix in the form (1), with regularization of dangling nodes and damping factor α , was introduced by (Brin and Page, 1998).

A PageRank vector of a Google matrix G^* with inverted directions of links has been considered by (Fogaras, 2003) and (Hrisitidis *et al.*, 2008), but no systematic statistical analysis of 2DRanking was presented there. An important step was done by (Chepelianskii, 2010) who analyzed $\lambda = 1$ eigenvectors of G for directed network and of G^* for network with inverted links. The comparative analysis of Linux Kernel network and WWW of University of Cambridge demonstrated a significant differences in correlator κ values on these networks and different functions of top nodes in K and K^* . The term CheiRank was coined in (Zhirov *et al.*, 2010) to have a clear distinction between eigenvectors of G and G^* . We note that top PageRank and CheiRank nodes have certain similarities with authorities and hubs appearing in the HITS algorithm (Kleinberg, 1999). However, the HITS is query dependent while the rank probabilities $P(K_i)$ and $P^*(K_i^*)$ classify all nodes of the network.

V. COMPLEX SPECTRUM AND FRACTAL WEYL LAW

The Weyl law (Weyl, 1912) gives a fundamental link between the properties of quantum eigenvalues in closed Hamiltonian systems, the Planck constant \hbar and the classical phase space volume. The number of states in this

case is determined by the phase volume of a system with dimension d . The case of Hermitian operators is now well understood both on mathematical and physical grounds (Dimassi and Sjöstrand, 1999; Landau and Lifshitz, 1989). Surprisingly, only recently it has been realized that the case of nonunitary operators describing open systems in the semiclassical limit has a number of new interesting properties and the concept of the fractal Weyl law (Sjöstrand, 1990; Zworski, 1999) has been introduced to describe the dependence of number of resonant Gamow eigenvalues (Gamow, 1928) on \hbar .

The Gamow eigenstates find important applications for decay of radioactive nuclei, quantum chemistry reactions, chaotic scattering and microlasers with chaotic resonators, open quantum maps (see (Gaspard, 1998, 2014; Shepelyansky, 2008) and Refs. therein). The spectrum of corresponding operators has a complex spectrum λ . The spread width $\gamma = -2 \ln |\lambda|$ of eigenvalues λ determines the life time of a corresponding eigenstate. The understanding of the spectral properties of related operators in the semiclassical limit represents an important challenge.

According to the fractal Weyl law (Lu *et al.*, 2003; Sjöstrand, 1990) the number of Gamow eigenvalues N_γ , which have escape rates γ in a finite band width $0 \leq \gamma \leq \gamma_b$, scales as

$$N_\gamma \propto \hbar^{-d/2} \propto N^{d/2} \quad (5)$$

where d is a fractal dimension of a classical strange repeller formed by classical orbits nonescaping in future and past times. In the context of eigenvalues λ of the Google matrix we have $\gamma = -2 \ln |\lambda|$. By numerical simulations it has been shown that the law (5) works for a scattering problem in 3-disk system (Lu *et al.*, 2003) and quantum chaos maps with absorption when the fractal dimension d is changed in a broad range $0 < d < 2$ (Erman and Shepelyansky, 2010b; Shepelyansky, 2008).

The fractal Weyl law (5) of open systems with a fractal dimension $d < 2$ leads to a striking consequence: only a relatively small fraction of eigenvalues $\mu_W \sim N_\gamma/N \propto \hbar^{(2-d)/2} \propto N^{(d-2)/2} \ll 1$ has finite values of $|\lambda|$ while almost all eigenstates of the matrix operator of size $N \propto 1/\hbar$ have $\lambda \rightarrow 0$. The eigenstates with finite $|\lambda| > 0$ are related to the classical fractal sets of orbits non-escaping neither in the future neither in the past. A fractal structure of these quantum fractal eigenstates has been investigated in (Shepelyansky, 2008). There it was conjectured that the eigenstates of a Google matrix with finite $|\lambda| > 0$ will select interesting specific communities of a network. We will see below that the fractal Weyl law can indeed be observed in certain directed networks and in particular we show in the next section that it naturally appears for Perron-Frobenius operators of dynamical systems and Ulam networks.

It is interesting to note that nontrivial complex spectra also naturally appear in systems of quantum chaos in presence of a contact with a measurement device (Bruzda *et al.*, 2010). The properties of complex spectra of small

size orthostochastic (unistochastic) matrices are analyzed in (Zyczkowski *et al.*, 2003). In such matrices the elements can be presented in a form $S_{ij} = O_{ij}^2$; ($S_{ij} = |U_{ij}|^2$) where O is an orthogonal matrix (U is a unitary matrix). We will see certain similarities of their spectra with the spectra of directed networks discussed in Sec. VIII.

Recent mathematical results for the fractal Weyl law are presented in (Nonnenmacher and Zworski , 2007; Nonnenmacher *et al.*, 2014).

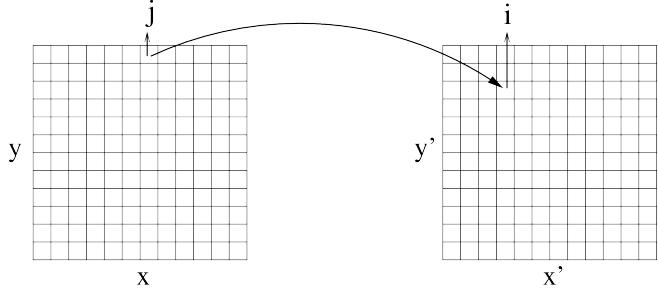


FIG. 8 Illustration of operation of the Ulam method: the phase space (x, y) is divided in $N = N_x \times N_y$ cells, N_c trajectories start from cell j and the number of trajectories N_{ij} arrived to a cell i from a cell j is collected after a map iteration. Then the matrix of Markov transitions is defined as $S_{ij} = N_{ij}/N_c$, by construction $\sum_{i=1}^N S_{ij} = 1$.

VI. ULAM NETWORKS

By construction the Google matrix belongs to the class of Perron-Frobenius operators which naturally appear in ergodic theory (Cornfeld *et al.*, 1982) and dynamical systems with Hamiltonian or dissipative dynamics (Brin and Stuck, 2002). In 1960 Ulam (Ulam, 1960) proposed a method, now known as the Ulam method, for a construction of finite size approximants for the Perron-Frobenius operators of dynamical maps. The method is based on discretization of the phase space and construction of a Markov chain based on probability transitions between such discrete cells given by the dynamics. Using as an example a simple chaotic map Ulam made a conjecture that the finite size approximation converges to the continuous limit when the cell size goes to zero. Indeed, it has been proven that for hyperbolic maps in one and higher dimensions the Ulam method converges to the spectrum of continuous system (Blank *et al.*, 2002; Li , 1976). The probability flows in dynamical systems have rich and non-trivial features of general importance, like simple and strange attractors with localized and delocalized dynamics governed by simple dynamical rules (Lichtenberg and Lieberman, 1992). Such objects are generic for nonlinear dissipative dynamics and hence can have relevance for actual WWW structure. The analysis of Ulam networks, generated by the Ulam method, allows to obtain a better intuition about the spectral properties of Google matrix. The term Ulam networks was introduced in (Shepelyan-

sky and Zhirov , 2010a).

A. Ulam method for dynamical maps

In Fig. 8 we show how the Ulam method works. The phase space of a dynamical map is divided in equal cells and a number of trajectories N_c is propagated by a map iteration. Thus a number of trajectories N_{ij} arrived from cell j to cell i is determined. Then the matrix of Markov transition is defined as $S_{ij} = N_{ij}/N_c$. By construction this matrix belongs to the class of Perron-Frobenius operators which includes the Google matrix.

The physical meaning of the coarse grain description by a finite number of cells is that it introduces in the system a noise of cell size amplitude. Due to that an exact time reversibility of dynamical equations of chaotic maps is destroyed due to exponential instability of chaotic dynamics. This time reversibility breaking is illustrated by an example of the Arnold cat map by (Ermann and Shepelyansky , 2012b). For the Arnold cat map on a long torus it is shown that the spectrum of the Ulam approximate of the Perron-Frobenius (UPFO) is composed of a large group of complex eigenvalues with $\gamma \sim 2h \approx 2$, and real eigenvalues with $|1 - \lambda| \ll 1$ corresponding to a statistical relaxation to the ergodic state at $\lambda = 1$ described by the Fokker-Planck equation (here h is the Kolmogorov-Sinai entropy of the map being here equal to the Lyapunov exponent, see e.g. (Chirikov , 1979)).

For fully chaotic maps the finite cell size, corresponding to added noise, does not significantly affect the dynamics and the discrete UPFO converges to the limiting case of continuous Perron-Frobenius operator (Blank *et al.*, 2002; Li , 1976). The Ulam method finds useful applications in studies of dynamics of molecular systems and coherent structures in dynamical flows (Froyland and Padberg , 2009). Additional Refs. can be found in (Frahm and Shepelyansky , 2010).

B. Chirikov standard map

However, for symplectic maps with a divided phase space, a noise present in the Ulam method significantly affects the original dynamics leading to a destruction of islands of stable motion and Kolmogorov-Arnold-Moser (KAM) curves. A famous example of such a map is the Chirikov standard map which describes the dynamics of many physical systems (Chirikov , 1979; Chirikov and Shepelyansky , 2008):

$$\bar{y} = \eta y + \frac{K_s}{2\pi} \sin(2\pi x) , \quad \bar{x} = x + \bar{y} \pmod{1} . \quad (6)$$

Here bars mark the variables after one map iteration and we consider the dynamics to be periodic on a torus so that $0 \leq x \leq 1$, $-1/2 \leq y \leq 1/2$; K_s is a dimensionless parameter of chaos. At $\eta = 1$ we have area-preserving symplectic map, considered in this SubSec., for $0 < \eta < 1$ we have a dissipative dynamics analyzed in next SubSec.

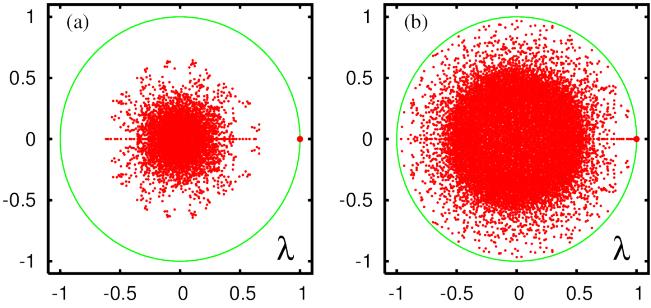


FIG. 9 (Color online) Complex spectrum of eigenvalues λ_j , shown by red/gray dots, for the UPFO of two variants of the Chirikov standard map (6); the unit circle $|\lambda| = 1$ is shown by a green (light gray) curve, the unit eigenvalue at $\lambda = 1$ is shown as larger red/gray dot. Panel (a) corresponds to the Chirikov standard map at dissipation $\eta = 0.3$ and $K_s = 7$; the phase space is covered by 110×110 cells and the UPFO is constructed by many trajectories with random initial conditions generating transitions from one cell into another (after (Erman and Shepelyansky , 2010b)). Panel (b) corresponds to the Chirikov standard map without dissipation at $K_s = 0.971635406$ with an UPFO constructed from a single trajectory of length 10^{12} in the chaotic domain and $280 \times 280/2$ cells to cover the phase space (after (Frahm and Shepelyansky , 2010)).

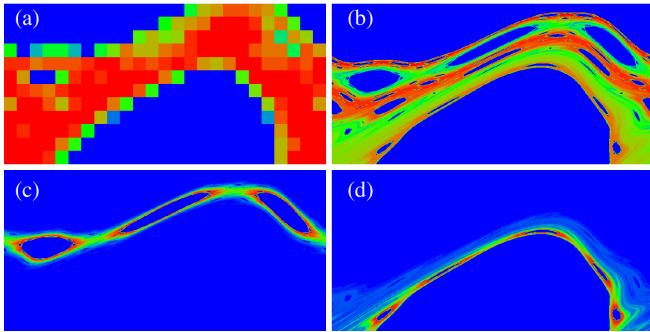


FIG. 10 (Color online) Density plots of absolute values of the eigenvectors of the UPFO obtained by the generalized Ulam method with a single trajectory of 10^{12} iterations of the Chirikov standard map at $K_s = 0.971635406$. The phase space is shown in the area $0 \leq x \leq 1$, $0 \leq y \leq 1/2$; the UPFO is obtained from $M \times M/2$ cells placed in this area. Panels represent: (a) eigenvector ψ_0 with eigenvalue $\lambda_0 = 1$; (b) eigenvector ψ_2 with real eigenvalue $\lambda_2 = 0.99878108$; (c) eigenvector ψ_6 with complex eigenvalue $\lambda_6 = -0.49699831 + i 0.86089756 \approx |\lambda_6| e^{i 2\pi/3}$; (d) eigenvector ψ_{13} with complex eigenvalue $\lambda_{13} = 0.30580631 + i 0.94120900 \approx |\lambda_{13}| e^{i 2\pi/5}$. Panel (a) corresponds to $M = 25$ while (b), (c) and (d) have $M = 800$. Color is proportional to amplitude with blue (black) for zero and red (gray) for maximal value. After (Frahm and Shepelyansky , 2010).

Since the finite cell size generates noise and destroys the KAM curves in the map (6) at $\eta = 1$, one should use the generalized Ulam method (Frahm and Shepelyansky

, 2010), where the transition probabilities N_{ij}/N_c are collected along one chaotic trajectory. In this construction a trajectory visits only those cells which belong to one connected chaotic component. Therefore the noise induced by the discretization of the phase space does not lead to a destruction of invariant curves, in contrast to the original Ulam method (Ulam, 1960), which uses all cells in the available phase space. Since a trajectory is generated by a continuous map it cannot penetrate inside the stability islands and on a physical level of rigor one can expect that, due to ergodicity of dynamics on one connected chaotic component, the UPFO constructed in such a way should converge to the Perron-Frobenius operator of the continuous map on a given subspace of chaotic component. The numerical confirmations of this convergence are presented in (Frahm and Shepelyansky , 2010).

We consider the map (6) at $K_s = 0.971635406$ when the golden KAM curve is critical. Due to the symmetry of the map with respect to $x \rightarrow 1-x$ and $y \rightarrow -y$ we can use only the upper part of the phase space with $y \geq 0$ dividing it in $M \times M/2$ cells. At that K_s we find that the number of cells visited by the trajectory in this half square scales as $N_d \approx C_d M^2/2$ with $C_d \approx 0.42$. This means that the chaotic component contains about 40% of the total area which is in good agreement with the known result of (Chirikov , 1979).

The spectrum of the UPFO matrix S for the phase space division by $280 \times 280/2$ cells is shown in Fig. 9(b). In a first approximation the spectrum λ of S is more or less homogeneously distributed in the polar angle φ defined as $\lambda_j = |\lambda_j| \exp(i\varphi_j)$. With the increase of matrix size N_d the two-dimensional density of states $\rho(\lambda)$ converges to a limiting distribution (Frahm and Shepelyansky , 2010). With the help of the Arnoldi method it is possible to compute a few thousands of eigenvalues with largest absolute values $|\lambda|$ for maximal $M = 1600$ with the total matrix size $N = N_d \approx 5.3 \times 10^5$.

The eigenstate at $\lambda = 1$ is homogeneously distributed over the chaotic component at $M = 25$ (Fig. 10) and higher M values (Frahm and Shepelyansky , 2010). This results from the ergodicity of motion and the fact that for symplectic maps the measure is proportional to the phase space area (Chirikov , 1979; Cornfeld *et al.*, 1982). Examples of other right eigenvalues of S at real and complex eigenvalues λ with $|\lambda| < 1$ are also shown in Fig. 10. For λ_2 the eigenstate corresponds to some diffusive mode with two nodal lines, while other two eigenstates are localized around certain resonant structures in phase space. This shows that eigenstates of the matrix G (and S) are related to specific communities of a network.

With the increase of number of cells $M^2/2$ there are eigenvalues which become more and more close to the unit eigenvalue. This is shown to be related to an algebraic statistics of Poincaré recurrences and long time sticking of trajectories in a vicinity of critical KAM curves. At the same time for symplectic maps the measure is proportional to area so that we have dimension $d = 2$ and hence we have a usual Weyl law with $N_\gamma \propto N$.

More details can be found at (Frahm and Shepelyansky , 2010, 2013).

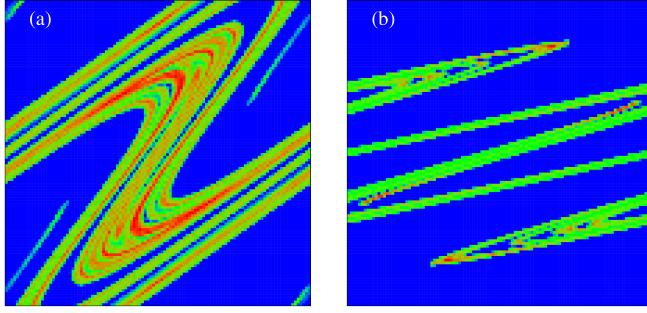


FIG. 11 (Color online) Phase space representation of eigenstates of the UFPO S for $N = 110 \times 110$ cells (color is proportional to absolute value $|\psi_i|$ with red/gray for maximum and blue/black for zero). Panel (a) shows an eigenstate with maximum eigenvalue $\lambda_1 = 0.756$ of the UFPO of map (6) with absorption at $K_s = 7$, $a = 2$, $\eta = 1$, the space region is $(-aK_s/4\pi \leq y \leq aK_s/4\pi, 0 \leq x \leq 1)$, the fractal dimension of the strange repeller set nonescaping in future is $d_e = 1 + d/2 = 1.769$. Panel (b) shows an eigenstate at $\lambda = 1$ of the UFPO of map (6) without absorption at $K_s = 7$, $\eta = 0.3$, the shown space region is $(-1/\pi \leq y \leq 1/\pi, 0 \leq x \leq 1)$ and the fractal dimension of the strange attractor is $d = 1.532$. After (Ermann and Shepelyansky , 2010b).

C. Dynamical maps with strange attractors

The fractal Weyl law (5) has initially been proposed for quantum systems with chaotic scattering. However, it is natural to assume that it should also work for Perron-Frobenius operators of dynamical systems. Indeed, the mathematical results for the Selberg zeta function indicated that the law (5) should remain valid for the UFPO (see Refs. at (Nonnenmacher *et al.*, 2014)). A detailed test of this conjecture (Ermann and Shepelyansky , 2010b) has been performed for the map (6) with dissipation at $0 < \eta < 1$, when at large K_s the dynamics converges to a strange attractor in the range $-2 < y < 2$, and for the nondissipative case $\eta = 1$ with absorption where all orbits leaving the interval $-aK_s/4\pi \leq y \leq aK_s/4\pi$ are absorbed after one iteration (in both cases there is no modulus in y).

An example of the spectrum of UPFO for the model with dissipation is shown in Fig. 9(a). We see that now, in contrast to the symplectic case of Fig. 9(b), the spectrum has a significant gap which separates the eigenvalue $\lambda = 1$ from the other eigenvalues with $|\lambda| < 0.7$. For the case with absorption the spectrum has a similar structure but now with $|\lambda| < 1$ for the leading eigenvalue λ since the total number of initial trajectories decreases with the number of map iterations due to absorption implying that for this case $\sum_i S_{ij} < 1$ with S being the UPFO.

It is established that the distribution of density of

states $dW/d\gamma$ (or $dW/d|\lambda|$) converges to a fixed distribution in the limit of large N or cell size going to zero (Ermann and Shepelyansky , 2010b) (see Fig.4 there). This demonstrates the validity of the Ulam conjecture for considered systems.

Examples of two eigenstates of the UFPO for these two models are shown in Fig. 11. The fractal structure of eigenstates is well visible. For the dissipative case without absorption we have eigenstates localized on the strange attractor. For the case with absorption eigenstates are located on a strange repeller corresponding to an invariant set of nonescaping orbits. The fractal dimension d of these classical invariant sets can be computed by the usual box-counting method for dynamical systems. It is important to note that for the case with absorption it is more natural to measure the dimension d_e of the set of orbits nonescaping in future. Due to the time reversal symmetry of the continuous map the dimension of the set of orbits nonescaping in the past is also d_e . Thus the phase space dimension 2 is composed of $2 = d_e + d_e - d$ and $d_e = 1 + d/2$ where d is the dimension of the invariant set of orbits nonescaping neither in the future neither in the past. For the case with dissipation without absorption all orbits drop on a strange attractor and we have the dimension of invariant set $d_e = d$.

D. Fractal Weyl law for Perron-Frobenius operators

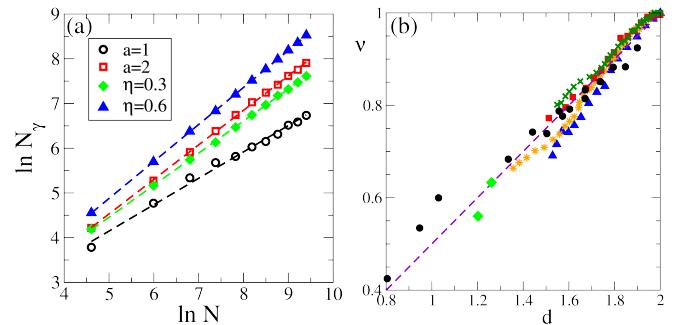


FIG. 12 (Color online) Panel (a) shows the dependence of the integrated number of states N_γ with decay rates $0 \leq \gamma \leq \gamma_b = 16$ on the size N of the UFPO matrix S for the map (6) at $K_s = 7$. The fits of numerical data, shown by dashed straight lines, give $\nu = 0.590, d_e = 1 + d/2 = 1.643$ (at $a = 1$); $\nu = 0.772, d_e = 1 + d/2 = 1.769$ (at $a = 2$); $\nu = 0.716, d = 1.532$ (at $\eta = 0.3$); $\nu = 0.827, d = 1.723$ (at $\eta = 0.6$). Panel (b) shows the fractal Weyl exponent ν as a function of fractal dimension d of the invariant fractal set for the map (6) with a strange attractor ($\eta < 1$) at $K_s = 15$ (green/gray crosses), $K_s = 12$ (red/gray squares), $K_s = 10$ (orange/gray stars), $K_s = 7$ (blue/black triangles); for a strange repeller ($\eta = 1$) at $K_s = 7$ (black points) and for a strange attractor for the Hénon map at standard parameters $a = 1.2, 1.4, b = 0.3$ (green diamonds). The straight dashed line shows the fractal Weyl law dependence $\nu = d/2$. After (Ermann and Shepelyansky , 2010b).

The direct verification of the validity of the fractal Weyl law (5) is presented in Fig. 12. The number of eigenvalues N_γ in a range with $0 \leq \gamma \leq \gamma_b$ ($\gamma = -2 \ln |\lambda|$) is numerically computed as a function of matrix size N . The fit of the dependence $N_\gamma(N)$, as shown in Fig. 12(a), allows to determine the exponent ν in the relation $N_\gamma \propto N^\nu$. The dependence of ν on the fractal dimension d , computed from the invariant fractal set by the box-counted method, is shown in Fig. 12(b). The numerical data are in good agreement with the theoretical fractal Weyl law dependence $\nu = d/2$. This law works for a variety of parameters for the system (6) with absorption and dissipation, and also for a strange attractor in the Hénon map ($\bar{x} = y + 1 - ax^2, \bar{y} = bx$). We attribute certain deviations, visible in Fig. 12 especially for $K_s = 7$, to the fact that at $K_s = 7$ there is a small island of stability at $\eta = 1$, which can produce certain influence on the dynamics.

The physical origin of the law (5) can be understood in a simple way: the number of states N_γ with finite values of γ is proportional to the number of cells $N_f \propto N^{d/2}$ on the fractal set of strange attractor. Indeed, the results for the overlap measure show that the eigenstates N_γ have a strong overlap with the steady state while the states with $\lambda \rightarrow 0$ have very small overlap. Thus almost all N states have eigenvalues $\lambda \rightarrow 0$ and only a small fraction of states on a strange attractor/repeller $N_\gamma \propto N_f \propto N^{d/2} \ll N$ has finite values of λ . We also checked that the participation ratio ξ of the eigenstate at $\lambda = 1$, grows as $\xi \sim N_f \propto N^{d/2}$ in agreement with the fractal Weyl law (Ermann and Shepelyansky , 2010b).

E. Intermittency maps

The properties of the Google matrix generated by one-dimensional intermittency maps are analyzed in (Ermann and Shepelyansky , 2010a). It is found that for such Ulam networks there are many eigenstates with eigenvalues $|\lambda|$ being very close to unity. The PageRank of such networks at $\alpha = 1$ is characterized by a power law decay with an exponent determined by the parameters of the map. It is interesting to note that usually for WWW the PageRank probability is proportional to a number of ingoing links distribution (see e.g. (Litvak *et al.*, 2008)). For the case of intermittency maps the decay of PageRank is independent of number of ingoing links. In addition, for α close to unity a decay of the PageRank has an exponent $\beta \approx 1$ but at smaller values $\alpha \leq 0.9$ the PageRank becomes completely delocalized. It is shown that the delocalization depends on the intermittency exponent of the map. This indicates that a rather dangerous phenomenon of PageRank delocalization can appear for certain directed networks. At the same time the one-dimensional intermittency map still generates a relatively simple structure of links with a typical number of links per node being close to unity. Such a case is probably not very typical for real networks. Therefore it is useful to analyze richer

Ulam networks with a larger number of links per node.

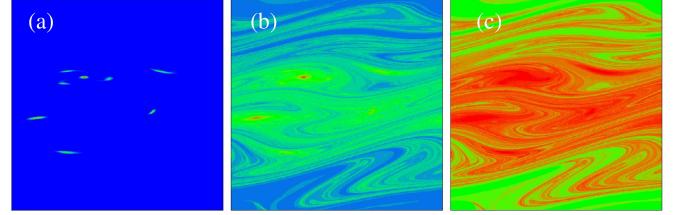


FIG. 13 (Color online) PageRank probability P_j for the Google matrix generated by the Chirikov typical map at $T = 10$, $k_s = 0.22$, $\eta = 0.99$ with $\alpha = 1$ (a), $\alpha = 0.95$ (b), and $\alpha = 0.85$ (c). The probability P_j is shown in the phase space region $0 \leq x < 2\pi; -\pi \leq y < \pi$ which is divided in $N = 3.6 \cdot 10^5$ cells; P_j is zero for blue/black and maximal for red/green. After (Shepelyansky and Zhirov , 2010a).

F. Chirikov typical map

With this aim we consider the Ulam networks generated by the Chirikov typical map with dissipation studied by (Shepelyansky and Zhirov , 2010a). The map introduced, by Chirikov in 1969 for description of continuous chaotic flows, has the form:

$$y_{t+1} = \eta y_t + k_s \sin(x_t + \theta_t), \quad x_{t+1} = x_t + y_{t+1}. \quad (7)$$

Here the dynamical variables x, y are taken at integer moments of time t . Also x has a meaning of phase variable and y is a conjugated momentum or action. The phases $\theta_t = \theta_{t+T}$ are T random phases periodically repeated along time t . We stress that their T values are chosen and fixed once and they are not changed during the dynamical evolution of x, y . We consider the map in the region of Fig. 13 ($0 \leq x < 2\pi, -\pi \leq y < \pi$) with the 2π -periodic boundary conditions. The parameter $0 < \eta < 1$ gives a global dissipation. The properties of the symplectic map at $\eta = 1$ have been studied in detail in (Frahm and Shepelyansky , 2009). The dynamics is globally chaotic for $k_s > k_c \approx 2.5/T^{3/2}$ and the Kolmogorov-Sinai entropy is $h \approx 0.29 k_s^{2/3}$ (more details about the Kolmogorov-Sinai entropy can be found in (Brin and Stuck, 2002; Chirikov , 1979; Cornfeld *et al.*, 1982)). A bifurcation diagram at $\eta < 1$ shows a series of transitions between fixed points, simple and strange attractors. Here we present results for $T = 10$, $k_s = 0.22$, $\eta = 0.99$ and a specific random set of θ_t given in (Shepelyansky and Zhirov , 2010a).

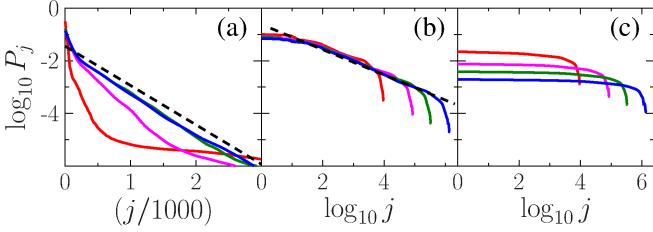


FIG. 14 (Color online) Dependence of PageRank probability P_j on PageRank index j for number of cells in the UFPO being $N = 10^4$, 9×10^4 , 3.6×10^5 and 1.44×10^6 (larger N have more dark and more long curves in (b), (c); in (a) this order of N is for curves from bottom to top (curves for $N = 3.6 \times 10^5$ and 1.44×10^6 practically coincide in this panel; for online version we note that the above order of N values corresponds to red, magenta, green, blue curves respectively). Dashed line in (a) shows an exponential Boltzmann decay (see text, line is shifted in j for clarity). The dashed straight line in (b) shows the fit $P_j \sim 1/j^\beta$ with $\beta = 0.48$. Other parameters, including the values of α , and panel order are as in Fig. 13. After (Shepelyansky and Zhirov , 2010a).

Due to exponential instability of motion one cell in the Ulam method gives transitions approximately to $k_{\text{cl}} \approx \exp(hT)$ other cells. According to this relation a large number of cells k_{cl} can be coupled at large T and h . For parameters of Fig. 13 one finds an approximate power law distribution of ingoing and outgoing links in the corresponding Ulam network with the exponents $\mu_{\text{in}} \approx \mu_{\text{out}} \approx 1.9$. The variation of the PageRank vector with the damping factor α is shown in Fig. 13 on the phase plane (x, y) . For $\alpha = 1$ the PageRank is concentrated in a vicinity of a simple attractor composed of several fixed points on the phase plane. Thus the dynamical attractors are the most popular nodes from the network view point. With a decrease of α down to 0.95, 0.85 values we find a stronger and stronger delocalization of PageRank over the whole phase space.

The delocalization with a decrease of α is also well seen in Fig. 14 where we show P_j dependence on PageRank index j with a monotonic decreasing probability P_j . At $\alpha = 1$ we have an exponential decay of P_j with j that corresponds to a Boltzmann type distribution where a noise produced by a finite cell size in the Ulam method is compensated by dissipation. For $\alpha = 0.95$ the random jumps of a network surfer, induced by the term $(1 - \alpha)/N$ in (1), produce an approximate power law decay of $P_j \propto 1/j^\beta$ with $\beta \approx 0.48$. For $\alpha = 0.85$ the PageRank probability is flat and completely delocalized over the whole phase space.

The analysis of the spectrum of S for the map (7) for the parameters of Fig. 14 shows the existence of eigenvalues being very close to $\lambda = 1$, however, there is no exact degeneracy as it is the case for UK universities which we will discuss below. The spectrum is characterized by the fractal Weyl law with the exponent $\nu \approx 0.85$. For eigenstates with $|\lambda| < 1$ the values of IPR ξ are less than 300 for a matrix size $N \approx 1.4 \times 10^4$ showing that eigenstates

are localized. However, for the PageRank the computations can be done with larger matrix sizes reaching a maximal value of $N = 6.4 \times 10^5$. The dependence of ξ on α shows that a delocalization transition of PageRank vector takes place for $\alpha < \alpha_c \approx 0.95$. Indeed, at $\alpha = 0.98$ we have $\xi \approx 30$ while at $\alpha \approx 0.8$ the IPR value of PageRank becomes comparable with the whole system size $\xi \approx 5 \times 10^5 \sim N = 6.4 \times 10^5$ (see Fig.9 at (Shepelyansky and Zhirov , 2010a)).

The example of Ulam networks considered here shows that a dangerous phenomenon of PageRank delocalization can take place under certain conditions. This delocalization may represent a serious danger for efficiency of search engines since for a delocalized flat PageRank the ranking of nodes becomes very sensitive to small perturbations and fluctuations.

VII. LINUX KERNEL NETWORKS

Modern software codes represent now complex large scale structures and analysis and optimization of their architecture become a challenge. An interesting approach to this problem, based on a directed network construction, has been proposed by (Chepelianskii, 2010). Here we present results obtained for such networks.

A. Ranking of software architecture

Following (Chepelianskii, 2010) we consider the Procedure Call Networks (PCN) for open source programs with emphasis on the code of Linux Kernel (Linux, 2010) written in the C programming language (Kernighan and Ritchie, 1978). In this language the code is structured as a sequence of procedures calling each other. Due to that feature the organization of a code can be naturally represented as a PCN, where each node represents a procedure and each directed link corresponds to a procedure call. For the Linux source code such a directed network is built by its lexical scanning with the identification of all the defined procedures. For each of them a list keeps track of the procedures calls inside their definition.

An example of the obtained network for a toy code with two procedures *start_kernel* and *printk* is shown in Fig. 15. The in/out-degrees of this model, noted as k and \bar{k} , are shown in Fig. 15. These numbers correspond to the number of out/in-going calls for each procedure. The obtained in/out-degree probability distributions $P_{\text{in}}(k)$, $P_{\text{out}}(\bar{k})$ are shown Fig. 15 for different Linux Kernel releases. These distributions are well described by power law dependencies $P_{\text{in}}(k) \propto 1/k^{\mu_{\text{in}}}$ and $P_{\text{out}}(\bar{k}) \propto 1/\bar{k}^{\mu_{\text{out}}}$ with $\mu_{\text{in}} = 2.0 \pm 0.02$, and $\mu_{\text{out}} = 3.0 \pm 0.1$. These values of exponents are close to those found for the WWW (Donato *et al.*, 2004; Pandurangan *et al.*, 2005). If only calls to distinct functions are counted in the outdegree distribution then the exponent drops to $\mu_{\text{out}} \approx 5$ whereas μ_{in} remains unchanged. It is important that the distribu-

tions for the different kernel releases remain stable even if the network size increases from $N = 2751$ for version V1.0 to $N = 285509$ for the latest version V2.6.32 taken into account in this study. This confirms the free-scale structure of software architecture of Linux Kernel network.

The probability distributions of PageRank and CheiRank vectors are also well described by power laws with exponents $\beta_{\text{in}} \approx 1$ and $\beta_{\text{out}} \approx 0.5$ being in good agreement with the usual relation $\beta = 1/(\mu - 1)$ (see Fig.2 in (Chepelianskii, 2010)). For V2.6.32 the top three procedures of PageRank at $\alpha = 0.85$ are *printk*, *memset*, *kfree* with probabilities 0.024, 0.012, 0.011 respectively, while at the top of CheiRank we have *start_kernel*, *btrfs_ioctl*, *menu_finalize* with respectively 0.000280, 0.000255, 0.000250. These procedures perform rather different tasks with *printk* reporting messages and *start_kernel* initializing the Kernel and managing the repartition of tasks. This gives an idea that both PageRank and CheiRank order can be useful to highlight en different aspects of directed and inverted flows on our network. Of course, in the context of WWW ingoing links related to PageRank are less vulnerable as compared to outgoing links related to CheiRank, which can be modified by a user rather easily. However, in other type of networks both directions of links appear in a natural manner and thus both vectors of PageRank and CheiRank play an important and useful role.

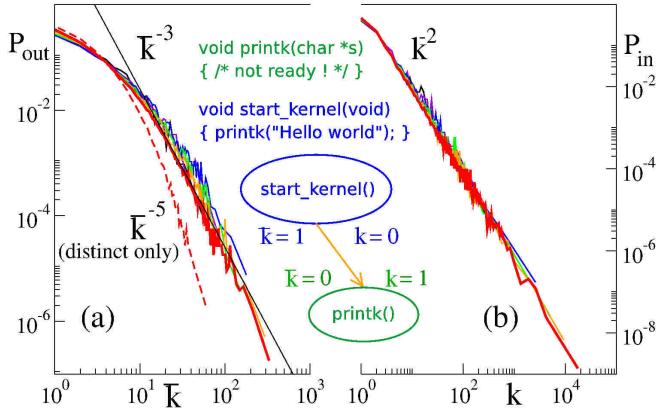


FIG. 15 (Color online) The diagram in the center represents the PCN of a toy kernel with two procedures written in C-programming language. The data on panels (a) and (b) show outdegree and indegree probability distributions $P_{\text{out}}(\bar{k})$ and $P_{\text{in}}(k)$ respectively. The colors correspond to different Kernel releases. The most recent version 2.6.32, with $N = 285509$ and an average 3.18 calls per procedure, is represented in red/gray. Older versions (2.4.37.6, 2.2.26, 2.0.40, 1.2.12, 1.0) with N respectively equal to (85756, 38766, 14079, 4358, 2751) follow the same behavior. The dashed curve in (a) shows the outdegree probability distribution if only calls to distinct destination procedures are kept. After (Chepelianskii, 2010).

For the Linux Kernel network the correlator κ (4) between PageRank and CheiRank vectors is close to zero

(see Fig. 6). This confirms the independence of two vectors. The density distribution of nodes of the Linux Kernel network, shown in Fig. 7(b), has a homogeneous distribution along $\ln K + \ln K^* = \text{const}$ lines demonstrating once more absence of correlations between $P(K_i)$ and $P^*(K_i^*)$. Indeed, such homogeneous distributions appear if nodes are generated randomly with factorized probabilities $P_i P_i^*$ (Chepelianskii, 2010; Zhirov *et al.*, 2010). Such a situation seems to be rather generic for software architecture. Indeed, other open software codes also have a small values of correlator, e.g. Open-Source software including Gimp 2.6.8 has $\kappa = -0.068$ at $N = 17540$ and X Windows server R7.1-1.1.0 has $\kappa = -0.027$ at $N = 14887$. In contrast to these software codes the Wikipedia networks have large values of κ and inhomogeneous distributions in (K, K^*) plane (see Figs. 6,7).

The physical reasons for absence of correlations between $P(K)$ and $P^*(K^*)$ have been explained in (Chepelianskii, 2010) on the basis of the concept of “separation of concerns” in software architecture (Dijkstra, 1982). It is argued that a good code should decrease the number of procedures that have high values of both PageRank and CheiRank since such procedures will play a critical role in error propagation since they are both popular and highly communicative at the same time. For example in the Linux Kernel, *do_fork*, that creates new processes, belongs to this class. Such critical procedures may introduce subtle errors because they entangle otherwise independent segments of code. The above observations suggest that the independence between popular procedures, which have high $P(K_i)$ and fulfill important but well defined tasks, and communicative procedures, which have high $P^*(K_i^*)$ and organize and assign tasks in the code, is an important ingredient of well structured software.

B. Fractal dimension of Linux Kernel Networks

The spectral properties the Linux Kernel network are analyzed in (Erman et al., 2011a). At large N the spectrum is obtained with the help of Arnoldi method from ARPACK library. This allows to find eigenvalues with $|\lambda| > 0.1$ for the maximal N at V2.6.32. An example of complex spectrum λ of G is shown in Fig. 16(a). There are clearly visible lines at real axis and polar angles $\varphi = \pi/2, 2\pi/3, 4\pi/3, 3\pi/2$. The later are related to certain cycles in procedure calls, e.g. an eigenstate at $\lambda_i = 0.85 \exp(i2\pi/3)$ is located only on 6 nodes. The spectrum of G^* has a similar structure.

The network size N grows with the version number of Linux Kernel corresponding to its evolution in time. We determine the total number of states N_λ with $0.1 < |\lambda| \leq 1$ and $0.25 < |\lambda| \leq 1$. The dependence of N_λ on N , shown in Fig. 16(b), clearly demonstrates the validity of the fractal Weyl law with the exponent $\nu \approx 0.63$ for G (we find $\nu^* \approx 0.65$ for G^*). We take the values of ν for $\lambda = 0.1$ where the number of eigenvalues N_λ gives a

better statistics. Within statistical errors the value of ν is not sensitive to the cutoff value at small λ . The matrix G^* has slightly higher values of ν . These results show that the PCN of Linux Kernel has a fractal dimension $d = 2\nu \approx 1.26$ for G and $d = 2\nu \approx 1.3$ for G^* .

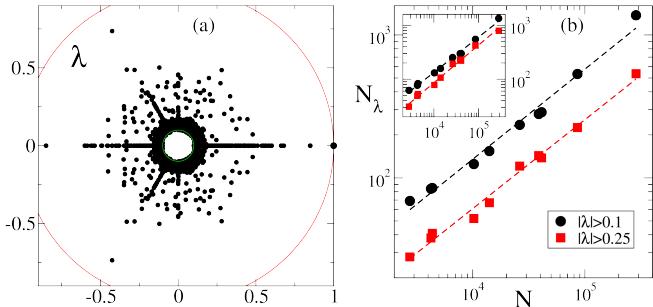


FIG. 16 (Color online) Panel (a) shows distribution of eigenvalues λ in the complex plane for the Google matrix G of the Linux Kernel version 2.6.32 with $N = 285509$ and $\alpha = 0.85$; the solid curves represent the unit circle and the lowest limit of computed eigenvalues. Panel (b) shows dependence of the integrated number of eigenvalues N_λ with $|\lambda| > 0.25$ (red/gray squares) and $|\lambda| > 0.1$ (black circles) as a function of the total number of processes N for versions of Linux Kernels. The values of N correspond (in increasing order) to Linux Kernel versions 1.0, 1.1, 1.2, 1.3, 2.0, 2.1, 2.2, 2.3, 2.4 and 2.6. The power law $N_\lambda \propto N^\nu$ has fitted values $\nu_{|\lambda|>0.25} = 0.622 \pm 0.010$ and $\nu_{|\lambda|>0.1} = 0.630 \pm 0.015$. Inset shows data for the Google matrix G^* with inverse link directions, the corresponding exponents are $\nu_{|\lambda|>0.25}^* = 0.696 \pm 0.010$ and $\nu_{|\lambda|>0.1}^* = 0.652 \pm 0.007$. After (Ermann *et al.*, 2011a).

To check that the fractal dimension of the PCN indeed has this value the dimension of the network is computed by another direct method known as the cluster growing method (see e.g. (Song *et al.*, 2005)). In this method the average mass or number of nodes $\langle M_c \rangle$ is computed as a function of the *network distance* l counted from an initial seed node with further averaging over all seed nodes. For a dimension d the mass $\langle M_c \rangle$ should grow as $\langle M_c \rangle \propto l^d$ that allows to determine the value of d for a given network. It should be noted that the above method should be generalized for the case of directed networks. For that the network distance l is computed following only outgoing links. The average of $\langle M_c(l) \rangle$ is done over all nodes. Due to global averaging the method gives the same result for the matrix with inverted link direction (indeed, the total number of outgoing links is equal to the number of ingoing links). However, as established in (Ermann *et al.*, 2011a), the fractal dimension obtained by this generalized method is very different from the case of converted undirected network, when each directed link is replaced by an undirected one. The average dimension obtained with this method for PCN is $d = 1.4$ even if a certain 20% increase of d appears for the latest Linux versions V2.6. We attribute this deviation for the version V2.6 to the well known fact that significant rearrangements in the Linux Kernel have been done after version V2.4

(Linux, 2010).

Thus in view of the above restrictions we consider that there is a rather good agreement of the fractal dimension obtained from the fractal Weyl law with $d \approx 1.3$ and the value obtained with the cluster growing method which gives an average $d \approx 1.4$. The fact that d is approximately the same for all versions up to V2.4 means that the Linux Kernel is characterized by a self-similar fractal growth in time. The closeness of d to unity signifies that procedure calls are almost linearly ordered that corresponds to a good code organization. Of course, the fractal Weyl law gives the dimension d obtained during time evolution of the network. This dimension is not necessary the same as for a given version of the network of fixed size. However, one can expect that the growth goes in a self-similar way (Dorogovtsev *et al.*, 2008) and that the static dimension is close to the dimension value emerging during the time evolution. This can be viewed as a some kind of ergodicity conjecture. Our data show that this conjecture works with a good accuracy up to the Linux Kernel V.2.6.

Thus the results obtained in (Ermann *et al.*, 2011a) and described here confirm the validity of the fractal Weyl law for the Linux Kernel network with the exponent $\nu \approx 0.65$ and the fractal dimension $d \approx 1.3$. It is important to note that the fractal Weyl exponent ν is not sensitive to the exponent β characterizing the decay of the PageRank. Indeed, the exponent β remains practically the same for the WWW (Donato *et al.*, 2004) and the PCN of Linux Kernel (Chepelianskii, 2010) while the values of fractal dimension are different with $d \approx 4$ for WWW and $d \approx 1.3$ for PCN (see (Ermann *et al.*, 2011a) and Refs. therein).

The analysis of the eigenstates of G and G^* shows that their IPR values remain small ($\xi < 70$) compared to the matrix size $N \approx 2.8 \times 10^5$ showing that they are well localized on certain selected nodes.

VIII. WWW NETWORKS OF UK UNIVERSITIES

The WWW networks of certain UK universities for years between 2002 and 2006 are publicly available at (UK universities, 2011). Due to their modest size, these networks are well suitable for a detail study of PageRank, CheiRank, complex eigenvalue spectra and eigenvectors (Frahm *et al.*, 2011).

A. Cambridge and Oxford University networks

We start our analysis of WWW university networks from those of Cambridge and Oxford 2006. For example, in Fig. 5 we show the dependence of PageRank (CheiRank) probabilities $P(P^*)$ on rank index K (K^*) for the WWW of Cambridge 2006 at $\alpha = 0.85$. The decay is satisfactory described by a power law with the exponent $\beta = 0.75$ ($\beta = 0.61$).

The complex eigenvalue spectrum and the invariant subspace structure (see section III.C) have been studied in great detail for the cases of Cambridge 2006 and Oxford 2006. For Cambridge 2006 (Oxford 2006) the network size is $N = 212710$ (200823) and the number of links is $N_\ell = 2015265$ (1831542). There are $n_{\text{inv}} = 1543$ (1889) invariant subspaces, with maximal dimension $d_{\text{max}} = 4656$ (1545), together they contain $N_s = 48239$ (30579) subspace nodes leading to 3508 (3275) eigenvalues (of the matrix S) with $|\lambda_j| = 1$ of which $n_1 = 1832$ (2360) are at $\lambda_j = 1$ (about 1% of N). The last number n_1 is larger than the number of invariant subspaces n_{inv} since each of the subspaces has at least one unit eigenvalue because each subspace is described by a full representation matrix of the Perron-Frobenius type. To determine the complex eigenvalue spectrum one can apply exact diagonalization on each subspace and the Arnoldi method on the remaining core space.

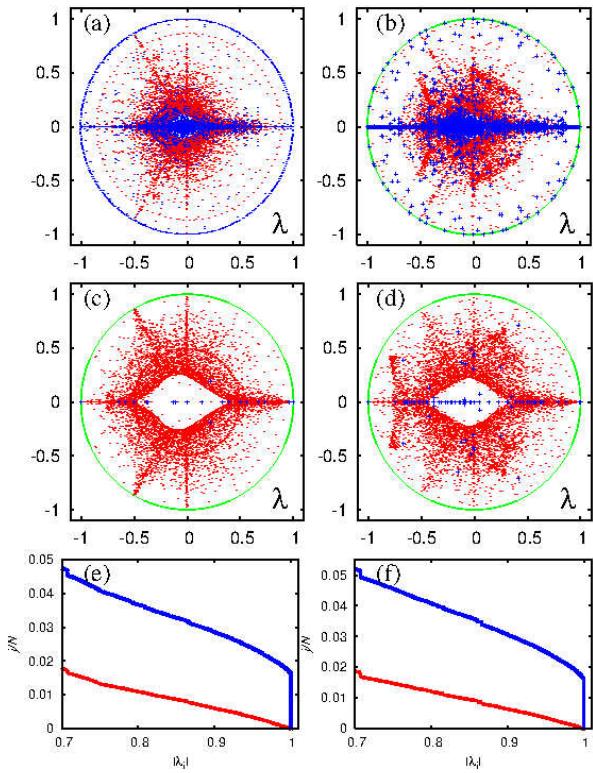


FIG. 17 (Color online) Panels (a) and (b) show the complex eigenvalue spectrum λ of matrix S for the University of Cambridge 2006 and Oxford 2006 respectively. The spectrum λ of matrix S^* for Cambridge 2006 and Oxford 2006 are shown in panels (c) and (d). Eigenvalues λ of the core space are shown by red/gray points, eigenvalues of isolated subspaces are shown by blue/black points and the green/gray curve (when shown) is the unit circle. Panels (e) and (f) show the fraction j/N of eigenvalues with $|\lambda| > |\lambda_j|$ for the core space eigenvalues (red/gray bottom curve) and all eigenvalues (blue/black top curve) from top row data for Cambridge 2006 and Oxford 2006. After (Frahm *et al.*, 2011).

The spectra of all subspace eigenvalues and $n_A =$

20000 core space eigenvalues of the matrices S and S^* are shown in Fig. 17. Even if the decay of PageRank and CheiRank probabilities with rank index is rather similar for both universities (see Fig.1 in (Frahm *et al.*, 2011)) the spectra of two networks are very different. Thus the spectrum contains much more detailed information about the network features compared to the rank vectors.

At the same time the spectra of two universities have certain similar features. Indeed, one can identify cross and triple-star structures. These structures are very similar to those seen in the spectra of random orthostochastic matrices of small size $N = 3, 4$ shown in Fig. 18 from (Zyczkowski *et al.*, 2003) (spectra of unistochastic matrices have a similar structure). The spectrum borders, determined analytically in (Zyczkowski *et al.*, 2003) for these N values, are also shown. The similarity is more visible for the spectrum of S^* case ((c) and (d) of Fig. 17). We attribute this to a larger randomness in outgoing links which have more fluctuations compared to ingoing links, as discussed in (Eom *et al.*, 2013b). The similarity of spectra of Fig. 17 with those of random matrices in Fig. 18 indicates that there are dominant triple and quadruple structures of nodes present in the University networks which are relatively weakly connected to other nodes.

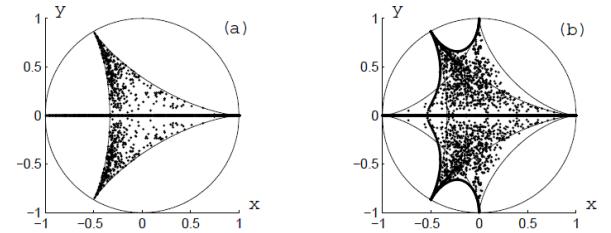


FIG. 18 Spectra λ of 800 random orthostochastic matrices of size $N = 3$ (a) and $N = 4$ (b) ($\text{Re}\lambda = x, \text{Im}\lambda = y$). Thin lines denote 3- and 4-hypocycloids, while the thick lines represent the 3-4 interpolation arc. After (Zyczkowski *et al.*, 2003).

The core space submatrix S_{cc} of Eq. (2) does not obey to the column sum normalization due to non-vanishing elements in the block S_{sc} which allow for a small but finite escape probability from core space to subspace nodes. Therefore the maximum eigenvalue of the core space (of the matrix S_{cc}) is below unity. For Cambridge 2006 (Oxford 2006) it is given by $\lambda_1^{(\text{core})} = 0.999874353718$ (0.999982435081) with a quite clear gap $1 - \lambda_1^{(\text{core})} \sim 10^{-4}$ ($\sim 10^{-5}$).

B. Universal emergence of PageRank

For $\alpha = 1$ the leading eigenvalue $\lambda = 1$ is highly degenerate due to the subspace structure. This degeneracy is lifted for $\alpha < 1$ with a unique eigenvector, the PageRank, for the leading eigenvalue. The question arises how the PageRank emerges if $1 - \alpha \ll 1$. Following (Frahm

et al., 2011), an answer is obtained from a formal matrix expression:

$$P = (1 - \alpha) (I - \alpha S)^{-1} e/N, \quad (8)$$

where the vector e has unit entries on each node and I is the unit matrix. Then, assuming that S is diagonalizable (with no nontrivial Jordan blocks) we can use the expansion:

$$P = \sum_{\lambda_j=1} c_j \psi_j + \sum_{\lambda_j \neq 1} \frac{1 - \alpha}{(1 - \alpha) + \alpha(1 - \lambda_j)} c_j \psi_j. \quad (9)$$

where ψ_j are the eigenvectors of S and c_j coefficients determined by the expansion $e/N = \sum_j c_j \psi_j$. Thus Eq. (9) indicates that in the limit $\alpha \rightarrow 1$ the PageRank converges to a particular linear combination of the eigenvectors with $\lambda_j = 1$, which are all localized in one of the subspaces. For a finite but very small value of $1 - \alpha \ll 1 - \lambda_1^{(\text{core})}$ the corrections for the contributions of the core space nodes are $\sim (1 - \alpha)/(1 - \lambda_1^{(\text{core})})$. This behavior is indeed confirmed by Fig. 19 (a) showing the evolution of the PageRank for different values of $1 - \alpha$ for the case of Cambridge 2006 and using a particular method, based on an alternate combination of the power iteration method and the Arnoldi method (Frahm et al., 2011), to determine numerically the PageRank for very small values of $1 - \alpha \sim 10^{-8}$.

However, for certain of the university networks the core space gap $1 - \lambda_1^{(\text{core})}$ is particularly small, for example $1 - \lambda_1^{(\text{core})} \sim 10^{-17}$, such that in standard double precision arithmetic the Arnoldi method, applied on the matrix S_{cc} , does not allow to determine this small gap. For these particular cases it is possible to determine rather accurately the core space gap and the corresponding eigenvector by another numerical approach called “projected power method” (Frahm et al., 2011). These eigenvectors, shown in Fig. 19 (b), are strongly localized on a modest number of nodes $\sim 10^2$ and with very small but non-vanishing values on the other nodes. Technically these vectors extend to the whole core space but practically they define small quasi-subspaces (in the core space domain) where the escape probability is extremely small (Frahm et al., 2011) and in the range $1 - \alpha \sim 10^{-8}$ they still contribute to the PageRank according to Eq. (9).

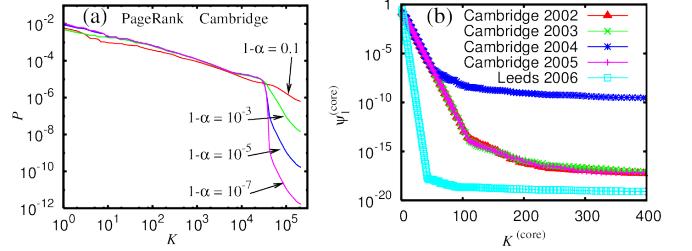


FIG. 19 (Color online) (a) PageRank $P(K)$ of Cambridge 2006 for $1 - \alpha = 0.1, 10^{-3}, 10^{-5}, 10^{-7}$. (b) First core space eigenvector $\psi_1^{(\text{core})}$ versus its rank index $K^{(\text{core})}$ for the UK university networks with a small core space gap $1 - \lambda_1^{(\text{core})} < 10^{-8}$. After (Frahm et al., 2011).

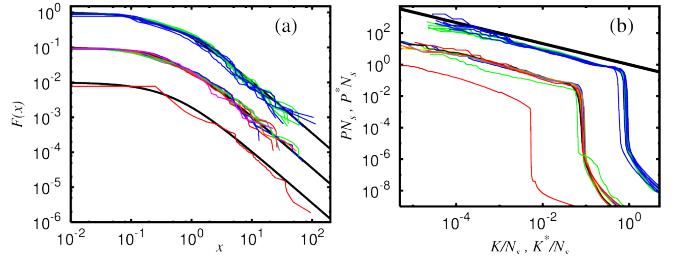


FIG. 20 (Color online) (a) Fraction of invariant subspaces F with dimensions larger than d as a function of the rescaled variable $x = d/\langle d \rangle$. Upper curves correspond to Cambridge (green/gray) and Oxford (blue/black) for years 2002 to 2006 and middle curves (shifted down by a factor of 10) correspond to the university networks of Glasgow, Cambridge, Oxford, Edinburgh, UCL, Manchester, Leeds, Bristol and Birkbeck for year 2006 with $\langle d \rangle$ between 14 and 31. Lower curve (shifted down by a factor of 100) corresponds to the matrix S^* of Wikipedia with $\langle d \rangle = 4$. The thick black line is $F(x) = (1 + 2x)^{-1.5}$. (b) Rescaled PageRank $P N_s$ versus rescaled rank index K/N_s for $1 - \alpha = 10^{-8}$ and $3974 \leq N_s \leq 48239$ for the same university networks as in (a) (upper and middle curves, the latter shifted down and left by a factor of 10). The lower curve (shifted down and left by a factor of 100) shows the rescaled CheiRank of Wikipedia $P^* N_s$ versus K^*/N_s with $N_s = 21198$. The thick black line corresponds to a power law with exponent $-2/3$. After (Frahm et al., 2011).

In Fig. 20(b) we show that for several of the university networks the PageRank at $1 - \alpha = 10^{-8}$ has actually a universal form when using the rescaled variables $P N_s$ versus K/N_s with a power law behavior close to $P \propto K^{-2/3}$ for $K/N_s < 1$. The rescaled data of Fig. 20 (a) show that the fraction of subspaces with dimensions larger than d is well described by the power law $F(x) \approx (1 + 2x)^{-1.5}$ with the dimensionless variable $x = d/\langle d \rangle$ where $\langle d \rangle$ is an average subspace dimension computed for WWW of a given university. The tables of all considered UK universities with the parameters of their WWW are given in (Frahm et al., 2011). We note that the CheiRank of S^* of Wikipedia 2009 also approximately follows the above universal distributions. How-

ever, for S matrix of Wikipedia the number of subspaces is small and statistical analysis cannot be performed for this case.

The origin of the universal distribution $F(x)$ still remains a puzzle. Possible links with a percolation on directed networks (see e.g. (Dorogovtsev *et al.*, 2008)) are still to be elucidated. It also remains unclear how stable this distribution really is. It works well for UK university networks 2002-2006. However, for the Twitter network (Frahm and Shepelyansky , 2012b) such a distribution becomes rather approximate. Also for the network of Cambridge in 2011, analyzed in (Ermann *et al.*, 2012a, 2013b) with $N \approx 8.9 \times 10^5$, $N_\ell \approx 1.5 \times 10^7$, the number of subspaces is significantly reduced and a statistical analysis of their size distribution becomes not relevant. It is possible that an increase of number of links per node N_ℓ/N from a typical value of 10 for UK universities to 35 for Twitter affects this distribution. For Cambridge 2011 the network entered in a regime when many links are generated by robots that apparently leads to a change of its statistical properties.

C. Two-dimensional ranking for University networks

Two-dimensional ranking of network nodes provides a new characterization of directed networks. Here we consider a density distribution of nodes (see Sec. IV.C) in the PageRank-CheiRank plane for examples of two WWW networks of Cambridge 2006 and ENS Paris 2011 shown in Fig. 21 from (Ermann *et al.*, 2012a).

The density distribution for Cambridge 2006 clearly shows that nodes with high PageRank have low CheiRank that corresponds to zero density at low K, K^* values. At large K, K^* values there is a maximum line of density which is located not very far from the diagonal $K \approx K^*$. The presence of correlations between $P(K_i)$ and $P^*(K_i^*)$ leads to a probability distribution with one main maximum along a diagonal at $\ln K + \ln K^* = \text{const}$. This is similar to the properties of the density distribution for the Wikipedia network shown in Fig. 7(a).

The 2DRanking might give new possibilities for information retrieval from large databases which are growing rapidly with time. Indeed, for example the size of the Cambridge network increased by a factor 4 from 2006 to 2011. At present, web robots start automatically to generate new web pages. These features can be responsible for the appearance of gaps in the density distribution in (K, K^*) plane at large $K, K^* \sim N$ values visible for large scale university networks such as ENS Paris in 2011 (see Fig. 21). Such an automatic generation of links can change the scale-free properties of networks. Indeed, for ENS Paris a large step in the PageRank distribution appears (Ermann *et al.*, 2012a) possibly indicating a delocalization transition tendency of the PageRank that can destroy the efficiency of information retrieval from the WWW.

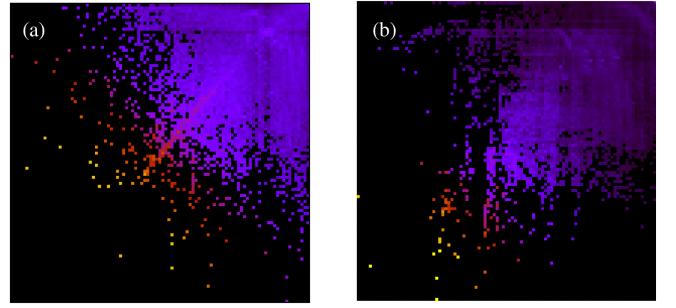


FIG. 21 (Color online) Density distribution $W(K, K^*) = dN_i/dKdK^*$ for networks of Universities in the plane of PageRank K and CheiRank K^* indexes in log-scale ($\log_N K, \log_N K^*$). The density is shown for 100×100 equidistant grid in $\log_N K, \log_N K^* \in [0, 1]$, the density is averaged over all nodes inside each cell of the grid, the normalization condition is $\sum_{K, K^*} W(K, K^*) = 1$. Color varies from black for zero to yellow/gray for maximum density value W_M with a saturation value of $W_s^{1/4} = 0.5W_M^{1/4}$ so that the same color is fixed for $0.5W_M^{1/4} \leq W^{1/4} \leq W_M^{1/4}$ to show in a better way low densities. The panels show networks of University of Cambridge 2006 with $N = 212710$ (a) and ENS Paris 2011 for crawling level 7 with $N = 1820015$ (b). After (Ermann *et al.*, 2012a).

IX. WIKIPEDIA NETWORKS

The free online encyclopedia Wikipedia is a huge repository of human knowledge. Its size is growing permanently accumulating enormous amount of information and becoming a modern version of *Library of Babel*, described by Jorge Luis Borges (Borges, 1962). The hyperlink citations between Wikipedia articles provides an important example of directed networks evolving in time for many different languages. In particular, the English edition of August 2009 has been studied in detail (Ermann *et al.*, 2012a, 2013b; Zhirov *et al.*, 2010). The effects of time evolution (Eom *et al.* , 2013b) and entanglement of cultures in multilingual Wikipedia editions have been investigated in (Aragón *et al.*, 2012; Eom and Shepelyansky , 2013a; Eom *et al.* , 2014).

A. Two-dimensional ranking of Wikipedia articles

The statistical distribution of links in Wikipedia networks has been found to follow a power law with the exponents $\mu_{\text{in}}, \mu_{\text{out}}$ (see e.g. (Capocci *et al.*, 2006; Muchnik *et al.*, 2007; Zhirov *et al.*, 2010; Zlatic *et al.*, 2006)). The probabilities of PageRank and CheiRank are shown in Fig. 5. They are satisfactorily described by a power law decay with exponents $\beta_{PR,CR} = 1/(\mu_{\text{in,out}} - 1)$ (Zhirov *et al.*, 2010).

The density distribution of articles over PageRank-CheiRank plane ($\log_N K, \log_N K^*$) is shown in Fig. 7(a) for English Wikipedia Aug 2009. We stress that the den-

sity is very different from those generated by the product of independent probabilities of P and P^* given in Fig. 5. In the latter case we obtain a density homogeneous along lines $\ln K^* = -\ln K + \text{const}$ being rather similar to the distribution for Linux network also shown in Fig. 7. This result is in good agreement with a fact that the correlator κ between PageRank and CheiRank vectors is rather large for Wikipedia $\kappa = 4.08$ while it is close to zero for Linux network $\kappa \approx -0.05$.

The difference between PageRank and CheiRank is clearly seen from the names of articles with highest ranks (ranks of all articles are given in (Zhirov *et al.*, 2010)). At the top of PageRank we have 1. *United States*, 2. *United Kingdom*, 3. *France* while for CheiRank we find 1. *Portal:Contents/Outline of knowledge/Geography and places*, 2. *List of state leaders by year*, 3. *Portal:Contents/Index/Geography and places*. Clearly PageRank selects first articles on a broadly known subject with a large number of ingoing links while CheiRank selects first highly communicative articles with many outgoing links. The 2DRank combines these two characteristics of information flow on directed network. At the top of 2DRank K_2 we find 1. *India*, 2. *Singapore*, 3. *Pakistan*. Thus, these articles are most known/popular and most communicative at the same time.

The top 100 articles in K, K_2, K^* are determined for several categories including countries, universities, people, physicists. It is shown in (Zhirov *et al.*, 2010) that PageRank recovers about 80% of top 100 countries from SJR data base (SJR, 2007), about 75% of top 100 universities of Shanghai university ranking (Shanghai ranking, 2010), and, among physicists, about 50% of top 100 Nobel winners in physics. This overlap is lower for 2DRank and even lower for CheiRank. However, as we will see below in more detail, 2DRank and CheiRank highlight other properties being complementary to PageRank.

Let us give an example of top three physicists among those of 754 registered in Wikipedia in 2010: 1. *Aristotle*, 2. *Albert Einstein*, 3. *Isaac Newton* from PageRank; 1. *Albert Einstein*, 2. *Nikola Tesla*, 3. *Benjamin Franklin* from 2DRank; 1. *Hubert Reeves*, 2. *Shen Kuo*, 3. *Stephen Hawking* from CheiRank. It is clear that PageRank gives most known, 2DRank gives most known and active in other areas, CheiRank gives those who are known and contribute to popularization of science. Indeed, e.g. *Hubert Reeves* and *Stephen Hawking* are very well known for their popularization of physics that increases their communicative power and place them at the top of CheiRank. *Shen Kuo* obtained recognized results in an enormous variety of fields of science that leads to the second top position in CheiRank even if his activity was about thousand years ago.

According to Wikipedia ranking the top universities are 1. *Harvard University*, 2. *University of Oxford*, 3. *University of Cambridge* in PageRank; 1. *Columbia University*, 2. *University of Florida*, 3. *Florida State University* in 2DRank and CheiRank. CheiRank and 2DRank

highlight connectivity degree of universities that leads to appearance of significant number of arts, religious and military specialized colleges (12% and 13% respectively for CheiRank and 2DRank) while PageRank has only 1% of them. CheiRank and 2DRank introduce also a larger number of relatively small universities who are keeping links to their alumni in a significantly better way that gives an increase of their ranks. It is established (Eom *et al.*, 2013b) that top 10 PageRank universities from English Wikipedia in years 2003, 2005, 2007, 2009, 2011 recover correspondingly 9, 9, 8, 7, 7 from top 10 of (Shanghai ranking, 2010).

The time evolution of probability distributions of PageRank, CheiRank and two-dimensional ranking is analyzed in (Eom *et al.*, 2013b) showing that they become stabilized for the period 2007-2011.

On the basis of these results we can conclude that the above algorithms provide correct and important ranking of huge information and knowledge accumulated at Wikipedia. It is interesting that even Dow-Jones companies are ranked via Wikipedia networks in a good manner (Zhirov *et al.*, 2010). We discuss ranking of top people of Wikipedia a bit later.

B. Spectral properties of Wikipedia network

The complex spectrum of eigenvalues of G for English Wikipedia network of Aug 2009 is shown in Fig. 22. As for university networks, the spectrum also has some invariant subspaces resulting in degeneracies of the leading eigenvalue $\lambda = 1$ of S (or S^*). However, due to the stronger connectivity of the Wikipedia network these subspaces are significantly smaller compared to university networks (Eom *et al.*, 2013b; Ermann *et al.*, 2013b). For example of Aug 2009 edition in Fig. 22 there are 255 invariant subspaces (of the matrix S) covering 515 nodes with 255 unit eigenvalues $\lambda_j = 1$ and 381 eigenvalues on the complex unit circle with $|\lambda_j| = 1$. For the matrix S^* of Wikipedia there are 5355 invariant subspaces with 21198 nodes, 5365 unit eigenvalues and 8968 eigenvalues on the unit circle (Ermann *et al.*, 2013b). The complex spectra of all subspace eigenvalues and the first $n_A = 6000$ core space eigenvalues of S and S^* are shown in Fig. 22. As in the university cases, in the spectrum we can identify cross and triple-star structures similar to those of orthostochastic matrices shown in Fig. 18. However, for Wikipedia (especially for S) the largest complex eigenvalues outside the real axis are more far away from the unit circle. For S of Wikipedia the two largest core space eigenvalues are $\lambda_1^{(\text{core})} = 0.999987$ and $\lambda_2^{(\text{core})} = 0.977237$ indicating that the core space gap $|1 - \lambda_1^{(\text{core})}| \sim 10^{-5}$ is much smaller than the secondary gap $|\lambda_1^{(\text{core})} - \lambda_2^{(\text{core})}| \sim 10^{-2}$. As a consequence the PageRank of Wikipedia (at $\alpha = 0.85$) is strongly influenced by the leading core space eigenvector and actually both vectors select the same 5 top nodes.

The time evolution of spectra of G and G^* for English Wikipedia is studied in (Eom *et al.*, 2013b). It is shown that the spectral structure remains stable for years 2007 - 2011.

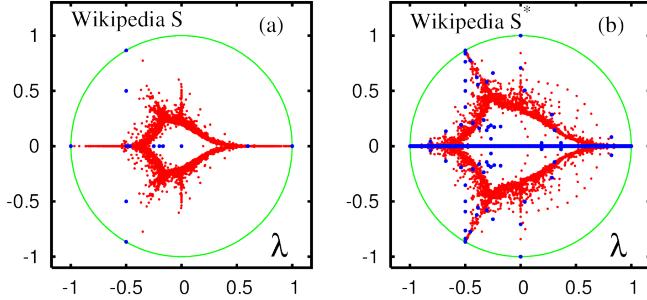


FIG. 22 (Color online) Complex eigenvalue spectra λ of S (a) and S^* (b) for English Wikipedia of Aug 2009 with $N = 3282257$ articles and $N_\ell = 71012307$ links. Red/gray dots are core space eigenvalues, blue/black dots are subspace eigenvalues and the full green/gray curve shows the unit circle. The core space eigenvalues are computed by the projected Arnoldi method with Arnoldi dimension $n_A = 6000$. After (Eom *et al.*, 2013b).

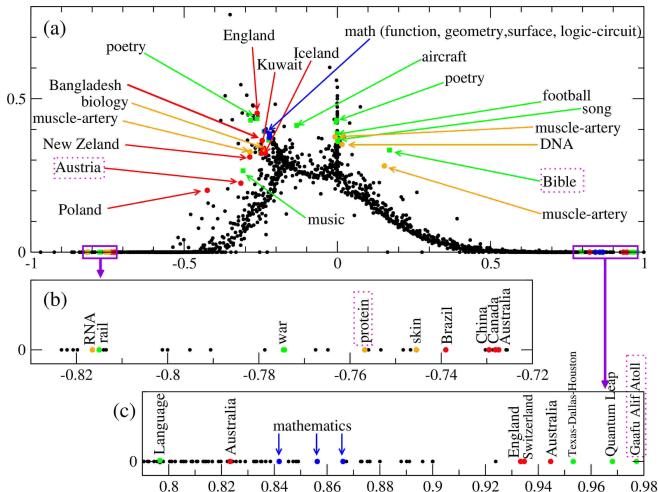


FIG. 23 (Color online) Complex eigenvalue spectrum of the matrices S for English Wikipedia Aug 2009. Highlighted eigenvalues represent different communities of Wikipedia and are labeled by the most repeated and important words following word counting of first 1000 nodes. Panel (a) shows complex plane for positive imaginary part of eigenvalues, while panels (b) and (c) zoom in the negative and positive real parts. After (Ermann *et al.*, 2013b).

C. Communities and eigenstates of Google matrix

The properties of eigenstates of Google matrix of Wikipedia Aug 2009 are analyzed in (Ermann *et al.*, 2013b). The global idea is that the eigenstates with large

values of $|\lambda|$ select certain specific communities. If $|\lambda|$ is close to unity then a relaxation of probability from such nodes is rather slow and we can expect that such eigenstates highlight some new interesting information even if these nodes are located on a tail of PageRank. The important advantage of the Wikipedia network is that its nodes are Wikipedia articles with a relatively clear meaning allowing to understand the origins of appearance of certain nodes in one community.

The localization properties of eigenvectors ψ_i of the Google matrix can be analyzed with the help of IPR ξ (see Sec. III.E). Another possibility is to fit a decay of an eigenstate amplitude by a power law $|\psi_i(K_i)| \sim K_i^b$ where K_i is the index ordering $|\psi_i(j)|$ by monotonically decreasing amplitude (similar to $P(K)$ for PageRank). The exponents b on the tails of $|\psi_i(j)|$ are found to be typically in the range $-2 < b < -1$ (Ermann *et al.*, 2013b). At the same time the eigenvectors with large complex eigenvalues or real eigenvalues close to ± 1 are quite well localized on $\xi \approx 10^2 - 10^3$ nodes that is much smaller than the whole network size $N \approx 3 \times 10^6$.

To understand the meaning of other eigenstates in the core space we order selected eigenstates by their decreasing value $|\psi_i(j)|$ and apply word frequency analysis for the first 1000 articles with $K_i \leq 1000$. The mostly frequent word of a given eigenvector is used to label the eigenvector name. These labels with corresponding eigenvalues are shown in Fig. 23. There are four main categories for the selected eigenvectors belonging to countries (red/gray), biology and medicine (orange/very light gray), mathematics (blue/black) and others (green/light gray). The category of others contains rather diverse articles about poetry, Bible, football, music, American TV series (e.g. Quantum Leap), small geographical places (e.g. Gaafu Alif Atoll). Clearly these eigenstates select certain specific communities which are relatively weakly coupled with the main bulk part of Wikipedia that generates relatively large modulus of $|\lambda_i|$.

For example, for the article *Gaafu Alif Atoll* the eigenvector is mainly localized on names of small atolls forming *Gaafu Alif Atoll*. Clearly this case represents well localized community of articles mainly linked between themselves that gives slow relaxation rate of this eigenmode with $\lambda = 0.9772$ being rather close to unity. Another eigenvector has a complex eigenvalue with $|\lambda| = 0.3733$ and the top article *Portal:Bible*. Another two articles are *Portal:Bible/Featured chapter/archives*, *Portal:Bible/Featured article*. These top 3 articles have very close values of $|\psi_i(j)|$ that seems to be the reason why we have $\varphi = \arg(\lambda_i) = 0.3496\pi$ being very close to $\pi/3$. Examples of other eigenvectors are discussed in (Ermann *et al.*, 2013b) in detail.

The analysis performed in (Ermann *et al.*, 2013b) for Wikipedia Aug 2009 shows that the eigenvectors of the Google matrix of Wikipedia clearly identify certain communities which are relatively weakly connected with the Wikipedia core when the modulus of corresponding eigenvalue is close to unity. For moderate values of $|\lambda|$

we still have well defined communities which are however have stronger links with some popular articles (e.g. countries) that leads to a more rapid decay of such eigenmodes. Thus the eigenvectors highlight interesting features of communities and network structure. However, a priori, it is not evident what is a correspondence between the numerically obtained eigenvectors and the specific community features in which someone has a specific interest. In fact, practically each eigenvector with a moderate value $|\lambda| \sim 0.5$ selects a certain community and there are many of them. So it remains difficult to target and select from eigenvalues λ a specific community one is interested.

The spectra and eigenstates of other networks like WWW of Cambridge 2011, Le Monde, BBC and PCN of Python are discussed in (Ermann *et al.*, 2013b). It is found that IPR values of eigenstates with large $|\lambda|$ are well localized with $\xi \ll N$. The spectra of each network have significant differences from one another.

D. Top people of Wikipedia

There is always a significant public interest to know who are most significant historical figures, or persons, of humanity. The Hart list of the top 100 people who, according to him, most influenced human history, is available at (Hart, 1992). Hart “ranked these 100 persons in order of importance: that is, according to the total amount of influence that each of them had on human history and on the everyday lives of other human beings” (Hart, 1992). Of course, a human ranking can be always objected arguing that an investigator has its own preferences. Also investigators from different cultures can have different view points on a same historical figure. Thus it is important to perform ranking of historical figures on purely mathematical and statistical grounds which exclude any cultural and personal preferences of investigators.

A detailed two-dimensional ranking of persons of English Wikipedia Aug 2009 has been done in (Zhirov *et al.*, 2010). Earlier studies had been done in a non-systematic way without any comparison with established top 100 lists (see these Refs. in (Wikipedia Top 100, 2014; Zhirov *et al.*, 2010)). Also at those times Wikipedia did not yet entered in its stabilized phase of development.

The top people of Wikipedia Aug 2009 are found to be 1. *Napoleon I of France*, 2. *George W. Bush*, 3. *Elizabeth II of the United Kingdom* for PageRank; 1. *Michael Jackson*, 2. *Frank Lloyd Wright*, 3. *David Bowie* for 2DRank; 1. *Kasey S. Pipes*, 2. *Roger Calmel*, 3. *Yury G. Chernavsky* for CheiRank (Zhirov *et al.*, 2010). For the PageRank list of 100 the overlap with the Hart list is at 35% (PageRank), 10% (2DRank) and almost zero for CheiRank. This is attributed to a very broad distribution of historical figures on 2D plane, as shown in Fig. 7, and a large variety of human activities. These activities are classified by 5 main categories: politics, religion, arts, sci-

ence, sport. For the top 100 PageRank persons we have the following distribution over these categories: 58, 10, 17, 15, 0 respectively. Clearly PageRank overestimates the significance of politicians which list is dominated by USA presidents not always much known to a broad public. For 2DRank we find respectively 24, 5, 62, 7, 2. Thus this rank highlights artistic sides of human activity. For CheiRank we have 15, 1, 52, 16, 16 so that the dominant contribution comes from arts, science and sport. The interesting property of this rank is that it selects many composers, singers, writers, actors. As an interesting feature of CheiRank we note that among scientists it selects those who are not so much known to a broad public but who discovered new objects, e.g. George Lyell who discovered many Australian butterflies or Nikolai Chernykh who discovered many asteroids. CheiRank also selects persons active in several categories of human activity.

For English Wikipedia Aug 2009 the distribution of top 100 PageRank, CheiRank and Hart’s persons on PageRank-CheiRank plane is shown in Fig. 7 (a).

The distribution of Hart’s top 100 persons on (K, K^*) plane for English Wikipedia in years 2003, 2005, 2007, Aug 2009, Dec 2009, 2011 is found to be stable for the period 2007-2011 even if certain persons change their ranks (Eom *et al.*, 2013b). The distribution of top 100 persons of Wikipedia Aug 2009 remains stable and compact for PageRank and 2DRank for the period 2007-2011 while for CheiRank the fluctuations of positions are large. This is due to the fact that outgoing links are easily modified and fluctuating.

The time evolution of distribution of top persons over fields of human activity is established in (Eom *et al.*, 2013b). PageRank persons are dominated by politicians whose percentage increases with time, while the percent of arts decreases. For 2DRank the arts are dominant but their percentage decreases with time. We also see the appearance of sport which is absent in PageRank. The mechanism of the qualitative ranking differences between two ranks is related to the fact that 2DRank takes into account via CheiRank a contribution of outgoing links. Due to that singers, actors, sportsmen improve their CheiRank and 2DRank positions since articles about them contain various music albums, movies and sport competitions with many outgoing links. Due to that the component of arts gets higher positions in 2DRank in contrast to dominance of politics in PageRank.

The interest to ranking of people via Wikipedia network is growing, as shows the recent study of English edition (Skiena and Ward, 2014).

E. Multilingual Wikipedia editions

The English edition allows to obtain ranking of historical people but as we saw the PageRank list is dominated by USA presidents that probably does not correspond to the global world view point. Hence, it is important to study multilingual Wikipedia editions which have now

287 languages and represent broader cultural views of the world.

One of the first cross-cultural study was done for 15 largest language editions constructing a network of links between set of articles of people biographies for each edition. However, the number of nodes and links in such a biographical network is significantly smaller compared to the whole network of Wikipedia articles and thus the fluctuations become rather large. For example, from the biographical network of the Russian edition one finds as the top person *Napoleon III* (and even not *Napoleon I*) (Aragón *et al.*, 2012), who has a rather low importance for Russia.

Another approach was used in (Eom and Shepelyansky , 2013a) ranking top 30 persons by PageRank, 2DRank and CheiRank algorithms for all articles of each of 9 editions and attributing each person to her/his native language. The selected editions are English (EN), French (FR), German (DE), Italian (IT), Spanish (ES), Dutch (NL), Russian (RU), Hungarian (HU) and Korean (KO). The aim here is to understand how different cultures evaluate a person? Is an important person in one culture is also important in the other culture? It is found that local heroes are dominant but also global heroes exist and create an effective network representing entanglement of cultures.

The top article of PageRank is usually *USA* or the name of country of a given language (FR, RU, KO). For NL we have at the top *beetle*, *species*, *France*. The top articles of CheiRank are various listings.

The distributions of articles density and top 30 persons for each rank algorithm are shown in Fig. 24 for four editions EN, FR, DE, RU. We see that in global the distributions have a similar shape that can be attributed to a fact that all editions describe the same world. However, local features of distributions are different corresponding to different cultural views on the same world (other 5 editions are shown in Fig.2 in (Eom and Shepelyansky , 2013a)). The top 30 persons for each edition are selected manually that represents a weak point of this study.

From the lists of top persons, the "fields" of activity are identified for each top 30 rank persons in which he/she is active on. The six activity fields are: politics, art, science, religion, sport and etc (here "etc" includes all other activities). As shown in Fig. 25, for PageRank, politics is dominant and science is secondarily dominant. The only exception is Dutch where science is the almost dominant activity field (politics has the same number of points). In case of 2DRank in Fig. 25, art becomes dominant and politics is secondarily dominant. In case of CheiRank, art and sport are dominant fields (see Fig.3 in (Eom and Shepelyansky , 2013a)). Thus for example, in CheiRank top 30 list we find astronomers who discovered a lot of asteroids, e.g. Karl Wilhelm Reinmuth (4th position in RU and 7th in DE), who was a prolific discoverer of about 400 of them. As a result, his article contains a long listing of asteroids discovered by him and giving him a high CheiRank. The distributions

of persons over activity fields are shown in Fig. 25 for 9 languages editions (marked by standard two letters used by Wikipedia).

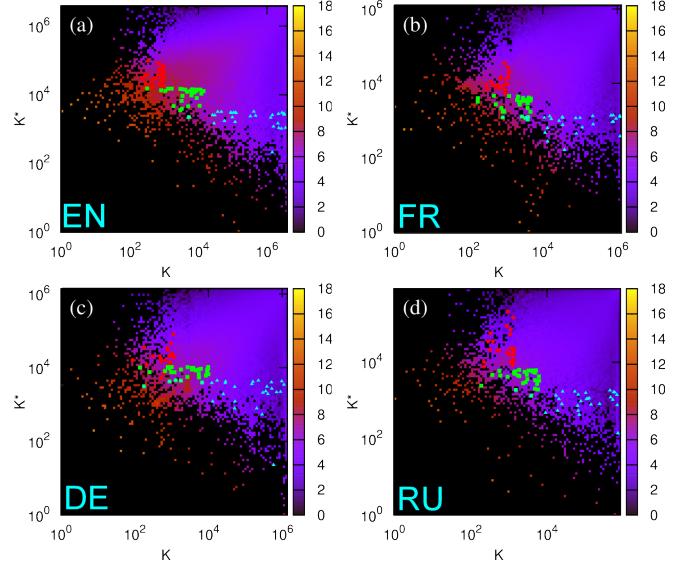


FIG. 24 (Color online) Density of Wikipedia articles in the PageRank-CheiRank plane (K, K^*) for four different language Wikipedia editions. The red (gray) points are top PageRank articles of persons, the green (light gray) squares are top 2DRank articles of persons and the cyan (dark gray) triangles are top CheiRank articles of persons. Wikipedia language editions are English EN (a), French FR (b), German DE (c), and Russian RU (d). Color bars show natural logarithm of density, changing from minimal nonzero density (dark) to maximal one (white), zero density is shown by black. After (Eom and Shepelyansky , 2013a).

The change of activity priority for different ranks is due to the different balance between incoming and outgoing links there. Usually the politicians are well known for a broad public, hence, the articles about politicians are pointed by many articles. However, the articles about politicians are not very communicative since they rarely point to other articles. In contrast, articles about persons in other fields like science, art and sport are more communicative because of listings of insects, planets, asteroids they discovered, or listings of song albums or sport competitions they gain.

On the basis of this approach one obtains local ranks of each of 30 persons $1 \leq K_{P,E,A} \leq 30$ for each edition E and algorithm A . Then an average ranking score of a person P is determined as $\Theta_{P,A} = \sum_E (31 - K_{P,E,A})$ for each algorithm. This method determines the global historical figures. The top global persons are 1. *Napoleon*, 2. *Jesus*, 3. *Carl Linnaeus* for PageRank; 1. *Micheal Jackson* , 2. *Adolf Hitler*, 3. *Julius Caesar* for 2DRank. For CheiRank the lists of different editions have rather low overlap and such an averaging is not efficient. The first positions reproduce top persons from English edition discussed in Sec. IX.D, however, the next ones are different.

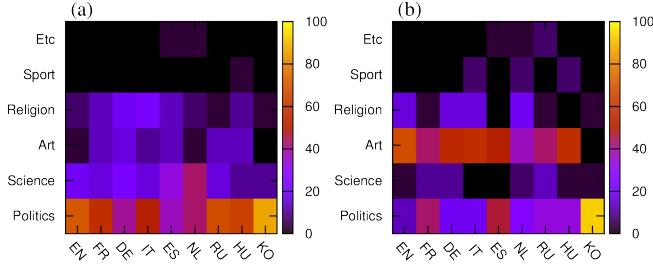


FIG. 25 (Color online) Distribution of top 30 persons over activity fields for PageRank (a) and 2DRank (b) for each of 9 Wikipedia editions. The color bar shows the values in percent. After (Eom and Shepelyansky , 2013a).

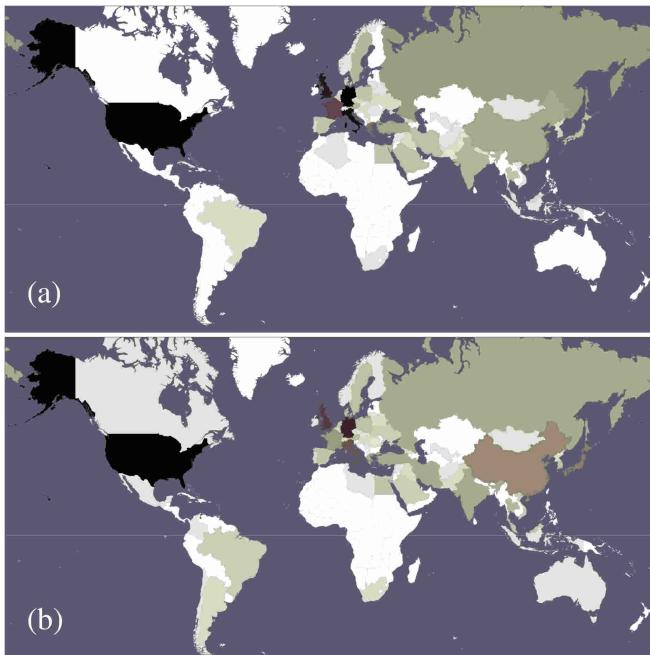


FIG. 26 Number of appearances of historical figures of a given country, obtained from 24 lists of top 100 persons of PageRank (a) and 2DRank (b), shown on the world map. Color changes from zero (white) to maximum (black), it corresponds to average number of person appearances per country. After (Eom *et al.* , 2014).

Since each person is attributed to her/his native language it is also possible for each edition to obtain top local heroes who have native language of the edition. For example, we find for PageRank for EN *George W. Bush*, *Barack Obama*, *Elizabeth II*; for FR *Napoleon*, *Louis XIV of France*, *Charles de Gaulle*; for DE *Adolf Hitler*, *Martin Luther*, *Immanuel Kant*; for RU *Peter the Great*, *Joseph Stalin*, *Alexander Pushkin*. For 2DRank we have for EN *Frank Sinatra*, *Paul McCartney*, *Michael Jackson*; for FR *Francois Mitterrand*, *Jacques Chirac*, *Honore de Balzac*; for DE *Adolf Hitler*, *Otto von Bismarck*, *Ludwig van Beethoven*; for RU *Dmitri Mendeleev*, *Peter the Great*,

Yaroslav the Wise. These ranking results are rather reasonable for each language. Results for other editions and CheiRank are given in (Eom and Shepelyansky , 2013a).

A weak point of above study is a manual selection of persons and a not very large number of editions. A significant improvement has been reached in a recent study (Eom *et al.* , 2014) where 24 editions have been analyzed. These 24 languages cover 59 percent of world population, and these 24 editions covers 68 percent of the total number of Wikipedia articles in all 287 available languages. Also the selection of people from the rank list of each edition is now done in an automatic computerized way. For that a list of about 1.1 million biographical articles about people with their English names is generated. From this list of persons, with their biographical article title in the English Wikipedia, the corresponding titles in other language editions are determined using the inter-language links provided by Wikipedia.

Using the corresponding articles, identified by the inter-languages links in different language editions, the top 100 persons are obtained from the rankings of all Wikipedia articles of each edition. A birth place, birth date, and gender of each top 100 ranked person are identified, based on DBpedia or a manual inspection of the corresponding Wikipedia biographical article, when for the considered person no DBpedia data were available. In this way 24 lists of top 100 persons for each edition are obtained in PageRank with 1045 unique names and in 2DRank with 1616 unique names. Each of the 100 historical figures is attributed to a birth place at the country level, to a birth date in year, to a gender, and to a cultural language group. The birth place is assigned according to the current country borders. The cultural group of historical figures is assigned by the most spoken language of their birth place at the current country level. The considered editions are: English EN, Dutch NL, German DE, French FR, Spanish, ES, Italian IT, Portuguese PT, Greek, EL, Danish DA, Swedish SV, Polish PL, Hungarian HU, Russian RU, Hebrew HE, Turkish TR, Arabic AR, Persian FA, Hindi HI, Malaysian MS, Thai TH, Vietnamese VI, Chinese ZH, Korean KO, Japanese JA (dated by February 2013). The size of network changes from maximal value $N = 4212493$ for EN to minimal one $N = 78953$ for TH.

All persons are ranked by their average rank score $\Theta_{P,A} = \sum_E (101 - K_{P,E,A})$ with $1 \leq K_{P,E,A} \leq 100$ similar to the study of 9 editions described above. For PageRank the top global historical figures are *Carl Linnaeus*, *Jesus*, *Aristotle* and for 2DRank we obtain *Adolf Hitler*, *Michael Jackson*, *Madonna (entertainer)*. Thus the averaging over 24 editions modifies the top ranking. The list of top 100 PageRank global persons has overlap of 43 persons with the Hart list (Hart, 1992). Thus the averaging over 24 editions gives a significant improvement compared to 35 persons overlap for the case of English edition only (Zhirov *et al.*, 2010). For comparison we note that the top 100 list of historical figures has been also determined recently by (Pantheon MIT project, 2014) having

overlap of 42 persons with the Hart list. This Pantheon MIT list is established on the basis of number of editions and number of clicks on an article of a given person without using rank algorithms discussed here. The overlap between top 100 PageRank list and top 100 Pantheon list is 44 percent. More data are available in (Eom *et al.*, 2014).

The fact that *Carl Linnaeus* is the top historical figure of Wikipedia PageRank list came out as a surprise for media and broad public (see (Wikipedia Top 100, 2014)). This ranking is due to the fact that *Carl Linnaeus* created a classification of world species including, animals, insects, herbs, trees etc. Thus all articles of these species point to the article *Carl Linnaeus* in various languages. As a result *Carl Linnaeus* appears on almost top positions in all 24 languages. Hence, even if a politician, like *Barak Obama*, takes the second position in his country language EN (*Napoleon* is at the first position in EN) he is usually placed at low ranking in other language editions. As a result *Carl Linnaeus* takes the first global PageRank position.

The number of appearances of historical persons in 24 lists of top 100 for each edition can be distributed over present world countries according to the birth place of each person. This geographical distribution is shown in Fig. 26 for PageRank and 2DRank. In PageRank the top countries are *DE, USA, IT* and in 2DRank *US, DE, UK*. The appearance of many UK and US singers improves the positions of English speaking countries in 2DRank.

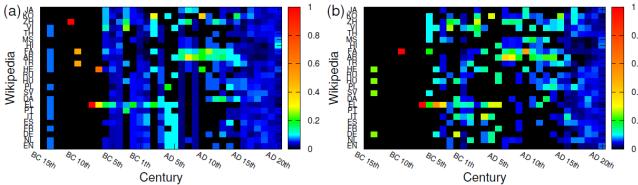


FIG. 27 (Color online) Birth date distributions over 35 centuries of top historical figures from each Wikipedia edition marked by two letters standard notation of Wikipedia. Panels: (a) column normalized birth date distributions of PageRank historical figures; (b) same as (a) for 2DRank historical figures. After (Eom *et al.*, 2014).

The distributions of the top PageRank and 2DRank historical figures over 24 Wikipedia editions for each century are shown in Fig. 27. Each person is attributed to a century according to the birth date covering the range of 35 centuries from BC 15th to AD 20th centuries. For each century the number of persons for each century is normalized to unity to see more clearly relative contribution of each language for each century.

The Greek edition has more historical figures in BC 5th century because of Greek philosophers. Also most of western-southern European language editions, including English, Dutch, German, French, Spanish, Italian, Portuguese, and Greek, have more top historical figures because they have Augustine the Hippo and Jus-

tinian I in common. The Persian (FA) and the Arabic (AR) Wikipedia have more historical figures comparing to other language editions (in particular European language editions) from the 6th to the 12th century that is due to Islamic leaders and scholars. The data of Fig. 27 clearly show well pronounced patterns, corresponding to strong interactions between cultures: from BC 5th century to AD 15th century for JA, KO, ZH, VI; from AD 6th century to AD 12th century for FA, AR; and a common birth pattern in EN, EL, PT, IT, ES, DE, NL (Western European languages) from BC 5th century to AD 6th century. A detailed analysis shows that even in BC 20th century each edition has a significant fraction of persons of its own language so that even with on going globalization there is a significant dominance of local historical figures for certain cultures. More data on the above points and gender distributions are available in (Eom *et al.*, 2014).

F. Networks and entanglement of cultures

We now know how a person of a given language is ranked by editions of other languages. Therefore, if a top person from a language edition *A* appears in another edition *B*, we can consider this as a 'cultural' influence from culture *A* to *B*. This generates entanglement in a network of cultures. Here we associate a language edition with its corresponding culture considering that a language is a first element of culture, even if a culture is not reduced only to a language. In (Eom and Shepelyansky, 2013a) a person is attributed to a given language, or culture, according to her/his native language fixed via corresponding Wikipedia article. In (Eom *et al.*, 2014) the attribution to a culture is done via a birth place of a person, each language is considered as a proxy for a cultural group and a person is assigned to one of these cultural groups based on the most spoken language of her/his birth place at the country level. If a person does not belong to any of studied editions then he/she is attributed to an additional cultural group world WR.

After such an attributions of all persons the two networks of cultures are constructed based on the top PageRank historical figures and top 2DRank historical figures respectively. Each culture (i.e. language) is represented as a node of the network, and the weight of a directed link from culture *A* to culture *B* is given by the number of historical figures belonging to culture *B* (e.g. French) appearing in the list of top 100 historical figures for a given culture *A* (e.g. English).

For example, according to (Eom *et al.*, 2014), there are 5 French historical figures among the top 100 PageRank historical figures of the English Wikipedia, so we can assign weight 5 to the link from English to French. Thus, Fig. 28(a) and Fig. 28(b) represent the constructed networks of cultures defined by appearances of the top PageRank historical figures and top 2DRank historical figures, respectively.

In total we have two networks with 25 nodes which include our 24 editions and an additional node WR for all other world cultures. Persons of a given culture are not taken into account in the rank list of language edition of this culture. Then following the standard rules (1) the Google matrix of network of cultures is constructed by normalization of sum of all elements in each column to unity. The matrix $G_{KK'}$, written in the PageRank indexes K, K' is shown in Fig. 29 for persons from PageRank (a) and 2DRank (b) lists. The matrix G^* is constructed in the same way as G for the network with inverted directions of links.

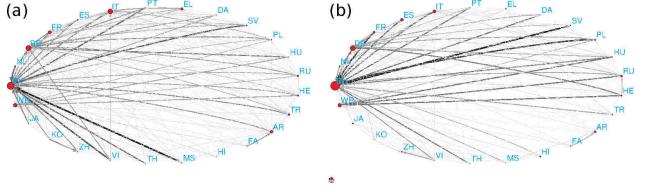


FIG. 28 (Color online) Network of cultures, obtained from 24 Wikipedia languages and the remaining world (WR), considering (a) top 100 PageRank historical figures and (b) top 100 2DRank historical figures. The link width and darkness are proportional to a number of foreign historical figures quoted in top 100 of a given culture, the link direction goes from a given culture to cultures of quoted foreign historical figures, quotations inside cultures are not considered. The size of nodes is proportional to their PageRank. After (Eom *et al.* , 2014).

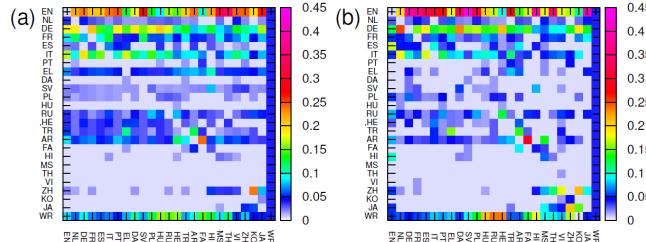


FIG. 29 (Color online) Google matrix of network of cultures shown in Fig 28 (a) and (b) respectively. The matrix elements G_{ij} are shown by color with damping factor $\alpha = 0.85$. After (Eom *et al.* , 2014).

From the obtained matrix G and G^* we determine PageRank and CheiRank vectors and then the PageRank-CheiRank plane (K, K^*), shown in Fig. 30, for networks of cultures from Fig. 28. Here K indicates the ranking of a given culture ordered by how many of its own top historical figures appear in other Wikipedia editions, and K^* indicates the ranking of a given culture according to how many of the top historical figures in the considered culture are from other cultures. It is important to note that for 24 editions the world node WR appears on positions $K = 3$ or $K = 4$, for panels (a), (b) in Fig. 30, signifying that the 24 editions capture

the main part of historical figures born in these cultures. We note that for 9 editions in (Eom and Shepelyansky , 2013a) the node WR was at the top position for PageRank so that a significant fraction of historical figures was attributed to other cultures.

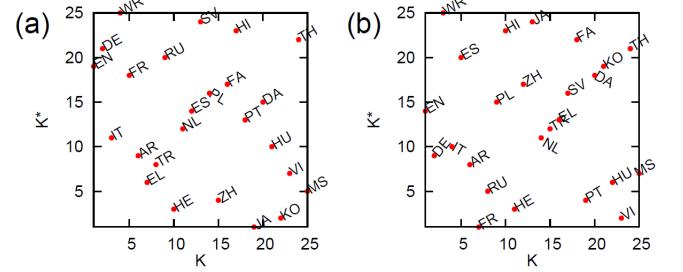


FIG. 30 (Color online) PageRank-CheiRank plane of cultures with corresponding indexes K and K^* obtained from the network of cultures based on (a) top 100 PageRank historical figures, (b) top 100 2DRank historical figures. After (Eom *et al.* , 2014).

From the data of Fig. 30 we obtain at the top positions of K cultures EN, DE, IT showing that other cultures strongly point to them. However, we can argue that for cultures it is also important to have strong communicative property and hence it is important to have 2DRank of cultures at top positions. On the top 2DRank position we have Greek, Turkish and Arabic (for PageRank persons) in Fig. 30(a) and French, Russian and Arabic (for 2DRank persons) in Fig. 30(b). This demonstrates the important historical influence of these cultures both via importance (incoming links) and communicative (outgoing links) properties present in a balanced manner.

Thus the described research across Wikipedia language editions suggests a rigorous mathematical way, based on Markov chains and Google matrix, for recognition of important historical figures and analysis of interactions of cultures at different historical periods and in different world regions. Such an approach recovers 43 percent of persons from the well established Hart historical study (Hart, 1992), that demonstrates the reliability of this method. We think that a further extension of this approach to a larger number of Wikipedia editions will provide a more detailed and balanced analysis of interactions of world cultures.

X. GOOGLE MATRIX OF SOCIAL NETWORKS

Social networks like Facebook, LiveJournal, Twitter, Vkontakte start to play a more and more important role in modern society. The Twitter network is a directed one and here we consider its spectral properties following mainly the analysis reported in (Frahm and Shepelyansky , 2012b).

A. Twitter network

Twitter is a rapidly growing online directed social network. For July 2009 a data set of this entire network is available with $N = 41652230$ nodes and $N_\ell = 1468365182$ links (for data sets see Refs. in (Frahm and Shepelyansky , 2012b)). For this case the spectrum and eigenstate properties of the corresponding Google matrix have been analyzed in detail using the Arnoldi method and standard PageRank and CheiRank computations (Frahm and Shepelyansky , 2012b). For the Twitter network the average number of links per node $\zeta = N_\ell/N \approx 35$ and the general inter-connectivity between top PageRank nodes are considerably larger than for other networks such as Wikipedia (Sec. IX) or UK universities (Sec. VIII) as can be seen in Figs. 31 and 32.

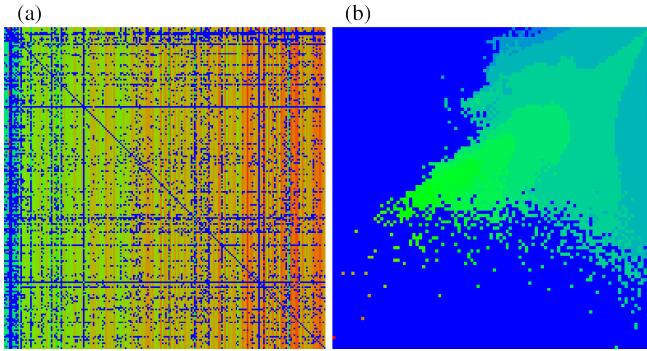


FIG. 31 (Color online) Panel (a): Google matrix of Twitter, matrix elements of G are shown in the basis of PageRank index K of matrix $G_{KK'}$. Here, x (and y) axis show K (and K') with the range $1 \leq K, K' \leq 200$. Panel (b) shows the density of nodes $W(K, K^*)$ of Twitter on PageRank-CheiRank plane (K, K^*) , averaged over 100×100 logarithmically equidistant grids for $0 \leq \ln K, \ln K^* \leq \ln N$ with the normalization condition $\sum_{K, K^*} W(K, K^*) = 1$. The x -axis corresponds to $\ln K$ and the y -axis to $\ln K^*$. In both panels color varies from blue/black at minimal value to red/gray at maximal value; here $\alpha = 0.85$. After (Frahm and Shepelyansky , 2012b).

The decay of PageRank probability can be approximately described by an algebraic decay with the exponent $\beta \approx 0.54$ while for CheiRank we have a larger value $\beta \approx 0.86$ (Frahm and Shepelyansky , 2012b) that is opposite to the usual situation. The image of top matrix elements of $G_{KK'}$ with $1 \leq K, K' \leq 200$ is shown in Fig. 31. The density distribution of nodes on (K, K^*) plane is also shown there. It is somewhat similar to those of Wikipedia case in Fig. 24, may be with a larger density concentration along the line $K \approx K^*$.

However, the most striking feature of G matrix elements is a very strong interconnectivity between top PageRank nodes. Thus for Twitter the top $K \leq 1000$ elements fill about 70% of the matrix and about 20% for size $K \leq 10^4$. For Wikipedia the filling factor is smaller by a factor $10 - 20$. In particular the number N_G of links between K top PageRank nodes behaves for $K \leq 10^3$

as $N_G \sim K^{1.993}$ while for Wikipedia $N_G \sim K^{1.469}$. The exponent for N_G , being close to 2 for Twitter, indicates that for the top PageRank nodes the Google matrix is macroscopically filled with a fraction $0.6 - 0.8$ of non-vanishing matrix elements (see also Figs. 31 and 32) and the very well connected top PageRank nodes can be considered as the Twitter elite (Kandiah and Shepelyansky , 2012). For Wikipedia the interconnectivity among top PageRank nodes has an exponent 1.5 being somewhat reduced but still stronger as compared to certain university networks where typical exponents are close to unity (for the range $10^2 \leq K \leq 10^4$). The strong interconnectivity of Twitter is also visible in its global logarithmic density distribution of nodes in the PageRank-CheiRank plane (K, K^*) (Fig. 31 (b)) which shows a maximal density along a certain ridge along a line $\ln K^* = \ln K + \text{const.}$ with a significant large number of nodes at small values $K, K^* < 1000$.

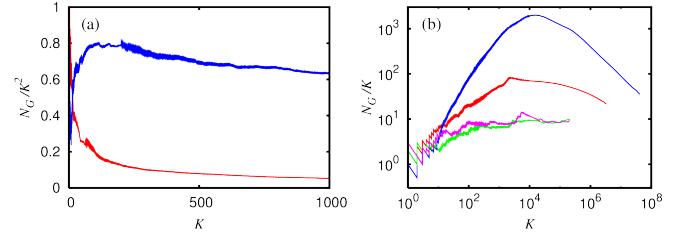


FIG. 32 (Color online) (a) Dependence of the area density $g_K = N_G/K^2$ of nonzero elements of the adjacency matrix among top PageRank nodes on the PageRank index K for Twitter (blue/black curve) and Wikipedia (red/gray curve) networks, data are shown in linear scale. (b) Linear density N_G/K of the same matrix elements shown for the whole range of K in log-log scale for Twitter (blue curve), Wikipedia (red curve), Oxford University 2006 (magenta curve) and Cambridge University 2006 (green curve) (curves from top to bottom at $K = 100$). After (Frahm and Shepelyansky , 2012b).

The decay exponent of the PageRank is for Twitter $\beta = 0.540$ (for $1 \leq K \leq 10^6$), which indicates a precursor of a delocalization transition as compared to Wikipedia ($\beta = 0.767$) or WWW ($\beta \approx 0.9$), caused by the strong interconnectivity (Frahm and Shepelyansky , 2012b). The Twitter network is also characterized by a large value of PageRank-CheiRank correlator $\kappa = 112.6$ that is by a factor $30 - 60$ larger than this value for Wikipedia and University networks. Such a larger value of κ results from certain individual large values $\kappa_i = NP(K(i))P^*(K^*(i)) \sim 1$. It is argued that this is related to a very strong inter-connectivity between top K PageRank users of the Twitter network (Frahm and Shepelyansky , 2012b).

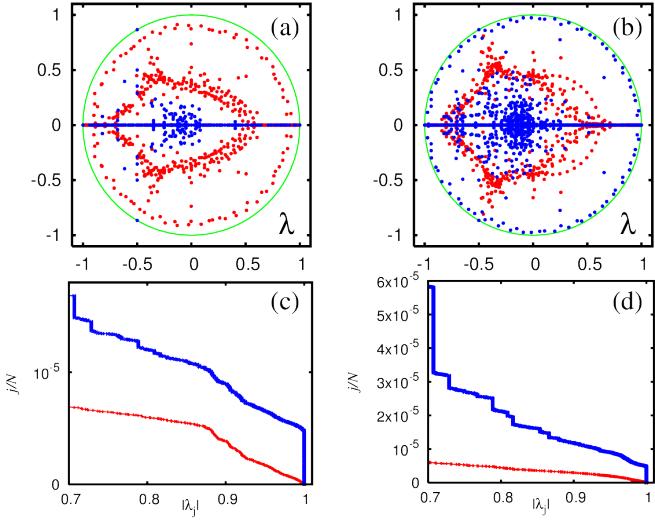


FIG. 33 (Color online) Spectrum of the Twitter matrix S (a) and (c), and S^* (b) and (d). Panels (a) and (b) show subspace eigenvalues (blue/black dots) and core space eigenvalues (red/gray dots) in λ -plane (green/gray curve shows unit circle); there are 17504 (66316) invariant subspaces, with maximal dimension 44 (2959) and the sum of all subspace dimensions is $N_s = 40307$ (180414). The core space eigenvalues are obtained from the Arnoldi method applied to the core space subblock S_{cc} of S with Arnoldi dimension $n_A = 640$. Panels (c) and (d) show the fraction j/N of eigenvalues with $|\lambda| > |\lambda_j|$ for the core space eigenvalues (red/gray bottom curve) and all eigenvalues (blue/black top curve) from raw data ((a) and (b) respectively). The number of eigenvalues with $|\lambda_j| = 1$ is 34135 (129185) of which 17505 (66357) are at $\lambda_j = 1$; this number is (slightly) larger than the number of invariant subspaces which have each at least one unit eigenvalue. Note that in panels (c) and (d) the number of eigenvalues with $|\lambda_j| = 1$ is artificially reduced to 200 in order to have a better scale on the vertical axis. The correct numbers of those eigenvalues correspond to $j/N = 8.195 \times 10^{-4}$ (c) and 3.102×10^{-3} (d) which are strongly outside the vertical panel scale. After (Frahm and Shepelyansky , 2012b).

The spectra of matrices S and S^* are obtained with the help of the Arnoldi method for a relatively modest Arnoldi dimension due to a very large matrix size. The largest n_A modulus eigenvalues $|\lambda|$ are shown in Fig. 33. The invariant subspaces (see Sec. III.C) for the Twitter network cover about $N_s = 4 \times 10^4$ (1.8×10^5) nodes for S (S^*) leading to 1.7×10^4 (6.6×10^4) eigenvalues with $\lambda_j = 1$ or even 3.4×10^4 (1.3×10^5) eigenvalues with $|\lambda_j| = 1$. However, for Twitter the fraction of subspace nodes $g_1 = N_s/N \approx 10^{-3}$ is smaller than the fraction $g_1 \approx 0.2$ for the university networks of Cambridge or Oxford (with $N \approx 2 \times 10^5$) since the size of the whole Twitter network is significantly larger. The complex spectra of S and S^* also show the cross and triple-star structures, as in the cases of Cambridge and Oxford 2006 (see Fig. 17), even though for the Twitter network they are significantly less pronounced.

B. Poisson statistics of PageRank probabilities

From a physical viewpoint one can conjecture that the PageRank probabilities are described by a steady-state quantum Gibbs distribution over certain quantum levels with energies E_i by the identification $P(i) = \exp(-E_i/T)/Z$ with $Z = \sum_i \exp(-E_i/T)$ (Frahm and Shepelyansky , 2014a). In some sense this conjecture assumes that the operator matrix G can be represented as a sum of two operators G_H and G_{NH} where G_H describes a Hermitian system while G_{NH} represents a non-Hermitian operator which creates a system thermalization at a certain effective temperature T with the quantum Gibbs distribution over energy levels E_i of the operator G_H .

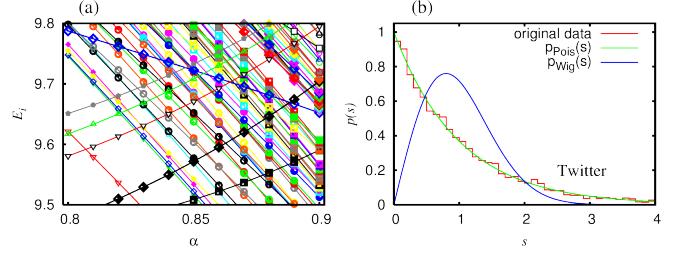


FIG. 34 (Color online) Panel (a) shows the dependence of certain top PageRank levels $E_i = -\ln(P_i)$ on the damping factor α for Twitter network. Data points on curves with one color corresponds to the same node i ; about 150 levels are shown close to the minimal energy $E \approx 7.5$. Panel (b) represents the histogram of unfolded level spacing statistics for Twitter at $10 < K \leq 10^4$. The Poisson distribution $p_{\text{Pois}}(s) = \exp(-s)$ and the Wigner surmise $p_{\text{Wig}}(s) = \frac{\pi}{2} s \exp(-\frac{\pi}{4} s^2)$ are also shown for comparison. After (Frahm and Shepelyansky , 2014a).

The identification of PageRank with an energy spectrum allows to study the corresponding level statistics which represents a well known concept in the framework of Random Matrix Theory (Guhr *et al.*, 1998; Mehta, 2004). The most direct characteristic is the probability distribution $p(s)$ of unfolded level spacings s . Here $s = (E_{i+1} - E_i)/\Delta E$ is a spacing between nearest levels measured in the units of average local energy spacing ΔE . The unfolding procedure (Guhr *et al.*, 1998; Mehta, 2004) requires the smoothed dependence of E_i on the index K which is obtained from a polynomial fit of $E_i \sim \ln(P_i)$ with $\ln(K)$ as argument (Frahm and Shepelyansky , 2014a).

The statistical properties of fluctuations of levels have been extensively studied in the fields of RMT (Mehta, 2004), quantum chaos (Haake, 2010) and disordered solid state systems (Evers and Mirlin , 2008). It is known that integrable quantum systems have $p(s)$ well described by the Poisson distribution $p_{\text{Pois}}(s) = \exp(-s)$. In contrast the quantum systems, which are chaotic in the classical limit (e.g. Sinai billiard), have $p(s)$ given by the RMT being close to the Wigner surmise $p_{\text{Wig}}(s) = \frac{\pi}{2} s \exp(-\frac{\pi}{4} s^2)$ (Bohigas *et al.*, 1984). Also the Ander-

son localized phase is characterized by $p_{\text{Pois}}(s)$ while in the delocalized regime one has $p_{\text{Wig}}(s)$ (Evers and Mirlin , 2008).

The results for the Twitter PageRank level statistics (Frahm and Shepelyansky , 2014a) are shown in Fig. 34. We find that $p(s)$ is well described by the Poisson distribution. Furthermore, the evolution of energy levels E_i with the variation of the damping factor α shows many level crossings which are typical for Poisson statistics. We may note that here each level has its own index so that it is rather easy to see if there is a real or avoided level crossing.

The validity of the Poisson statistics for PageRank probabilities is confirmed also for the networks of Wikipedia editions in English, French and German from Fig. 24 (Frahm and Shepelyansky , 2014a). We argue that due to absence of level repulsion the PageRank order of nearby nodes can be easily interchanged. The obtained Poisson law implies that the nearby PageRank probabilities fluctuate as random independent variables.

XI. GOOGLE MATRIX ANALYSIS OF WORLD TRADE

During the last decades the trade between countries has been developed in an extraordinary way. Usually countries are ranked in the world trade network (WTN) taking into account their exports and imports measured in *USD* (CIA, 2009). However, the use of these quantities, which are local in the sense that countries know their total imports and exports, could hide the information of the centrality role that a country plays in this complex network. In this section we present the two-dimensional Google matrix analysis of the WTN introduced in (Ermann and Shepelyansky , 2011b). Some previous studies of global network characteristics were considered in (Garlaschelli and Loffredo , 2005; Serrano *et al.*, 2007), degree centrality measures were analyzed in (De Benedictis and Tajoli , 2011) and a time evolution of network global characteristics was studied in (He and Deem , 2010). Topological and clustering properties of multiplex network of various commodities were discussed in (Barigozzi *et al.*, 2010), and an ecological ranking based on the nestedness of countries and products was presented in (Ermann and Shepelyansky , 2013a).

The money exchange between countries defines a directed network. Therefore Google matrix analysis can be introduced in a natural way. PageRank and CheiRank algorithms can be easily applied to this network with a straightforward correspondence with imports and exports. Two-dimensional ranking, introduced in Sec. IV, gives an illustrative representation of global importance of countries in the WTN. The important element of Google ranking of WTN is its democratic treatment of all world countries, independently of their richness, that follows the main principle of the United Nations (UN).

A. Democratic ranking of countries

The WTN is a directed network that can be constructed considering countries as nodes and money exchange as links. We follow the definition of the WTN of (Ermann and Shepelyansky , 2011b) where trade information comes from (UN COMTRADE, 2011). These data include all trades between countries for different products (using Standard International Trade Classification of goods, SITC1) from 1962 to 2009.

All useful information of the WTN is expressed via the *money matrix* M , which definition, in terms of its matrix elements M_{ij} , is defined as the money transfer (in *USD*) from country j to country i in a given year. This definition can be applied to a given specific product or to *all commodities*, which represent the sum over all products.

In contrast to the binary adjacency matrix A_{ij} of WWW (as the ones analyzed in SVIII and SX for example) M has weighted elements. This corresponds to a case when there are in principle multiple number of links from j to i and this number is proportional to *USD* amount transfer. Such a situation appears in Sec. VI for Ulam networks and Sec. VII for Linux PCN with a main difference that for the WTN case there is a very large variation of mass matrix elements M_{ij} , related to the fact that there is a very strong variation of richness of various countries.

Google matrices G and G^* are constructed according to the usual rules and relation (1) with M_{ij} and its transposed: $S_{ij} = M_{ij}/m_j$ and $S_{ij}^* = M_{ji}/m_j^*$ where $S_{ij} = 1/N$ and $S_{ij}^* = 1/N$, if for a given j all elements $M_{ij} = 0$ and $M_{ji} = 0$ respectively. Here $m_j = \sum_i M_{ij}$ and $m_j^* = \sum_i M_{ji}$ are the total export and import mass for country j . Thus the sum in each column of G or G^* is equal to unity. In this way Google matrices G and G^* of WTN allow to treat all countries on equal grounds independently of the fact if a given country is rich or poor. This kind of analysis treats in a democratic way all world countries in consonance with the standards of the UN.

The probability distributions of ordered PageRank $P(K)$ and CheiRank $P^*(K^*)$ depend on their indexes in a rather similar way with a power law decay given by β . For the fit of top 100 countries and *all commodities* the average exponent value is close to $\beta = 1$ corresponding to the Zipf law (Zipf, 1949).

The distribution of countries on PageRank-CheiRank plane for trade in *all commodities* in year 2008 is shown in panels (a) and (b) of Fig. 35 at $\alpha = 0.5$. Even if the Google matrix approach is based on a democratic ranking of international trade, being independent of total amount of export-import and PIB for a given country, the top ranks K and K^* belong to the group of industrially developed countries. This means that these countries have efficient trade networks with optimally distributed trade flows. Another striking feature of global distribution is that it is concentrated along the main diagonal $K = K^*$. This feature is not present in other networks

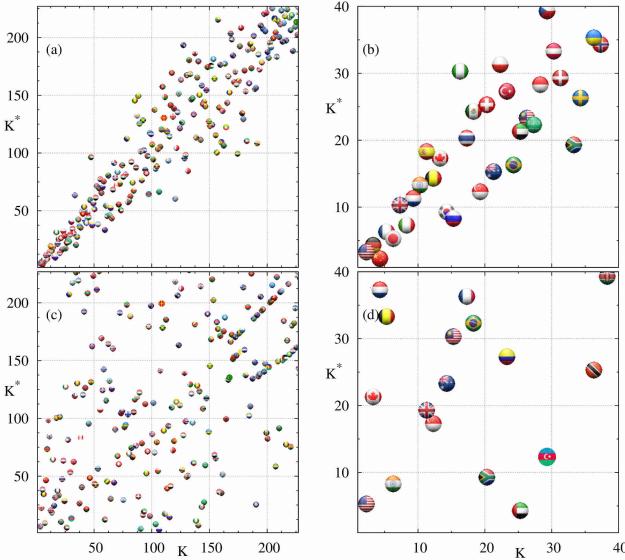


FIG. 35 (Color online) Country positions in PageRank-CheiRank plane (K, K^*) for world trade in various commodities in 2008. Each country is shown by circle with its own flag (for a better visibility the circle center is slightly displaced from its integer position (K, K^*) along direction angle $\pi/4$). The panels show the ranking for trade in the following commodities: *all commodities* (a) and (b); and *crude petroleum* (c) and (d). Panels (a) and (c) show a global scale with all 227 countries, while (b) and (d) give a zoom in the region of 40×40 top ranks. After (Ermann and Shepelyansky , 2011b).

studied before. The origin of this density concentration is related to a simple economy reason: for each country the total import is approximately equal to export since each country should keep in average an economic balance. This balance does not imply a symmetric money matrix, used in gravity model of trade (see e.g. (De Benedictis and Tajoli , 2011; Krugman *et al.*, 2011)), as can be seen in the significant broadening of distribution of Fig. 35 (especially at middle values of $K \sim 100$).

For a given country its trade is doing well if its $K^* < K$ so that the country exports more than it imports. The opposite relation $K^* > K$ corresponds to a bad trade situation (e.g. Greece being significantly above the diagonal). We also can say that local minima in the curve of $(K^* - K)$ vs. K correspond to a successful trade while maxima mark bad traders. In 2008 most successful were China, R of Korea, Russia, Singapore, Brazil, South Africa, Venezuela (in order of K for $K \leq 50$) while among bad traders we note UK, Spain, Nigeria, Poland, Czech Rep, Greece, Sudan with especially strong export drop for two last cases.

A comparison between local and global rankings of countries for both imports and exports gives a new tool to analyze countries economy. For example, in 2008 the most significant differences between CheiRank and the rank given by total exports are for *Canada* and *Mexico*

with corresponding money export ranks $\tilde{K}^* = 11$ and 13 and with $K^* = 16$ and $K^* = 23$ respectively. These variations can be explained in the context that the export of these two countries is too strongly oriented on *USA*. In contrast *Singapore* moves up from $\tilde{K}^* = 15$ export position to $K^* = 11$ that shows the stability and broadness of its export trade, a similar situation appears for *India* moving up from $\tilde{K}^* = 19$ to $K^* = 12$ (see (Ermann and Shepelyansky , 2011b) for more detailed analysis).

B. Ranking of countries by trade in products

If we focus on the two-dimensional distribution of countries in a specific product we obtain a very different information. The symmetry approximately visible for *all commodities* is absolutely absent: the points are scattered practically over the whole square $N \times N$ (see Fig. 35). The reason of such a strong scattering is clear: e.g. for *crude petroleum* some countries export this product while other countries import it. Even if there is some flow from exporters to importers it remains relatively low. This makes the Google matrix to be very asymmetric. Indeed, the asymmetry of trade flow is well visible in panels (c) and (d) of Fig. 35.

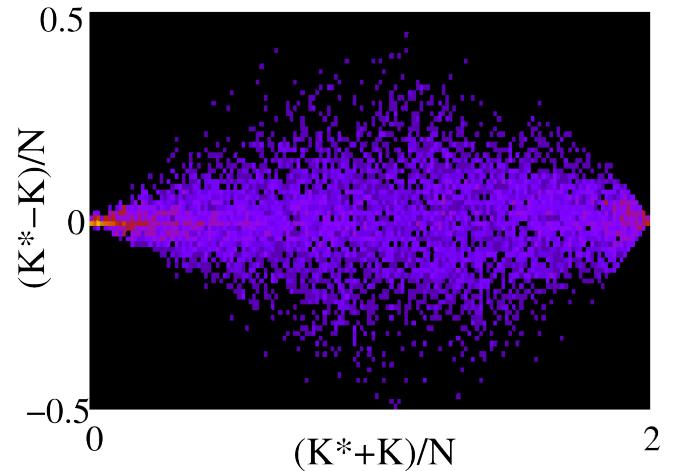


FIG. 36 (Color online) Spindle distribution for WTN of *all commodities* for all countries in the period 1962 - 2009 shown in the plane of $((K^* - K)/N, (K^* + K)/N)$ (coarse-graining inside each of 76×152 cells); data from the UN COMTRADE database. After (Ermann and Shepelyansky , 2011b).

The same comparison of global and local rankings done before for *all commodities* can be applied to specific products obtaining even more strong differences. For example for *crude petroleum* Russia moves up from $\tilde{K}^* = 2$ export position to $K^* = 1$ showing that its trade network in this product is better and broader than the one of Saudi Arabia which is at the first export position $\tilde{K}^* = 1$ in money volume. Iran moves in opposite direction from $\tilde{K}^* = 5$ money position down to $K^* = 14$ showing that its trade

network is restricted to a small number of nearby countries. A significant improvement of ranking takes place for Kazakhstan moving up from $\tilde{K}^* = 12$ to $K^* = 2$. The direct analysis shows that this happens due to an unusual fact that Kazakhstan is practically the only country which sells *crude petroleum* to the CheiRank leader in this product Russia. This puts Kazakhstan on the second position. It is clear that such direction of trade is more of political or geographical origin and is not based on economic reasons.

The same detailed analysis can be applied to all specific products given by SITC1. For example for trade of *cars* France goes up from $\tilde{K}^* = 7$ position in exports to $K^* = 3$ due to its broad export network.

C. Ranking time evolution and crises

The WTN has evolved during the period 1962 - 2009. The number of countries is increased by 38%, while the number of links per country for *all commodities* is increased in total by 140% with a significant increase from 50% to 140% during the period 1993 - 2009 corresponding to economy globalization. At the same time for a specific commodity the average number of links per country remains on a level of 3-5 links being by a factor 30 smaller compared to *all commodities* trade. During the whole period the total amount M_T of trade in *USD* shows an average exponential growth by 2 orders of magnitude.

A statistical density distribution of countries in the plane $(K^* - K, K^* + K)$ in the period 1962 - 2009 for *all commodities* is shown in Fig. 36. The distribution has a form of *spindle* with maximum density at the vertical axis $K^* - K = 0$. We remind that good exporters are on the lower side of this axis at $K^* - K < 0$, while the good importers (bad exporters) are on the upper side at $K^* - K > 0$.

The evolution of the ranking of countries for *all commodities* reflects their economical changes. The countries that occupy top positions tend to move very little in their ranks and can be associated to a *solid phase*. On the other hand, the countries in the middle region of $K^* + K$ have a gas like phase with strong rank fluctuations.

Examples of ranking evolution K and K^* for Japan, France, Fed R of Germany and Germany, Great Britain, USA, and for Argentina, India, China, USSR and Russian Fed are shown in Fig. 37. It is interesting to note that sharp increases in K mark crises in 1991, 1998 for Russia and in 2001 for Argentina (import is reduced in period of crises). It is also visible that in recent years the solid phase is perturbed by entrance of new countries like China and India. Other regional or global crisis could be highlighted due to the big fluctuations in the evolution of ranks. For example, in the range $81 \leq K + K^* \leq 120$, during the period of 1992 - 1998 some financial crises as Black Wednesday, Mexico crisis, Asian crisis and Russian crisis are appreciated with this ranking evolution.

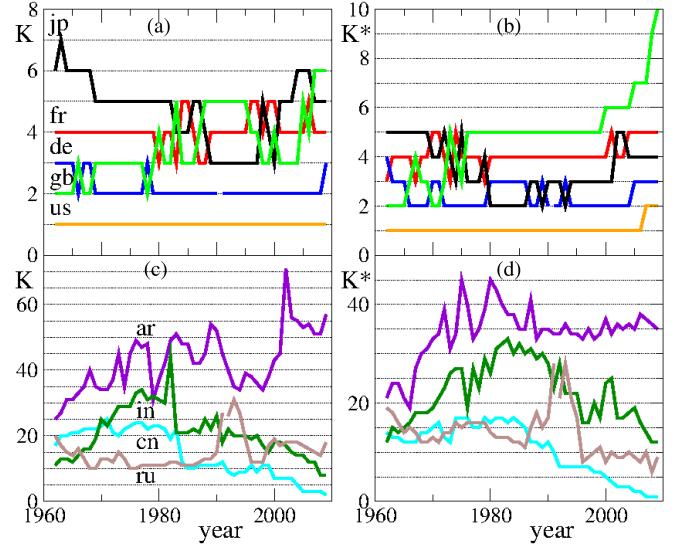


FIG. 37 (Color online) Time evolution of CheiRank and PageRank indexes K , K^* for some selected countries for *all commodities*. The countries shown panels (a) and (b) are: Japan (jp-black), France (fr-red), Fed R of Germany and Germany (de - both in blue), Great Britain (gb - green), USA (us - orange) [curves from top to bottom in 1962 in (a)]. The countries shown panels (c) and (d) are: Argentina (ar - violet), India (in - dark green), China (cn - cyan), USSR and Russian Fed (ru - both in gray) [curves from top to bottom in 1975 in (c)]. After (Ermann and Shepelyansky , 2011b).

D. Ecological ranking of world trade

Interesting parallels between multi-product world trade and interactions between species in ecological systems has been traced in (Ermann and Shepelyansky , 2013a). This approach is based on analysis of strength of transitions forming the Google matrix for the multi-product world trade network.

Ecological systems are characterized by high complexity and biodiversity (May, 2001) linked to nonlinear dynamics and chaos emerging in the process of their evolution (Lichtenberg and Lieberman, 1992). The interactions between species form a complex network whose properties can be analyzed by the modern methods of scale-free networks. The analysis of their properties uses a concept of mutualistic networks and provides a detailed understanding of their features being linked to a high nestedness of these networks (Bastolla *et al.*, 2009; Burgos *et al.*, 2007, 2008; Saverda *et al.*, 2011). Using the UN COMTRADE database we show that a similar ecological analysis gives a valuable description of the world trade: countries and trade products are analogous to plants and pollinators, and the whole trade network is characterized by a high nestedness typical for ecological networks.

An important feature of ecological networks is that they are highly structured, being very different from randomly interacting species (Bascompte *et al.*, 2003). Recently it has been shown that the mutualistic networks

between plants and their pollinators (Bascompte *et al.*, 2003; Memmott *et al.*, 2004; Olesen *et al.*, 2007; Rezende *et al.*, 2007; Vázquez and Aizen, 2004) are characterized by high nestedness which minimizes competition and increases biodiversity (Bastolla *et al.*, 2009; Burgos *et al.*, 2007, 2008; Saverda *et al.*, 2011).

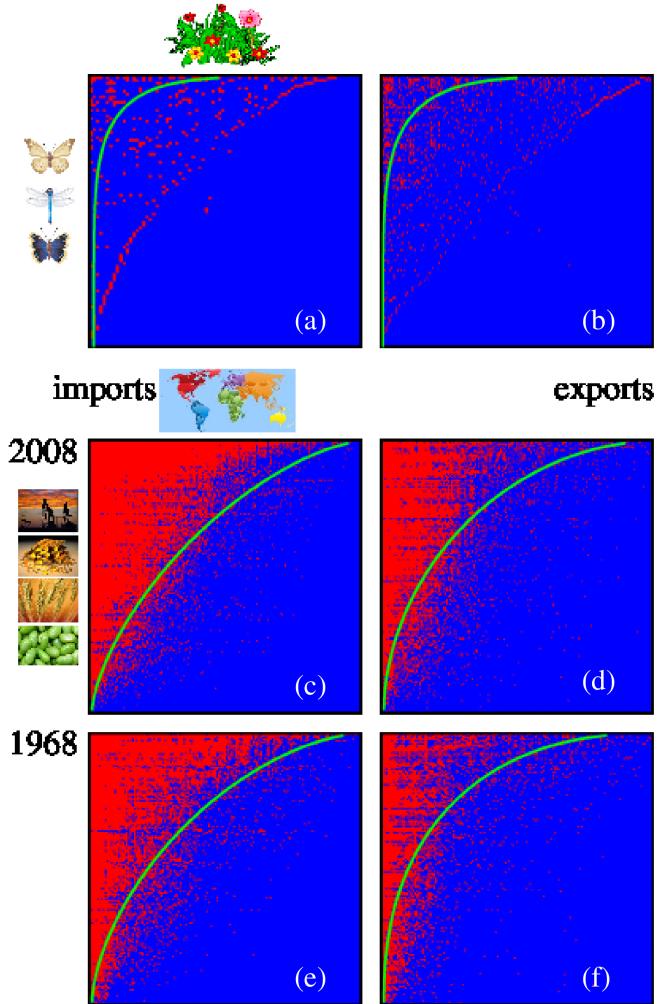


FIG. 38 (Color online) Nestedness matrices for the plant-animal mutualistic networks on top panels, and for the WTN of countries-products on middle and bottom panels. Panels (a) and (b) represent data of *ARR1* and *WES* networks from (Rezende *et al.*, 2007). The WTN matrices are computed with the threshold $\mu = 10^{-3}$ and corresponding $\varphi \approx 0.2$ for years 2008 (c,d) and 1968 (e,f) and 2008 for import (c,e) and export (d,f) panels. Red/gray and blue/black represent unit and zero elements respectively; only lines and columns with nonzero elements are shown. The order of plants-animals, countries-products is given by the nestedness algorithm (Rodríguez-Gironés *et al.*, 2006), the perfect nestedness is shown by green/gray curves for the corresponding values of φ . After (Ermann and Shepelyansky, 2013a).

The mutualistic WTN is constructed on the basis of the UN COMTRADE database from the matrix of trade transactions $M_{c',c}^p$ expressed in USD for a given prod-

uct (commodity) p from country c to country c' in a given year (from 1962 to 2009). For product classification we use 3-digits SITC Rev.1 discussed above with the number of products $N_p = 182$. All these products are described in (UN COMTRADE, 2011) in the commodity code document SITC Rev1. The number of countries varies between $N_c = 164$ in 1962 and $N_c = 227$ in 2009. The import and export trade matrices are defined as $M_{p,c}^{(i)} = \sum_{c'=1}^{N_c} M_{c,c'}^p$ and $M_{p,c}^{(e)} = \sum_{c'=1}^{N_c} M_{c',c}^p$ respectively. We use the dimensionless matrix elements $m^{(i)} = M^{(i)}/M_{max}$ and $m^{(e)} = M^{(e)}/M_{max}$ where for a given year $M_{max} = \max\{\max[M_{p,c}^{(i)}], \max[M_{p,c}^{(e)}]\}$. The distribution of matrix elements $m^{(i)}$, $m^{(e)}$ in the plane of indexes p and c , ordered by the total amount of import/export in a decreasing order, are shown and discussed in (Ermann and Shepelyansky, 2013a). In global, the distributions of $m^{(i)}$, $m^{(e)}$ remain stable in time especially in a view of 100 times growth of the total trade volume during the period 1962-2009. The fluctuations of $m^{(e)}$ are larger compared to $m^{(i)}$ case since certain products, e.g. petroleum, are exported by only a few countries while it is imported by almost all countries.

To use the methods of ecological analysis we construct the mutualistic network matrix for import $Q^{(i)}$ and export $Q^{(e)}$ whose matrix elements take binary value 1 or 0 if corresponding elements $m^{(i)}$ and $m^{(e)}$ are respectively larger or smaller than a certain trade threshold value μ . The fraction φ of nonzero matrix elements varies smoothly in the range $10^{-6} \leq \mu \leq 10^{-2}$ and the further analysis is not really sensitive to the actual μ value inside this broad range.

In contrast to ecological systems (Bastolla *et al.*, 2009) the world trade is described by a directed network and hence we characterize the system by two mutualistic matrices $Q^{(i)}$ and $Q^{(e)}$ corresponding to import and export. Using the standard nestedness BINMATNEST algorithm (Rodríguez-Gironés *et al.*, 2006) we determine the nestedness parameter η of the WTN and the related nestedness temperature $T = 100(1 - \eta)$. The algorithm re-orders lines and columns of a mutualistic matrix concentrating nonzero elements as much as possible in the top left corner and thus providing information about the role of immigration and extinction in an ecological system. A high level of nestedness and ordering can be reached only for systems with low T . It is argued that the nested architecture of real mutualistic networks increases their biodiversity.

The nestedness matrices generated by the BINMATNEST algorithm (Rodríguez-Gironés *et al.*, 2006) are shown in Fig. 38 for ecology networks *ARR1* ($N_{pl} = 84$, $N_{anim} = 101$, $\varphi = 0.043$, $T = 2.4$) and *WES* ($N_{pl} = 207$, $N_{anim} = 110$, $\varphi = 0.049$, $T = 3.2$) from (Rezende *et al.*, 2007). Using the same algorithm we generate the nestedness matrices of WTN using the mutualistic matrices for import $Q^{(i)}$ and export $Q^{(e)}$ for the WTN in years 1968 and 2008 using a fixed typical threshold $\mu = 10^{-3}$ (see Fig. 38). As for ecological systems, for the WTN data we also obtain rather small nestedness

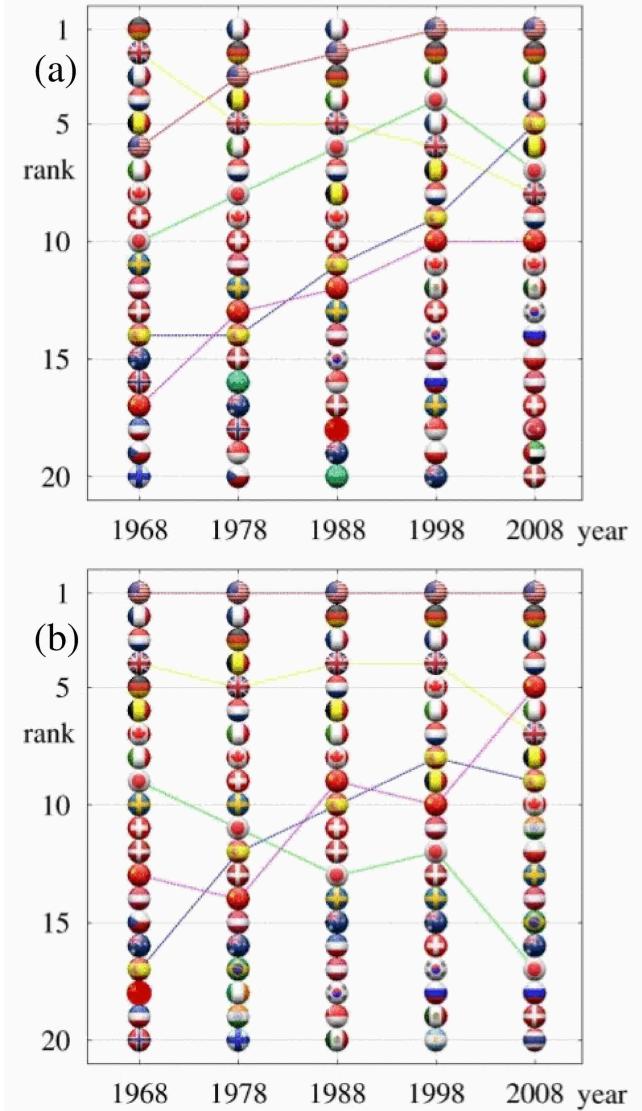


FIG. 39 (Color online) Top 20 EcoloRank countries as a function of years for the WTN import (a) and export (b) panels. The ranking is given by the nestedness algorithm for the trade threshold $\mu = 10^{-3}$; each country is represented by its corresponding flag. As an example, dashed lines show time evolution of the following countries: USA, UK, Japan, China, Spain. After (Ermann and Shepelyansky , 2013a).

temperature ($T \approx 6/8$ for import/export in 1968 and $T \approx 4/8$ in 2008 respectively). These values are by a factor 9/4 of times smaller than the corresponding T values for import/export from random generated networks with the corresponding values of φ .

The small value of nestedness temperature obtained for the WTN confirms the validity of the ecological analysis of WTN structure: trade products play the role of pollinators which produce exchange between world countries, which play the role of plants. Like in ecology the WTN evolves to the state with very low nestedness temperature that satisfies the ecological concept of system

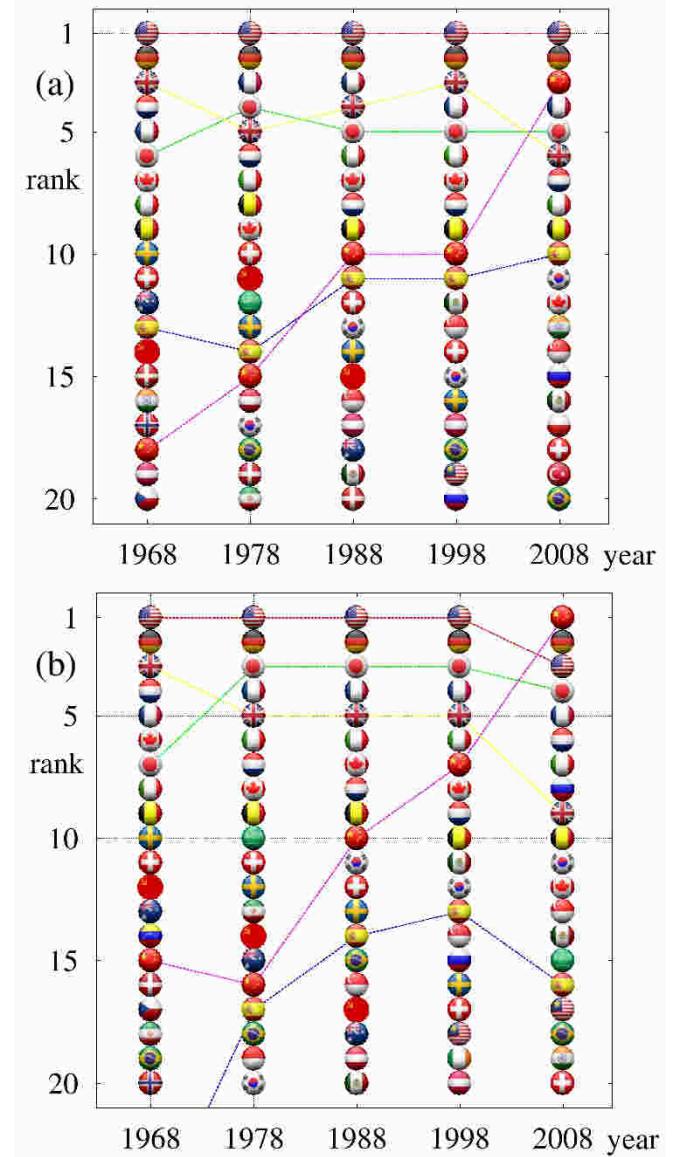


FIG. 40 (Color online) Top 20 countries as a function of years ranked by the total monetary trade volume of the WTN in import (a) and export (b) panels respectively; each country is represented by its corresponding flag. Dashed lines show time evolution of the same countries as in Fig. 39. After (Ermann and Shepelyansky , 2013a).

stability appearing as a result of high network nestedness (Bastolla *et al.*, 2009).

The nestedness algorithm creates effective ecological ranking (EcoloRanking) of all UN countries. The evolution of 20 top ranks throughout the years is shown in Fig. 39 for import and export. This ranking is quite different from the more commonly applied ranking of countries by their total import/export monetary trade volume (CIA, 2009) (see corresponding data in Fig. 40) or the democratic ranking of WTN based on the Google matrix

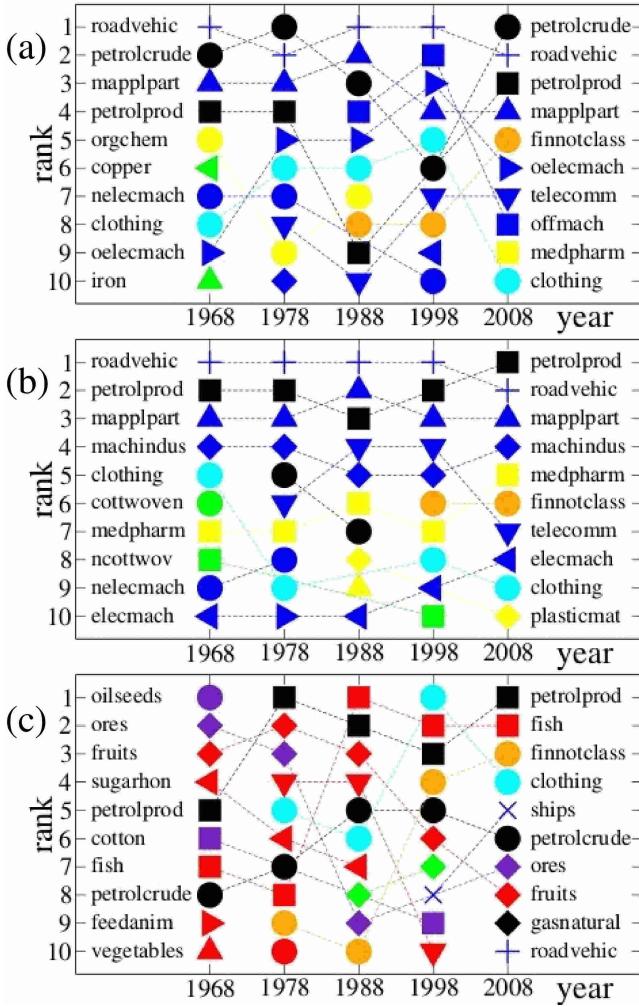


FIG. 41 (Color online) Top 10 ranks of trade products as a function of years for the WTN. Panel (a): ranking of products by monetary trade volume. Panels (b), (c): ranking is given by the nestedness algorithm for import (b) and export (c) with the trade threshold $\mu = 10^{-3}$. Each product is shown by its own symbol with short name written at years 1968, 2008; symbol color marks 1st SITC digit; SITC codes of products and their names are given in (UN COMTRADE, 2011) and Table 2 in (Ermann and Shepelyansky , 2013a). After (Ermann and Shepelyansky , 2013a).

analysis discussed above. Indeed, in 2008 China is at the top rank for total export volume but it is only at 5th position in EcoloRank (see Fig. 39, Fig. 40). In a similar way Japan moves down from 4th to 17th position while USA raises up from 3rd to 1st rank.

The same nestedness algorithm generates not only the ranking of countries but also the ranking of trade products for import and export which is presented in Fig. 41. For comparison we also show there the standard ranking of products by their trade volume. In Fig. 41 the color of symbol marks the 1st SITC digit described in figure, (UN COMTRADE, 2011) and Table 2 in (Ermann and

Shepelyansky , 2013a).

The origin of such a difference between EcoloRanking and trade volume ranking of countries is related to the main idea of mutualistic ranking in ecological systems: the nestedness ordering stresses the importance of mutualistic pollinators (products for WTN) which generate links and exchange between plants (countries for WTN). In this way generic products, which participate in the trade between many countries, become of primary importance even if their trade volume is not at the top lines of import or export. In fact such mutualistic products glue the skeleton of the world trade while the nestedness concept allows to rank them in order of their importance. The time evolution of this EcoloRanking of products of WTN is shown in Fig. 41 for import/export in comparison with the product ranking by the monetary trade volume (since the trade matrix is diagonal in product index the ranking of products in the latter case is the same for import/export). The top and middle panels have dominate colors corresponding to machinery (SITC Rev. 1 code 7; blue) and mineral fuels (3; black) with a moderate contribution of chemicals (5; yellow) and manufactured articles (8; cyan) and a small fraction of goods classified by material (6; green). Even if the global structure of product ranking by trade volume has certain similarities with import EcoloRanking there are also important new elements. Indeed, in 2008 the mutualistic significance of petroleum products (code 332), *machindus* (machines for special industries code 718) and *medpharm* (medical-pharmaceutic products code 541) is much higher compared to their volume ranking, while petroleum crude (code 331) and office machines (code 714) have smaller mutualistic significance compared to their volume ranking.

The new element of EcoloRanking is that it differentiates between import and export products while for trade volume they are ranked in the same way. Indeed, the dominant colors for export (Fig. 41 bottom panel) correspond to food (SITC Rev. 1 code 0; red) with contribution of black (present in import) and crude materials (code 2; violet); followed by cyan (present in import) and more pronounced presence of *finnotclass* (commodities/transactions not classified code 9; brown). EcoloRanking of export shows a clear decrease tendency of dominance of SITC codes 0 and 2 with time and increase of importance of codes 3,7. It is interesting to note that the code 332 of petroleum products is very vulnerable in volume ranking due to significant variations of petroleum prices but in EcoloRanking this product keeps the stable top positions in all years showing its mutualistic structural importance for the world trade. EcoloRanking of export shows also importance of fish (code 031), clothing (code 841) and fruits (code 051) which are placed on higher positions compared to their volume ranking. At the same time *roadvehic* (code 732), which are at top volume ranking, have relatively low ranking in export since only a few countries dominate the production of road vehicles.

It is interesting to note that in Fig. 41 petroleum crude is at the top of trade volume ranking e.g. in 2008 (top panel) but it is absent in import EcoloRanking (middle panel) and it is only on 6th position in export EcoloRanking (bottom panel). A similar feature is visible for years 1968, 1978. On a first glance this looks surprising but in fact for mutualistic EcoloRanking it is important that a given product is imported from top EcoloRank countries: this is definitely not the case for petroleum crude which practically is not produced inside top 10 import EcoloRank countries (the only exception is USA, which however also does not export much). Due to that reason this product has low mutualistic significance.

The mutualistic concept of product importance is at the origin of significant difference of EcoloRanking of countries compared to the usual trade volume ranking (see Fig. 39, Fig. 40). Indeed, in the latter case China and Japan are at the dominant positions but their trade is concentrated in specific products which mutualistic role is relatively low. In contrast USA, Germany and France keep top three EcoloRank positions during almost 40 years clearly demonstrating their mutualistic power and importance for the world trade.

Thus our results show the universal features of eco-logic ranking of complex networks with promising future applications to trade, finance and other areas.

E. Remarks on world trade and banking networks

The new approach to the world trade, based on the Google matrix analysis, gives a democratic type of ranking being independent of the trade amount of a given country. In this way rich and poor countries are treated on equal democratic grounds. In a certain sense PageRank probability for a given country is proportional to its rescaled import flows while CheiRank is proportional to its rescaled export flows inside of the WTN.

The global characteristics of the world trade are analyzed on the basis of this new type of ranking. Even if all countries are treated now on equal democratic grounds still we find at the top rank the group of industrially developed countries approximately corresponding to *G-20* and recover 74% of countries listed in *G-20*. The Google matrix analysis demonstrates an existence of two solid state domains of rich and poor countries which remain stable during the years of consideration. Other countries correspond to a gas phase with ranking strongly fluctuating in time. We propose a simple random matrix model which well describes the statistical properties of rank distribution for the WTN (Ermann and Shepelyansky , 2011b).

The comparison between usual ImportRank–ExportRank (see e.g. (CIA, 2009)) and our PageRank–CheiRank approach shows that the later highlights the trade flows in a new useful manner which is complementary to the usual analysis. The important difference between these two approaches is due to the fact

that ImportRank–ExportRank method takes into account only global amount of money exchange between a country and the rest of the world while PageRank–CheiRank approach takes into account all links and money flows between all countries.

The future developments should consider a matrix with all countries and all products which size becomes significantly larger ($N \sim 220 \times 10^4 \sim 2 \times 10^6$) comparing to a modest size $N \approx 227$ considered here. However, some new problems of this multiplex network analysis should be resolved combining a democracy in countries with volume importance of products which role is not democratic. It is quite possible that such an improved analysis will generate an asymmetric ranking of products in contrast to their symmetric ranking by volume in export and import. The ecological ranking of the WTN discussed in the previous SubSec. indicates preferences and asymmetry of trade in multiple products (Ermann and Shepelyansky , 2013a).

It is also important to note that usually in economy researchers analyze time evolution of various indexes studying their correlations. The results presented above for the WTN show that in addition to time evolution there is also evolution in space of the network. Like for waves in an ocean time and space are both important and we think that time and space study of trade captures important geographical factors which will play a dominant role for analysis of contamination propagation over the WTN in case of crisis. We think that the WTN data capture many essential elements which will play a rather similar role for financial flows in the interbank payment networks. We expect that the analysis of financial flows between bank units would prevent important financial crisis shaking the world in last years. Unfortunately, in contrast to WWW and UN COMTRADE, the banks keep hidden their financial flows. Due to this secrecy of banks the society is still suffering from financial crises. And all this for a network of very small size estimated on a level of 50 thousands bank units for the whole world being by a factor million smaller than the present size of WWW (e.g. Fed-wire interbank payment network of USA contains only 6600 nodes (Soramaki *et al.*, 2007)). In a drastic contrast with bank networks the WWW provided a public access to its nodes changing the world on a scale of 20 years. A creation of the World Bank Web (WBW) with information accessible for authorized investigators would allow to understand and control financial flows in an efficient manner preventing the society from bank crises. We note that the methods of network analysis and ranking start to attract interest of researchers in various banks (see e.g. (Craig and von Peter, 2010; Garratt *et al.*, 2011)).

XII. NETWORKS WITH NILPOTENT ADJACENCY MATRIX

A. General properties

In certain networks (Frahm *et al.*, 2012a, 2014b) it is possible to identify an ordering scheme for the nodes such that the adjacency matrix has non-vanishing elements A_{mn} only for nodes $m < n$ providing a triangular matrix structure. In these cases it is possible to provide a semi-analytical theory (Frahm *et al.*, 2012a, 2014b) which allows to simplify the numerical calculation of the non-vanishing eigenvalues of the matrix S introduced in Sec. III.A. It is useful to write this matrix in the form

$$S = S_0 + (1/N) e d^T \quad (10)$$

where the vector e has unit entries for all nodes and the *dangling vector* d has unit entries for dangling nodes and zero entries for the other nodes. The extra contribution $e d^T / N$ just replaces the empty columns (of S_0) with $1/N$ entries at each element. For a triangular network structure the matrix S_0 is nilpotent, i.e. $S_0^l = 0$ for some integer $l > 0$ and $S_0^{l-1} \neq 0$. Furthermore for the network examples studied previously (Frahm *et al.*, 2012a, 2014b) we have $l \ll N$ which has important consequences for the eigenvalue spectrum of S .

There are two groups of (right) eigenvectors ψ of S with eigenvalue λ . For the first group the quantity $C = d^T \psi$ vanishes and ψ is also an eigenvector of S_0 and if S_0 is nilpotent we have $\lambda = 0$ (there are also many higher order generalized eigenvectors associated to $\lambda = 0$). For the second group we have $C \neq 0$, $\lambda \neq 0$ and the eigenvector is given by $\psi = (\lambda \mathbb{1} - S_0)^{-1} C e / N$. Expanding the matrix inverse in a finite geometric series (for nilpotent S_0) and applying the condition $C = d^T \psi$ on this expression one finds that the eigenvalue must be a zero of the *reduced polynomial* of degree l :

$$\mathcal{P}_r(\lambda) = \lambda^l - \sum_{j=0}^{l-1} \lambda^{l-1-j} c_j = 0, \quad c_j = d^T S_0^j e / N. \quad (11)$$

This shows that there are at most l non-vanishing eigenvalues of S with eigenvectors $\psi \propto \sum_{j=0}^{l-1} \lambda^{-j-1} v^{(j)}$ where $v^{(j)} = S_0^j e / N$ for $j = 0, \dots, l-1$. Actually, the vectors $v^{(j)}$ generate an S -invariant l -dimensional subspace and from $S v^{(j)} = c_j v^{(0)} + v^{(j+1)}$ (using the identification $v^{(l)} = 0$) one obtains directly the $l \times l$ representation matrix \bar{S} of S with respect to $v^{(j)}$ (Frahm *et al.*, 2012a). Furthermore, the characteristic polynomial of \bar{S} is indeed given by the reduced polynomial (11) and the sum rule $\sum_{j=0}^{l-1} c_j = 1$ ensures that $\lambda = 1$ is indeed a zero of $\mathcal{P}_r(\lambda)$ (Frahm *et al.*, 2012a). The corresponding eigenvector (PageRank P at $\alpha = 1$) is given by $P \propto \sum_{j=0}^{l-1} v^{(j)}$. The remaining $N - l$ (generalized) eigenvectors of S are associated to many different Jordan blocks of S_0 for the eigenvalue $\lambda = 0$.

These l non-vanishing complex eigenvalues can be numerically computed as the zeros of the reduced polynomial by the Newton-Maehly method, by a numerical diagonalization of the “small” representation matrix \bar{S} (or better a more stable transformed matrix with identical eigenvalues) or by the Arnoldi method using the uniform vector e as initial vector. In the latter case the Arnoldi method should theoretically (in absence of rounding errors) exactly explore the l -dimensional subspace of the vectors $v^{(j)}$ and break off after l iterations with l exact eigenvalues.

However, numerical rounding errors may have a strong effect due to the Jordan blocks for the zero eigenvalue (Frahm *et al.*, 2012a). Indeed, an error ϵ appearing in a left bottom corner of a Jordan matrix of size D with zero eigenvalue leads to numerically induced eigenvalues on a complex circle of radius

$$|\lambda_\epsilon| = \epsilon^{1/D}. \quad (12)$$

Such an error can become significant with $|\lambda| > 0.1$ even for $\epsilon \sim 10^{-15}$ as soon as $D > 15$. We call this phenomenon the Jordan error enhancement. Furthermore, also the numerical determination of the zeros of $\mathcal{P}_r(\lambda)$ for large values of $l \sim 10^2$ can be numerically rather difficult. Thus, it may be necessary to use a high precision library such as the GNU GMP library either for the determination of the zeros of $\mathcal{P}_r(\lambda)$ or for the Arnoldi method (Frahm *et al.*, 2014b).

B. PageRank of integers

A network for integer numbers (Frahm *et al.*, 2012a) can be constructed by linking an integer number $n \in \{1, \dots, N\}$ to its divisors m different from 1 and n itself by an adjacency matrix $A_{mn} = M(n, m)$ where the multiplicity $M(n, m)$ is the number of times we can divide n by m , i.e. the largest integer such that $m^{M(n, m)}$ is a divisor of n , and $A_{mn} = 0$ for all other cases. The number 1 and the prime numbers are not linked to any other number and correspond to dangling nodes. The total size N of the matrix is fixed by the maximal considered integer. According to numerical data the number of links $N_\ell = \sum_{mn} A_{mn}$ is approximately given by $N_\ell = N(a_\ell + b_\ell \ln N)$ with $a_\ell = -0.901 \pm 0.018$, $b_\ell = 1.003 \pm 0.001$.

The matrix elements A_{mn} are different from zero only for $n \geq 2m$ and the associated matrix S_0 is therefore nilpotent with $S_0^l = 0$ and $l = [\log_2(N)] \ll N$. This triangular matrix structure can be seen in Fig. 42(a) which shows the amplitudes of S . The vertical green/gray lines correspond to the extra contribution due to the dangling nodes. These l non-vanishing eigenvalues of S can be efficiently calculated as the zeros of the reduced polynomial (11) up to $N = 10^9$ with $l = 29$. For $N = 10^9$ the largest eigenvalues are $\lambda_1 = 1$, $\lambda_{2,3} \approx -0.27178 \pm i 0.42736$, $\lambda_4 \approx -0.17734$ and $|\lambda_j| < 0.1$ for $j \geq 5$. The dependence of the eigenvalues on N seems to scale with

the parameter $1/\ln(N)$ for $N \rightarrow \infty$ and in particular $\gamma_2(N) = -2 \ln |\lambda_2(N)| \approx 1.020 + 7.14/\ln N$ (Frahm *et al.*, 2012a). Therefore the first eigenvalue is clearly separated from the second eigenvalue and one can chose the damping factor $\alpha = 1$ without any problems to define a unique PageRank.

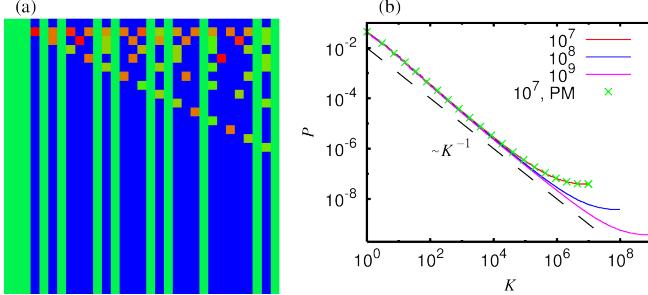


FIG. 42 (Color online) Panel (a): the Google matrix of integers, the amplitudes of matrix elements S_{mn} are shown by color with blue/black for minimal zero elements and red/gray for maximal unity elements, with $1 \leq n \leq 31$ corresponding to x -axis (with $n = 1$ corresponding to the left column) and $1 \leq m \leq 31$ for y -axis (with $m = 1$ corresponding to the upper row). Panel (b): the full lines correspond to the dependence of PageRank probability $P(K)$ on index K for the matrix sizes $N = 10^7, 10^8, 10^9$ with the PageRank evaluated by the exact expression $P \propto \sum_{j=0}^{l-1} v^{(j)}$. The green/gray crosses correspond to the PageRank obtained by the power method for $N = 10^7$; the dashed straight line shows the Zipf law dependence $P \sim 1/K$. After (Frahm *et al.*, 2012a).

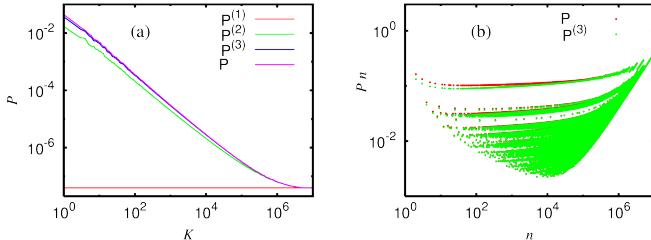


FIG. 43 (Color online) Panel (a): comparison of the first three PageRank approximations $P^{(i)} \propto \sum_{j=0}^{i-1} v^{(j)}$ for $i = 1, 2, 3$ and the exact PageRank dependence $P(K)$. Panel (b): comparison of the dependence of the rescaled probabilities nP and $nP^{(3)}$ on n . Both panels correspond to the case $N = 10^7$. After (Frahm *et al.*, 2012a).

The large values of N are possible because the vector iteration $v^{(j+1)} = S_0 v^{(j)}$ can actually be computed without storing the $N_\ell \sim N \ln N$ non-vanishing elements of S_0 by using the relation:

$$v_n^{(j+1)} = \sum_{m=2}^{[N/n]} \frac{M(mn, m)}{Q(mn)} v_{mn}^{(j)}, \quad \text{if } n \geq 2 \quad (13)$$

and $v_1^{(j+1)} = 0$ (Frahm *et al.*, 2012a). The initial vec-

tor is given by $v^{(0)} = e/N$ and $Q(n) = \sum_{m=2}^{n-1} M(m, n)$ is the number of divisors of n (taking into account the multiplicity). The multiplicity $M(mn, n)$ can be recalculated during each iteration and one needs only to store $N(\ll N_\ell)$ integer numbers $Q(n)$. It is also possible to reformulate (13) in a different way without using $M(mn, n)$ (Frahm *et al.*, 2012a). The vectors $v^{(j)}$ allow to compute the coefficients $c_j = d^T v^{(j)}$ in the reduced polynomial and the PageRank $P \propto \sum_{j=0}^{l-1} v^{(j)}$. Fig. 42(b) shows the PageRank for $N \in \{10^7, 10^8, 10^9\}$ obtained in this way and for comparison also the result of the power method for $N = 10^7$.

Actually Fig. 43 shows that in the sum $P \propto \sum_{j=0}^{l-1} v^{(j)}$ already the first three terms give a quite satisfactory approximation to the PageRank allowing a further analytical simplified evaluation (Frahm *et al.*, 2012a) with the result $P(n) \approx C_N/(b_n n)$ for $n \ll N$, where C_N is the normalization constant and $b_n = 2$ for prime numbers n and $b_n = 6 - \delta_{p_1, p_2}$ for numbers $n = p_1 p_2$ being a product of two prime numbers p_1 and p_2 . The behavior $P(n) n \approx C_N/b_n$, which takes approximately constant values on several branches, is also visible in Fig. 43 with C_N/b_n decreasing if n is a product of many prime numbers. The numerical results up to $N = 10^9$ show that the numbers n , corresponding to the leading PageRank values for $K = 1, 2, \dots, 32$, are $n = 2, 3, 5, 7, 4, 11, 13, 17, 6, 19, 9, 23, 29, 8, 31, 10, 37, 41, 43, 14, 47, 15, 53, 59, 61, 25, 67, 12, 71, 73, 22, 21$ with about 30% of non-primes among these values (Frahm *et al.*, 2012a).

A simplified model for the network for integer numbers with $M(n, m) = 1$ if m is divisor of n and $1 < m < n$ has also been studied with similar results (Frahm *et al.*, 2012a).

C. Citation network of Physical Review

Citation networks for Physical Review and other scientific journals can be defined by taking published articles as nodes and linking an article A to another article B if A cites B. PageRank and similar analysis of such networks are efficient to determine influential articles (Newman, 2001; Radicchi *et al.*, 2009; Redner, 1998, 2005).

In citation network links go mostly from newer to older articles and therefore such networks have, apart from the dangling node contributions, typically also a (nearly) triangular structure as can be seen in Fig. 44 which shows a coarse-grained density of the corresponding Google matrix for the citation network of Physical Review from the very beginning until 2009 (Frahm *et al.*, 2014b). However, due to the delay of the publication process in certain rare instances a published paper may cite another paper that is actually published a little later and sometimes two papers may even cite mutually each other. Therefore the matrix structure is not exactly triangular but in the coarse-grained density in Fig. 44 the rare “future citations” are not well visible.

The nearly triangular matrix structure implies large

dimensional Jordan blocks associated to the eigenvalue $\lambda = 0$. This creates the Jordan error enhancement (12) with severe numerical problems for accurate computation of eigenvalues in the range $|\lambda| < 0.3 - 0.4$ when using the Arnoldi method with standard double-precision arithmetic (Frahm *et al.*, 2014b).

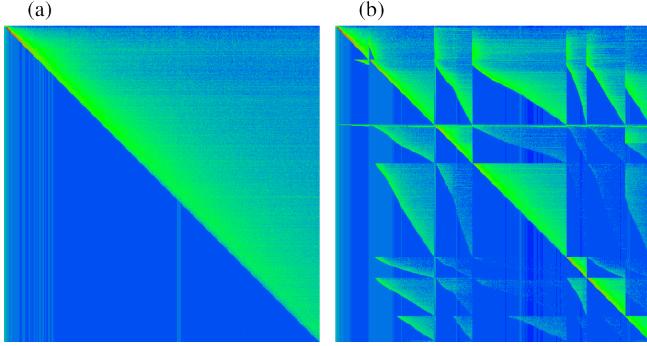


FIG. 44 (Color online) Different representations of the Google matrix structure for the Physical Review network until 2009. (a) Density of matrix elements $G_{tt'}$ in the basis of the publication time index t (and t'). (b) Density of matrix elements in the basis of journal ordering according to: Phys. Rev. Series I, Phys. Rev., Phys. Rev. Lett., Rev. Mod. Phys., Phys. Rev. A, B, C, D, E, Phys. Rev. STAB and Phys. Rev. STPER, and with time index ordering inside each journal. Note that the journals Phys. Rev. Series I, Phys. Rev. STAB and Phys. Rev. STPER are not clearly visible due to a small number of published papers. Also Rev. Mod. Phys. appears only as a thick line with 2-3 pixels (out of 500) due to a limited number of published papers. The different blocks with triangular structure correspond to clearly visible seven journals with considerable numbers of published papers. Both panels show the coarse-grained density of matrix elements on 500×500 square cells for the entire network. Color shows the density of matrix elements (of G at $\alpha = 1$) changing from blue/black for minimum zero value to red/gray at maximum value. After (Frahm *et al.*, 2014b).

One can eliminate the small number of future citations (12126 which is 0.26 % of the total number of links $N_\ell = 4691015$) and determine the complex eigenvalue spectrum of a triangular reduced citation network using the semi-analytical theory presented in previous subsection. It turns out that in this case the matrix S_0 is nilpotent $S_0^l = 0$ with $l = 352$ which is much smaller than the total network size $N = 463348$. The 352 non-vanishing eigenvalues can be determined numerically as the zeros of the polynomial (11) but due to an alternate sign problem with a strong loss of significance it is necessary to use the high precision library GMP with 256 binary digits (Frahm *et al.*, 2014b).

The semi-analytical theory can also be generalized to the case of *nearly* triangular networks, i.e. the full citation network including the future citations. In this case the matrix S_0 is no longer nilpotent but one can still generalize the arguments of previous subsection and discuss the two cases where the quantity $C = d^T \psi$ either van-

ishes (eigenvectors of first group) or is different from zero (eigenvectors of second group). The eigenvalues λ for the first group, which may now be different from zero, can be determined by a quite complicated but numerically very efficient procedure using the subspace eigenvalues of S and degenerate subspace eigenvalues of S_0 (due to absence of dangling node contributions the matrix S_0 produces much larger invariant subspaces than S) (Frahm *et al.*, 2014b). The eigenvalues of the second group are given as the complex zeros of the rational function:

$$\mathcal{R}(\lambda) = 1 - d^T \frac{\mathbb{1}}{\lambda \mathbb{1} - S_0} e/N = 1 - \sum_{j=0}^{\infty} c_j \lambda^{-1-j} \quad (14)$$

with c_j given as in (11) and now the series is not finite since S_0 is not nilpotent. For the citation network of Physical Review the coefficients c_j behave as $c_j \propto \rho_1^j$ where $\rho_1 \approx 0.902$ is the largest eigenvalue of the matrix S_0 with an eigenvector non-orthogonal to d . Therefore the series in (14) converges well for $|\lambda| > \rho_1$ but in order to determine the spectrum the rational function $\mathcal{R}(\lambda)$ needs to be evaluated for smaller values of $|\lambda|$. This problem can be solved by interpolating $\mathcal{R}(\lambda)$ with (another) rational function using a certain number of support points on the complex unit circle, where (14) converges very well, and determining the complex zeros, well inside the unit circle, of the numerator polynomial using again the high precision library GMP (Frahm *et al.*, 2014b). In this way using 16384 binary digits one may obtain 2500 reliable eigenvalues of the second group.

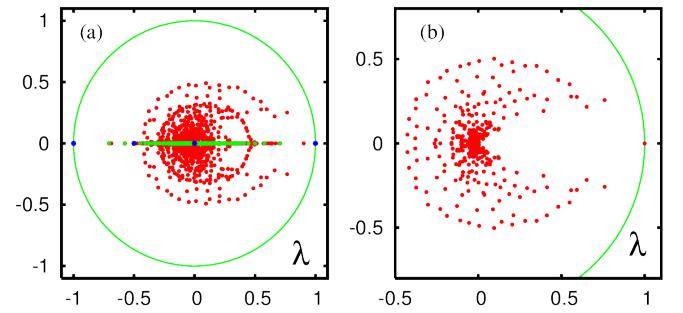


FIG. 45 (Color online) (a) Most accurate spectrum of eigenvalues for the full Physical Review network; red/gray dots represent the core space eigenvalues obtained by the rational interpolation method with the numerical precision of $p = 16384$ binary digits, $n_R = 2500$ eigenvalues; green (light gray) dots show the degenerate subspace eigenvalues of the matrix S_0 which are also eigenvalues of S with a degeneracy reduced by one (eigenvalues of the first group); blue/black dots show the direct subspace eigenvalues of S . (b) Spectrum of numerically accurate 352 non-vanishing eigenvalues of the Google matrix for the triangular reduced Physical Review network determined by the Newton-Maehly method applied to the reduced polynomial (11) with a high-precision calculation of 256 binary digits; note the absence of subspace eigenvalues for this case. In both panels the green/light gray curve represents the unit circle. After (Frahm *et al.*, 2014b).

The numerical high precision spectra obtained by the semi-analytic methods for both cases, triangular reduced and full citation network, are shown in Fig. 45. One may mention that it is also possible to implement the Arnoldi method using the high precision library GMP for both cases and the resulting eigenvalues coincide very accurately with the semi-analytic spectra for both cases (Frahm *et al.*, 2014b).

When the spectrum of G is determined with a good accuracy we can test the validity of the fractal Weyl law (5) changing the matrix size N_t by considering articles published from the beginning to a certain time moment t measured in years. The data presented in Fig. 46 show that the network size grows approximately exponentially as $N_t = 2^{(t-t_0)/\tau}$ with the fit parameters $t_0 = 1791$, $\tau = 11.4$. The time interval considered in Fig. 46 is $1913 \leq t \leq 2009$ since the first data point corresponds to $t = 1913$ with $N_t = 1500$ papers published between 1893 and 1913. The results, for the number N_λ of eigenvalues with $|\lambda_i| > \lambda$, show that its growth is well described by the relation $N_\lambda = a(N_t)^\nu$ for the range when the number of articles becomes sufficiently large $3 \times 10^4 \leq N_t < 5 \times 10^5$. This range is not very large and probably due to that there is a certain dependence of the exponent ν on the range parameter λ_c . At the same time we note that the maximal matrix size N studied here is probably the largest one used in numerical studies of the fractal Weyl law. We have $0.47 < \nu < 0.6$ for all $\lambda_c \geq 0.4$ that is definitely smaller than unity and thus the fractal Weyl law is well applicable to the Phys. Rev. network. The value of ν increases up to 0.7 for the data points with $\lambda_c < 0.4$ but this is due to the fact here N_λ also includes some numerically incorrect eigenvalues related to the numerical instability of the Arnoldi method at standard double-precision (52 binary digits) as discussed above.

We conclude that the most appropriate choice for the description of the data is obtained at $\lambda_c = 0.4$ which from one side excludes small, partly numerically incorrect, values of λ and on the other side gives sufficiently large values of N_λ . Here we have $\nu = 0.49 \pm 0.02$ corresponding to the fractal dimension $d = 0.98 \pm 0.04$. Furthermore, for $0.4 \leq \lambda_c \leq 0.7$ we have a rather constant value $\nu \approx 0.5$ with $d_f \approx 1.0$. Of course, it would be interesting to extend this analysis to a larger size N of citation networks of various type and not only for Phys. Rev. We expect that the fractal Weyl law is a generic feature of citation networks.

Further studies of the citation network of Physical Review concern the properties of eigenvectors (different from the PageRank) associated to relatively large complex eigenvalues, the fractal Weyl law, the correlations between PageRank and CheiRank (see also subsection IV.C) and the notion of “ImpactRank” (Frahm *et al.*, 2014b). To define the ImpactRank one may ask the question how a paper influences or has been influenced by other papers. For this one considers an initial vector v_0 , localized on a one node/paper. Then the modified Google

matrix $\tilde{G} = \gamma G + (1 - \gamma) v_0 e^T$ (with a damping factor $\gamma \sim 0.5 - 0.9$) produces a “PageRank” v_f by the propagator $v_f = (1 - \gamma)/(1 - \gamma G) v_0$. In the vector v_f the leading nodes/papers have strongly influenced the initial paper represented in v_0 . Doing the same for G^* one obtains a vector v_f^* where the leading papers have been influenced by the initial paper represented in v_0 . This procedure has been applied to certain historically important papers (Frahm *et al.*, 2014b).

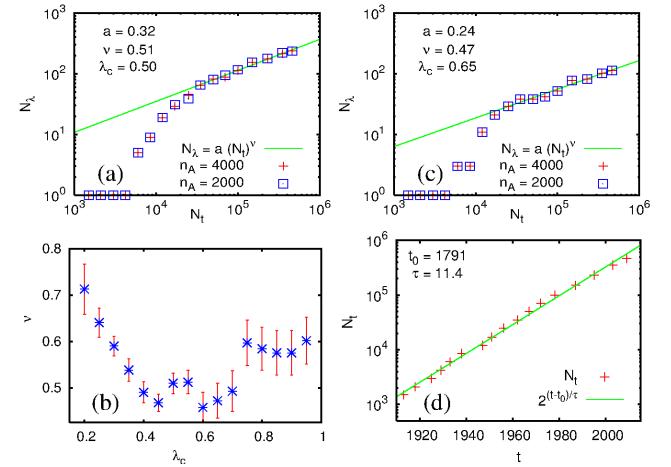


FIG. 46 (Color online) Data for the whole CNPR at different moments of time. Panel (a) (or (c)): shows the number N_λ of eigenvalues with $\lambda_c \leq \lambda \leq 1$ for $\lambda_c = 0.50$ (or $\lambda_c = 0.65$) versus the effective network size N_t where the nodes with publication times after a cut time t are removed from the network. The green/gray line shows the fractal Weyl law $N_\lambda = a(N_t)^\nu$ with parameters $a = 0.32 \pm 0.08$ ($a = 0.24 \pm 0.11$) and $\nu = 0.51 \pm 0.02$ ($\nu = 0.47 \pm 0.04$) obtained from a fit in the range $3 \times 10^4 \leq N_t < 5 \times 10^5$. The number N_λ includes both exactly determined invariant subspace eigenvalues and core space eigenvalues obtained from the Arnoldi method with double-precision (52 binary digits) for $n_A = 4000$ (red/gray crosses) and $n_A = 2000$ (blue/black squares). Panel (b): exponent ν with error bars obtained from the fit $N_\lambda = a(N_t)^\nu$ in the range $3 \times 10^4 \leq N_t < 5 \times 10^5$ versus cut value λ_c . Panel (d): effective network size N_t versus cut time t (in years). The green/gray line shows the exponential fit $2^{(t-t_0)/\tau}$ with $t_0 = 1791 \pm 3$ and $\tau = 11.4 \pm 0.2$ representing the number of years after which the size of the network (number of papers published in all Physical Review journals) is effectively doubled. After (Frahm *et al.*, 2014b).

In summary, the results of this section show that the phenomenon of the Jordan error enhancement (12), induced by finite accuracy of computations with a finite number of digits, can be resolved by advanced numerical methods described above. Thus the accurate eigenvalues λ can be obtained even for the most difficult case of quasi-triangular matrices. We note that for other networks like WWW of UK universities, Wikipedia and Twitter the triangular structure of S is much less pronounced (see e.g. Fig. 1) that gives a reduction of Jordan blocks so that the Arnoldi method with double precision computes

accurate values of λ .

XIII. RANDOM MATRIX MODELS OF MARKOV CHAINS

A. Albert-Barabási model of directed networks

There are various preferential attachment models generating complex scale-free networks (see e.g. (Albert and Barabási , 2002; Dorogovtsev, 2010)). Such undirected networks are generated by the Albert-Barabási (AB) procedure (Albert and Barabási , 2000) which builds networks by an iterative process. Such a procedure has been generalized to generate directed networks in (Giraud *et al.*, 2009) with the aim to study properties of the Google matrix of such networks. The procedure is working as follows: starting from m nodes, at each step m links are added to the existing network with probability p , or m links are rewired with probability q , or a new node with m links is added with probability $1 - p - q$. In each case the end node of new links is chosen with preferential attachment, i.e. with probability $(k_i + 1) / \sum_j (k_j + 1)$ where k_i is the total number of ingoing and outgoing links of node i . This mechanism generates directed networks having the small-world and scale-free properties, depending on the values of p and q . The results are averaged over N_r random realizations of the network to improve the statistics.

The studies (Giraud *et al.*, 2009) are done mainly for $m = 5$, $p = 0.2$ and two values of q corresponding to scale-free ($q = 0.1$) and exponential ($q = 0.7$) regimes of link distributions (see Fig. 1 in (Albert and Barabási , 2000) for undirected networks). For the generated directed networks at $q = 0.1$, one finds properties close to the behavior for the WWW with the cumulative distribution of ingoing links showing algebraic decay $P_c^{\text{in}}(k) \sim 1/k$ and average connectivity $\langle k \rangle \approx 6.4$. For $q = 0.7$ one finds $P_c^{\text{in}}(k) \sim \exp(-0.03k)$ and $\langle k \rangle \approx 15$. For outgoing links, the numerical data are compatible with an exponential decay in both cases with $P_c^{\text{out}}(k) \sim \exp(-0.6k)$ for $q = 0.1$ and $P_c^{\text{out}}(k) \sim \exp(-0.1k)$ for $q = 0.7$. It is found that small variations of parameters m, p, q near the chosen values do not qualitatively affect the properties of G matrix.

It is found that the eigenvalues of G for the AB model have one $\lambda = 1$ with all other $|\lambda_i| < 0.3$ at $\alpha = 0.85$ (see Fig. 1 in (Giraud *et al.*, 2009)). This distribution shows no significant modification with the growth of matrix size $2^{10} \leq N \leq 2^{14}$. However, the values of IPR ξ are growing with N for typical values $|\lambda| \sim 0.2$. This indicates a delocalization of corresponding eigenstates at large N . At the same time the PageRank probability is well described by the algebraic dependence $P \sim 1/K$ with ξ being practically independent of N .

These results for directed AB model network shows that it captures certain features of real directed networks, as e.g. a typical PageRank decay with the exponent

$\beta \approx 1$. However, the spectrum of G in this model is characterized by a large gap between $\lambda = 1$ and other eigenvalues which have $\lambda \leq 0.35$ at $\alpha = 1$. This feature is in a drastic difference with spectra of such typical networks at WWW of universities, Wikipedia and Twitter (see Figs. 17,22,32). In fact the AB model has no subspaces and no isolated or weakly coupled communities. In this network all sites can be reached from a given site in a logarithmic number of steps that generates a large gap in the spectrum of Google matrix and a rapid relaxation to PageRank eigenstate. In real networks there are plenty of isolated or weakly coupled communities and the introduction of damping factor $\alpha < 1$ is necessary to have a single PageRank eigenvalue at $\lambda = 1$. Thus the results obtained in (Giraud *et al.*, 2009) show that the AB model is not able to capture the important spectral features of real networks.

Additional studies in (Giraud *et al.*, 2009) analyzed the model of a real WWW university network with rewiring procedure of links, which consists in randomizing the links of the network keeping fixed the number of links at any given node. Starting from a single network, this creates an ensemble of randomized networks of same size, where each node has the same number of ingoing and outgoing links as for the original network. The spectrum of such randomly rewired networks is also characterized by a large gap in the spectrum of G showing that rewiring destroys the communities existing in original networks. The spectrum and eigenstate properties are studied in the related work on various real networks of moderate size $N < 2 \times 10^4$ which have no spectral gap (Georgeot *et al.*, 2010).

B. Random matrix models of directed networks

Above we saw that the standard models of scale-free networks are not able to reproduce the typical properties of spectrum of Google matrices of real large scale networks. At the same time we believe that it is important to find realistic matrix models of WWW and other networks. Here we discuss certain results for certain random matrix models of G .

Analytical and numerical studies of random unistochastic or orthostochastic matrices of size $N = 3$ and 4 lead to triplet and cross structures in the complex eigenvalue spectra (Zyczkowski *et al.*, 2003) (see also Fig. 18). However, the size of such matrices is too small.

Here we consider other examples of random matrix models of Perron-Frobenius operators characterized by non-negative matrix elements and column sums normalized to unity. We call these models Random Perron-Frobenius Matrices (RPFM). A number of RPFM, with arbitrary size N , can be constructed by drawing N^2 independent matrix elements $0 \leq G_{ij} \leq 1$ from a given distribution $p(G_{ij})$ with finite variance $\sigma^2 = \langle G_{ij}^2 \rangle - \langle G_{ij} \rangle^2$ and normalizing the column sums to unity (Frahm *et al.*, 2014b). The average matrix $\langle G_{ij} \rangle = 1/N$ is just a pro-

jector on the vector e (with unity entries on each node, see also Sec. XII.A) and has the two eigenvalues $\lambda_1 = 1$ (of multiplicity 1) and $\lambda_2 = 0$ (of multiplicity $N - 1$). Using an argument of degenerate perturbation theory on $\delta G = G - \langle G \rangle$ and known results on the eigenvalue density of non-symmetric random matrices (Akemann *et al.*, 2011; Guhr *et al.*, 1998; Mehta, 2004) one finds that an arbitrary realization of G has the leading eigenvalue $\lambda_1 = 1$ and the other eigenvalues are uniformly distributed on the complex unit circle of radius $R = \sqrt{N}\sigma$ (see Fig. 47).

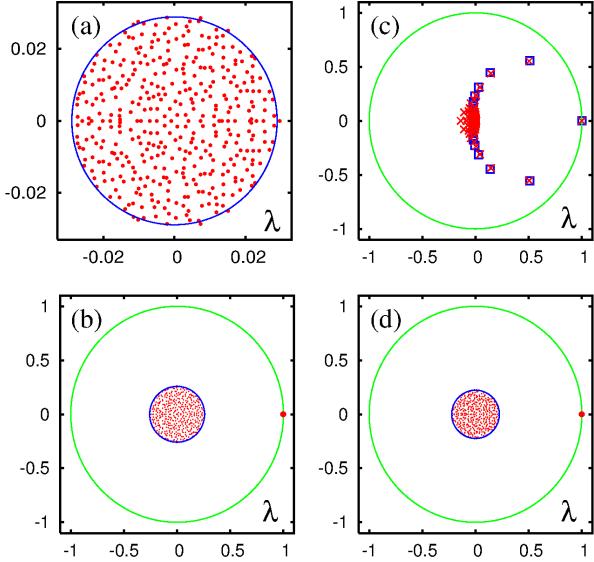


FIG. 47 (Color online) Panel (a) shows the spectrum (red/gray dots) of one realization of a full uniform RPBM with dimension $N = 400$ and matrix elements uniformly distributed in the interval $[0, 2/N]$; the blue/black circle represents the theoretical spectral border with radius $R = 1/\sqrt{3N} \approx 0.02887$. The unit eigenvalue $\lambda = 1$ is not shown due to the zoomed presentation range. Panel (c) shows the spectrum of one realization of triangular RPBM (red/gray crosses) with non-vanishing matrix elements uniformly distributed in the interval $[0, 2/(j-1)]$ and a triangular matrix with non-vanishing elements $1/(j-1)$ (blue/black squares); here $j = 2, 3, \dots, N$ is the index-number of non-empty columns and the first column with $j = 1$ corresponds to a dangling node with elements $1/N$ for both triangular cases. Panels (b), (d) show the complex eigenvalue spectrum (red/gray dots) of a sparse RPBM with dimension $N = 400$ and $Q = 20$ non-vanishing elements per column at random positions. Panel (b) (or (d)) corresponds to the case of uniformly distributed non-vanishing elements in the interval $[0, 2/Q]$ (constant non-vanishing elements being $1/Q$); the blue/black circle represents the theoretical spectral border with radius $R = 2/\sqrt{3Q} \approx 0.2582$ ($R = 1/\sqrt{Q} \approx 0.2236$). In panels (b), (d) $\lambda = 1$ is shown by a larger red dot for better visibility. The unit circle is shown by green/gray curve (panels (b), (c), (d)). After (Frahm *et al.*, 2014b).

Choosing different distributions $p(G_{ij})$ one obtains different variants of the model (Frahm *et al.*, 2014b), for example $R = 1/\sqrt{3N}$ using a full matrix with uniform

$G_{ij} \in [0, 2/N]$. Sparse models with $Q \ll N$ non-vanishing elements per column can be modeled by a distribution where the probability of $G_{ij} = 0$ is $1 - Q/N$ and for non-zero G_{ij} (either uniform in $[0, 2/Q]$ or constant $1/Q$) is Q/N leading to $R = 2/\sqrt{3Q}$ (for uniform non-zero elements) or $R = 1/\sqrt{Q}$ (for constant non-zero elements). The circular eigenvalue density with these values of R is also very well confirmed by numerical simulations in Fig. 47. Another case is a power law $p(G) = D/(1 + aG)^{-b}$ (for $0 \leq G \leq 1$) with D and a to be determined by normalization and the average $\langle G_{ij} \rangle = 1/N$. For $b > 3$ this case is similar to a full matrix with $R \sim 1/\sqrt{N}$. However for $2 < b < 3$ one finds that $R \sim N^{1-b/2}$.

The situation changes when one imposes a triangular structure on G in which case the complex spectrum of $\langle G \rangle$ is already quite complicated and, due to non-degenerate perturbation theory, close to the spectrum of G with modest fluctuations, mostly for the smallest eigenvalues (Frahm *et al.*, 2014b). Following the above discussion about triangular networks (with $G_{ij} = 0$ for $i \geq j$) we also study numerically a triangular RPBM where for $j \geq 2$ and $i < j$ the matrix elements G_{ij} are uniformly distributed in the interval $[0, 2/(j-1)]$ and for $i \geq j$ we have $G_{ij} = 0$. Then the first column is empty, that means it corresponds to a dangling node and it needs to be replaced by $1/N$ entries. For the triangular RPBM the situation changes completely since here the average matrix $\langle G_{ij} \rangle = 1/(j-1)$ (for $i < j$ and $j \geq 2$) has already a nontrivial structure and eigenvalue spectrum. Therefore the argument of degenerate perturbation theory which allowed to apply the results of standard full non-symmetric random matrices does not apply here. In Fig. 47 one clearly sees that for $N = 400$ the spectra for one realization of a triangular RPBM and its average are very similar for the eigenvalues with large modulus but both do not have at all a uniform circular density in contrast to the RPBM models without the triangular constraint discussed above. For the triangular RPBM the PageRank behaves as $P(K) \sim 1/K$ with the ranking index K being close to the natural order of nodes $\{1, 2, 3, \dots\}$ that reflects the fact that the node 1 has the maximum of $N - 1$ incoming links etc.

The above results show that it is not so simple to propose a good random matrix model which captures the generic spectral features of real directed networks. We think that investigations in this direction should be continued.

C. Anderson delocalization of PageRank?

The phenomenon of Anderson localization of electron transport in disordered materials (Anderson, 1958) is now a well-known effect studied in detail in physics (see e.g. (Evers and Mirlin, 2008)). In one and two dimensions even a small disorder leads to an exponential localization of electron diffusion that corresponds to an insu-

lating phase. Thus, even if a classical electron dynamics is diffusive and delocalized over the whole space, the effects of quantum interference generates a localization of all eigenstates of the Schrödinger equation. In higher dimensions a localization is preserved at a sufficiently strong disorder, while a delocalized metallic phase appears for a disorder strength being smaller a certain critical value dependent on the Fermi energy of electrons. This phenomenon is rather generic and we can expect that a somewhat similar delocalization transition can appear in the small-world networks.

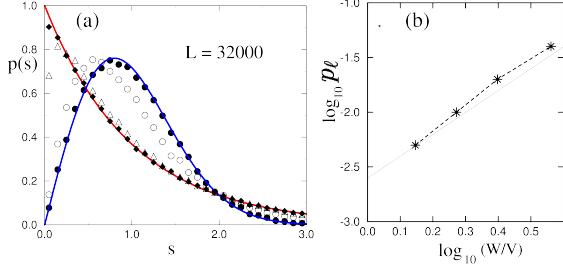


FIG. 48 (Color online) (a) The red/gray and blue/black curves represent the Poisson and Wigner surmise distributions. Diamonds, triangles, circles and black disks represent respectively the level spacing statistics $p(s)$ at $W/V = 4, 3, 2, 1; p_\ell = 0.02, L = 32000$; averaging is done over 60 network realizations. (b) Stars give dependence of p_ℓ on a disorder strength W/V at the critical point when $\eta_c(W, p_\ell) = 0.8$, and $p_\ell = 0.005, 0.01, 0.02, 0.04$ at fixed $L = 8000$; the straight line corresponds to $p_\ell = p_c = 1/4\ell_1 \approx (W/V)^2/400$; the dashed curve is drawn to adapt an eye. After (Chepelianskii and Shepelyansky , 2001).

Indeed, it is useful to consider the 1D Anderson model on a ring with a certain number of shortcut links, described by the Schrödinger equation

$$\epsilon_n \psi_n + V(\psi_{n+1} + \psi_{n-1}) + V \sum_S (\psi_{n+S} + \psi_{n-S}) = E \psi_n , \quad (15)$$

where ϵ_n are random on site energies homogeneously distributed within the interval $-W/2 \leq \epsilon_n \leq W/2$, and V is the hopping matrix element. The sum over S is taken over randomly established shortcuts from a site n to any other random site of the network. The number of such shortcuts is $S_{\text{tot}} = p_\ell L$, where L is the total number of sites on a ring and p_ℓ is the density of shortcut links. This model had been introduced in (Chepelianskii and Shepelyansky , 2001). The numerical study, reported there, showed that the level-spacing statistics $p(s)$ for this model has a transition from the Poisson distribution $p_{\text{Pois}}(s) = \exp(-s)$, typical for the Anderson localization phase, to the Wigner surmise distribution $p_{\text{Wig}}(s) = \pi s / 2 \exp(-\pi s^2 / 4)$, typical for the Anderson metallic phase (Evers and Mirlin , 2008; Guhr *et al.*, 1998). The numerical diagonalization was done via the Lanczos algorithm for the sizes up to $L = 32000$ and the typical parameter range $0.005 \leq p_\ell < 0.1$ and

$1 \leq W/V \leq 4$. An example, of the variation of $p_\ell(s)$ with a decrease of W/V is shown in Fig. 48(a). We see that the Wigner surmise provides a good description of the numerical data at $W/V = 1$, when the maximal localization length $\ell_1 \approx 96(V/W)^2 \approx 96$ in the 1D Anderson model (see e.g. (Evers and Mirlin , 2008)) is much smaller than the system size L .

To identify a transition from one limiting case $p_{\text{Pois}}(s)$ to another $p_{\text{Wig}}(s)$ it is convenient to introduce the parameter $\eta_s = \int_0^{s_0} (p(s) - p_{\text{Wig}}(s)) ds / \int_0^{s_0} (p_{\text{Pois}}(s) - p_{\text{Wig}}(s)) ds$, where $s_0 = 0.4729\dots$ is the intersection point of $p_{\text{Pois}}(s)$ and $p_{\text{Wig}}(s)$. In this way η_s varies from 1 (for $p(s) = p_{\text{Pois}}(s)$) to 0 (for $p(s) = p_{\text{Wig}}(s)$) (see e.g. (Shepelyansky , 2001)). From the variation of η_s with system parameters and size L , the critical density $p_\ell = p_c$ can be determined by the condition $\eta_s(p_c, W/V) = \eta_c = 0.8 = \text{const}$. being independent of L . The obtained dependence of p_c on W/V obtained at a fixed critical point $\eta_c = 0.8$ is shown in Fig. 48(b). The Anderson delocalization transition takes place when the density of shortcuts becomes larger than a critical density $p_\ell > p_c \approx 1/(4\ell_1)$ where $\ell_1 \approx 96(V/W)^2$ is the length of Anderson localization in 1D. A simple physical interpretation of this result is that the delocalization takes place when the localization length ℓ_1 becomes larger than a typical distance $1/(4p_\ell)$ between shortcuts. The further studies of time evolution of wave function $\psi_n(t)$ and IPR ξ variation also confirmed the existence of quantum delocalization transition on this quantum small-world network (Giraud *et al.*, 2005).

Thus the results obtained for the quantum small-world networks (Chepelianskii and Shepelyansky , 2001; Giraud *et al.*, 2005) show that the Anderson transition can take place in such systems. However, the above model represents an undirected network corresponding to a symmetric matrix with a real spectrum while the typical directed networks are characterized by asymmetric matrix G and complex spectrum. The possibility of existence of localized states of G for WWW networks was also discussed by (Perra *et al.*, 2009) but the fact that in a typical case the spectrum of G is complex has not been analyzed in detail.

Above we saw certain indications on a possibility of Anderson type delocalization transition for eigenstates of the G matrix. Our results clearly show that certain eigenstates in the core space are exponentially localized (see e.g. Fig 19(b)). Such states are localized only on a few nodes touching other nodes of network only by an exponentially small tail. A similar situation would appear in the 1D Anderson model if an absorption would be introduced on one end of the chain. Then the eigenstates located far away from this place would feel this absorption only by exponentially small tails so that the imaginary part of the eigenenergy would have for such far away states only an exponentially small imaginary part. It is natural to expect that such localization can be destroyed by some parameter variation. Indeed, certain eigenstates with $|\lambda| < 1$ for the directed network of the AB model have IPR ξ growing with the matrix

size N (see Sec. XIII.A and (Giraud *et al.*, 2009)) even if for the PageRank the values of ξ remain independent of N . The results for the Ulam network from Figs. 13, 14 provide an example of directed network where the PageRank vector becomes delocalized when the damping factor is decreased from $\alpha = 0.95$ to 0.85 (Zhirov *et al.*, 2010). This example demonstrates a possibility of PageRank delocalization but a deeper understanding of the conditions required for such a phenomenon to occur are still lacking. The main difficulty is an absence of well established random matrix models which have properties similar to the available examples of real networks.

Indeed, for Hermitian and unitary matrices the theories of random matrices, mesoscopic systems and quantum chaos allow to capture main universal properties of spectra and eigenstates (Akemann *et al.*, 2011; Evers and Mirlin , 2008; Guhr *et al.*, 1998; Haake, 2010; Mehta, 2004). For asymmetric Google matrices the spectrum is complex and at the moment there are no good random matrix models which would allow to perform analytical analysis of various parameter dependencies. It is possible that non-Hermitian Anderson models in 1D, which naturally generates a complex spectrum and may have delocalized eigenstates, will provide new insights in this direction (Goldsheid and Khoruzhenko , 1998).

XIV. OTHER EXAMPLES OF DIRECTED NETWORKS

In this section we discuss additional examples of real directed networks.

A. Brain neural networks

In 1958 John von Neumann traced first parallels between architecture of the computer and the brain (von Neumann , 1958). Since that time computers became an unavoidable element of the modern society forming a computer network connected by the WWW with about 4×10^9 indexed web pages spread all over the world (see e.g. <http://www.worldwidewebsize.com/>). This number starts to become comparable with 10^{10} neurons in a human brain where each neuron can be viewed as an independent processing unit connected with about 10^4 other neurons by synaptic links (see e.g. (Sporns , 2007)). About 20% of these links are unidirectional (Felleman and van Essen , 1991) and hence the brain can be viewed as a directed network of neuron links. At present, more and more experimental information about neurons and their links becomes available and the investigations of properties of neuronal networks attract an active interest (see e.g. (Bullmore and Sporns , 2009; Zuo *et al.*, 2012)). The fact that enormous sizes of WWW and brain networks are comparable gives an idea that the Google matrix analysis should find useful application in brain science as it is the case of WWW.

First applications of methods of Google matrix meth-

ods to brain neural networks was done in (Shepelyansky and Zhirov , 2010b) for a large-scale thalamocortical model (Izhikevich and Edelman , 2008) based on experimental measures in several mammalian species. The model spans three anatomic scales. (i) It is based on global (white-matter) thalamocortical anatomy obtained by means of diffusion tensor imaging of a human brain. (ii) It includes multiple thalamic nuclei and six-layered cortical microcircuitry based on *in vitro* labeling and three-dimensional reconstruction of single neurons of cat visual cortex. (iii) It has 22 basic types of neurons with appropriate laminar distribution of their branching dendritic trees. According to (Izhikevich and Edelman , 2008) the model exhibits behavioral regimes of normal brain activity that were not explicitly built-in but emerged spontaneously as the result of interactions among anatomical and dynamic processes.

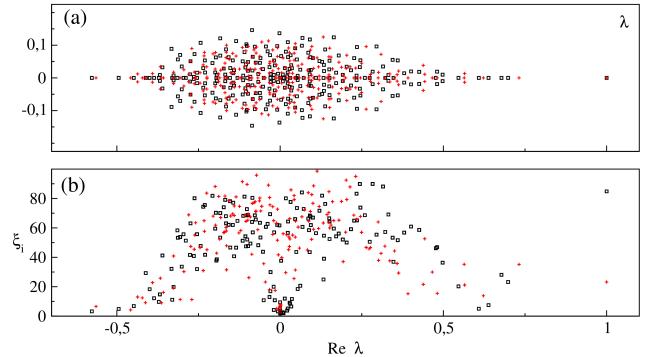


FIG. 49 (Color online) (a) Spectrum of eigenvalues λ for the Google matrices G and G^* at $\alpha = 0.85$ for the neural network of *C.elegans* (black and red/gray symbols). (b) Values of IPR ξ_i of eigenvectors ψ_i are shown as a function of corresponding $\text{Re } \lambda$ (same colors). After (Kandiah and Shepelyansky , 2014a).

The model studied in (Shepelyansky and Zhirov , 2010b) contains $N = 10^4$ neuron with $N_\ell = 1960108$. The obtained results show that PageRank and CheiRank vectors have rather large ξ being comparable with the whole network size at $\alpha = 0.85$. The corresponding probabilities have very flat dependence on their indexes showing that they are close to a delocalized regime. We attribute these features to a rather large number of links per node $\zeta \approx 196$ being even larger than for the Twitter network. At the same time the PageRank-CheiRank correlator is rather small $\kappa = -0.065$. Thus this network is structured in such a way that functions related to order signals (outgoing links of CheiRank) and signals bringing orders (ingoing links of PageRank) are well separated and independent of each other as it is the case for the Linux Kernel software architecture. The spectrum of G has a gapless structure showing that long living excitations can exist in this neuronal network.

Of course, model systems of neural networks can provide a number of interesting insights but it is much more

important to study examples of real neural networks. In (Kandiah and Shepelyansky , 2014a) such an analysis is performed for the neural network of *C.elegans* (worm). The full connectivity of this directed network is known and well documented at WormAtlas (Altun *et al.*, 2012). The number of linked neurons (nodes) is $N = 279$ with the number of synaptic connections and gap junctions (links) between them being $N_\ell = 2990$.

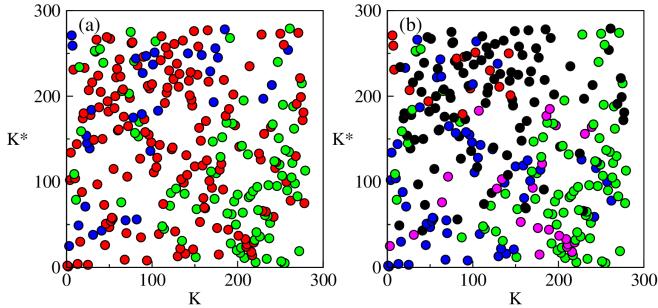


FIG. 50 (Color online) PageRank - CheiRank plane (K, K^*) showing distribution of neurons according to their ranking. (a): soma region coloration - head (red/gray), middle (green/light gray), tail (blue/dark gray). (b): neuron type coloration - sensory (red/gray), motor (green/light gray), interneuron (blue/dark gray), polymodal (purple/light-dark gray) and unknown (black). The classifications and colors are given according to WormAtlas (Altun *et al.*, 2012). After (Kandiah and Shepelyansky , 2014a).

The Google matrix G of *C.elegans* is constructed using the connectivity matrix elements $S_{ij} = S_{\text{syn},ij} + S_{\text{gap},ij}$, where S_{syn} is an asymmetric matrix of synaptic links whose elements are 1 if neuron j connects to neuron i through a chemical synaptic connection and 0 otherwise. The matrix part S_{gap} is a symmetric matrix describing gap junctions between pairs of cells, $S_{\text{gap},ij} = S_{\text{gap},ji} = 1$ if neurons i and j are connected through a gap junction and 0 otherwise. Then the matrices G and G^* are constructed following the standard rule (1) at $\alpha = 0.85$. The connectivity properties of this network are similar to those of WWW of Cambridge and Oxford with approximately the same number of links per node.

The spectra of G and G^* are shown in Fig. 49 with corresponding IPR values of eigenstates. The imaginary part of λ is relatively small $|\text{Im}(\lambda)| < 0.2$ due to a large fraction of symmetric links. The second by modulus eigenvalues are $\lambda_2 = 0.8214$ for G and $\lambda_2 = 0.8608$ for G^* . Thus the network relaxation time $\tau = 1/|\ln \lambda_2|$ is approximately 5, 6.7 iterations of G, G^* . Certain IPR values ξ_i of eigenstates of G, G^* have rather large $\xi \approx N/3$ while others have ξ located only on about ten nodes.

We have a large value $\xi \approx 85$ for PageRank and a more moderate value $\xi \approx 23$ for CheiRank vectors. Here we have the algebraic decay exponents being $\beta \approx 0.33$ for $P(K)$ and $\beta \approx 0.50$ for $P^*(K^*)$. Of course, the network size is not large and these values are only approximate. However, they indicate an interchange between PageRank and CheiRank showing importance of outgoing links.

It is possible that such an inversion is related to a significant importance of outgoing links in neural systems: in a sense such links transfer orders, while ingoing links bring instructions to a given neuron from other neurons. The correlator $\kappa = 0.125$ is small and thus, the network structure allows to perform a control of information flow in a more efficient way without interference of errors between orders and executions. We saw already in Sec. VII.A that such a separation of concerns emerges in software architecture. It seems that the neural networks also adopt such a structure.

We note that a somewhat similar situation appears for networks of Business Process Management where *Principals* of a company are located at the top CheiRank position while the top PageRank positions belong to company *Contacts* (Abel and Shepelyansky , 2011). Indeed, a case study of a real company structure analyzed in (Abel and Shepelyansky , 2011) also stress the importance of company managers who transfer orders to other structural units. For this network the correlator is also small being $\kappa = 0.164$. We expect that brain neural networks may have certain similarities with company organization.

Each neuron i belongs to two ranks K_i and K_i^* and it is convenient to represent the distribution of neurons on PageRank-CheiRank plane (K, K^*) shown in Fig. 50. The plot confirms that there are little correlations between both ranks since the points are scattered over the whole plane. Neurons ranked at top K positions of PageRank have their soma located mainly in both extremities of the worm (head and tail) showing that neurons in those regions have important connections coming from many other neurons which control head and tail movements. This tendency is even more visible for neurons at top K^* positions of CheiRank but with a preference for head and middle regions. In general, neurons, that have their soma in the middle region of the worm, are quite highly ranked in CheiRank but not in PageRank. The neurons located at the head region have top positions in CheiRank and also PageRank, while the middle region has some top CheiRank indexes but rather large indexes of PageRank (Fig. 50 (a)). The neuron type coloration (Fig. 50 (b)) also reveals that sensory neurons are at top PageRank positions but at rather large CheiRank indexes, whereas in general motor neurons are in the opposite situation.

Top nodes of PageRank and CheiRank favor important signal relaying neurons such as *AVA* and *AVB* that integrate signals from crucial nodes and in turn pilot other crucial nodes. Neurons *AVAL*, *AVAR*, *AVBL*, *AVBR* and *AVEL*, *AVER* are considered to belong to the rich club analyzed in (Towlson *et al.*, 2013). The top neurons in 2DRank are *AVAL*, *AVAR*, *AVBL*, *AVBR*, *PVCR* that corresponds to a dominance of interneurons. More details can be found in (Kandiah and Shepelyansky , 2014a).

The technological progress allows to obtain now more and more detailed information about neural networks (see e.g. (Bullmore and Sporns , 2009; Towlson *et al.*, 2013; Zuo *et al.*, 2012)) even if it is not easy to get infor-

mation about link directions. In view of that we expect that the methods of directed network analysis described here will find useful future applications for brain neural networks.

B. Google matrix of DNA sequences

The approaches of Markov chains and Google matrix can be also efficiently used for analysis of statistical properties of DNA sequences. The data sets are publicly available at (Ensemble Genome database, 2011). The analysis of Poincaré recurrences in these DNA sequences (Frahm and Shepelyansky , 2012c) shows their similarities with the statistical properties of recurrences for dynamical trajectories in the Chirikov standard map and other symplectic maps (Frahm and Shepelyansky , 2010). Indeed, a DNA sequence can be viewed as a long symbolic trajectory and hence, the Google matrix, constructed from it, highlights the statistical features of DNA from a new viewpoint.

An important step in the statistical analysis of DNA sequences was done in (Mantegna *et al.*, 1995) applying methods of statistical linguistics and determining the frequency of various words composed of up to 7 letters. A first order Markovian models have been also proposed and briefly discussed in this work. The Google matrix analysis provides a natural extension of this approach. Thus the PageRank eigenvector gives most frequent words of given length. The spectrum and eigenstates of G characterize the relaxation processes of different modes in the Markov process generated by a symbolic DNA sequence. Thus the comparison of word ranks of different species allows to identify their proximity.

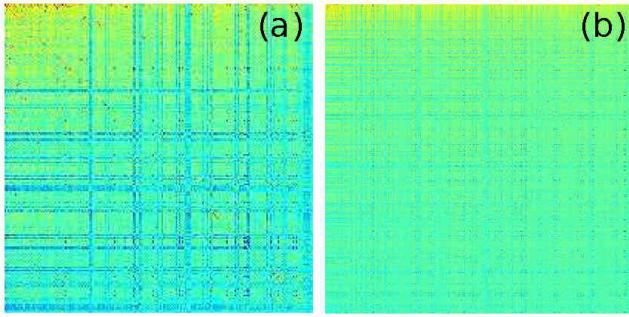


FIG. 51 (Color online) DNA Google matrix of *Homo sapiens* (HS) constructed for words of 6-letters length. Matrix elements $G_{KK'}$ are shown in the basis of PageRank index K (and K'). Here, x and y axes show K and K' within the range $1 \leq K, K' \leq 200$ (a) and $1 \leq K, K' \leq 1000$ (b). The element G_{11} at $K = K' = 1$ is placed at top left corner. Color marks the amplitude of matrix elements changing from blue/black for minimum zero value to red/gray at maximum value. After (Kandiah and Shepelyansky , 2013).

The statistical analysis is done for DNA sequences of the species: *Homo sapiens* (HS, human), *Canis familiaris*

(CF, dog), *Loxodonta africana* (LA, elephant), *Bos Taurus* (bull, BT), *Danio rerio* (DR, zebrafish) (Kandiah and Shepelyansky , 2013). For HS DNA sequences are represented as a single string of length $L \approx 1.5 \cdot 10^{10}$ base pairs (bp) corresponding to 5 individuals. Similar data are obtained for BT ($2.9 \cdot 10^9$ bp), CF ($2.5 \cdot 10^9$ bp), LA ($3.1 \cdot 10^9$ bp), DR ($1.4 \cdot 10^9$ bp). All strings are composed of 4 letters A, G, C, T and undetermined letter N_l . The strings can be found from (Kandiah and Shepelyansky , 2013).

For a given sequence we fix the words W_k of m letters length corresponding to the number of states $N = 4^m$. We consider that there is a transition from a state j to state i inside this basis N when we move along the string from left to right going from a word W_k to a next word W_{k+1} . This transition adds one unit in the transition matrix element $T_{ij} \rightarrow T_{ij} + 1$. The words with letter N_l are omitted, the transitions are counted only between nearby words not separated by words with N_l . There are approximately $N_t \approx L/m$ such transitions for the whole length L since the fraction of undetermined letters N_l is small. Thus we have $N_t = \sum_{i,j=1}^N T_{ij}$. The Markov matrix of transitions S_{ij} is obtained by normalizing matrix elements in such a way that their sum in each column is equal to unity: $S_{ij} = T_{ij} / \sum_i T_{ij}$. If there are columns with all zero elements (dangling nodes) then zeros of such columns are replaced by $1/N$. Then the Google matrix G is constructed from S by the standard rule (1). It is found that the spectrum of G has a significant gap and a variation of α in a range (0.5, 1) does not affect significantly the PageRank probability. Thus all DNA results are shown at $\alpha = 1$.

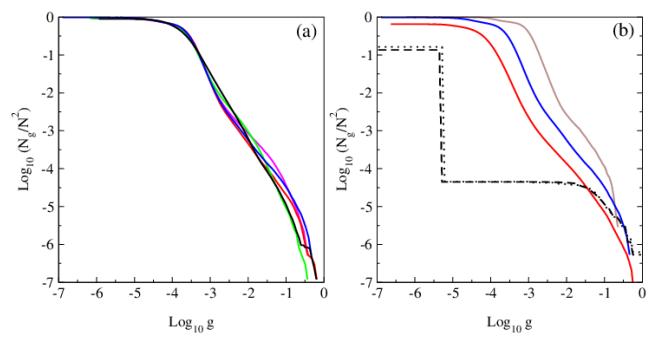


FIG. 52 (Color online) Integrated fraction N_g/N^2 of Google matrix elements with $G_{ij} > g$ as a function of g . (a) Various species with 6-letters word length: elephant LA (green), zebrafish DR (black), dog CF (red), bull BT (magenta), and *Homo sapiens* HS (blue) (from left to right at $y = -5.5$). (b) Data for HS sequence with words of length $m = 5$ (brown), 6 (blue), 7 (red) (from right to left at $y = -2$); for comparison black dashed and dotted curves show the same distribution for the WWW networks of Universities of Cambridge and Oxford in 2006 respectively. After (Kandiah and Shepelyansky , 2013).

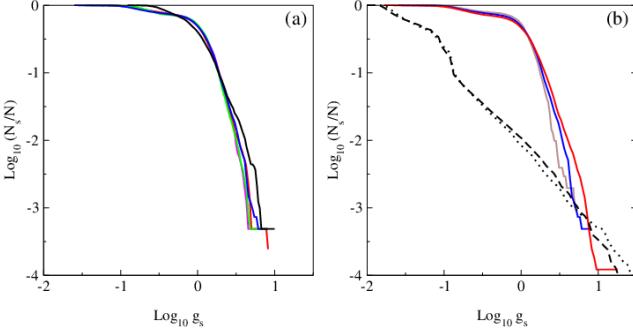


FIG. 53 (Color online) Integrated fraction N_s/N of sum of ingoing matrix elements with $\sum_{j=1}^N G_{i,j} \geq g_s$. Panels (a) and (b) show the same cases as in Fig. 52 in same colors. The dashed and dotted curves are shifted in x -axis by one unit left to fit the figure scale. After (Kandiah and Shepelyansky , 2013).

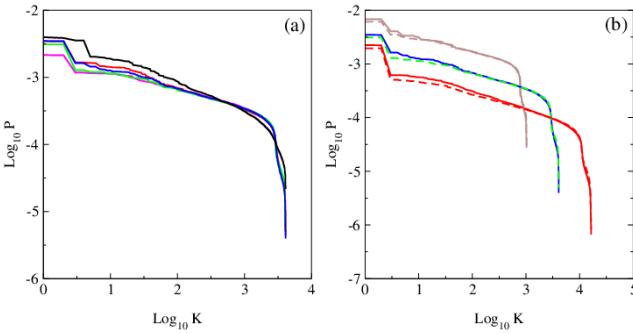


FIG. 54 (Color online) Dependence of PageRank probability $P(K)$ on PageRank index K . (a) Data for different species for word length of 6-letters: zebrafish DR (black), dog CF (red), Homo sapiens HS (blue), elephant LA (green) and bull BT (magenta) (from top to bottom at $x = 1$). (b) Data for HS (full curve) and LA (dashed curve) for word length $m = 5$ (brown), 6 (blue/green), 7 (red) (from top to bottom at $x = 1$). After (Kandiah and Shepelyansky , 2013).

The image of matrix elements $G_{KK'}$ is shown in Fig. 51 for HS with $m = 6$. We see that almost all matrix is full that is drastically different from the WWW and other networks considered above. The analysis of statistical properties of matrix elements G_{ij} shows that their integrated distribution follows a power law as it is seen in Fig. 52. Here N_g is the number of matrix elements of the matrix G with values $G_{ij} > g$. The data show that the number of nonzero matrix elements G_{ij} is very close to N^2 . The main fraction of elements has values $G_{ij} \leq 1/N$ (some elements $G_{ij} < 1/N$ since for certain j there are many transitions to some node i' with $T_{i'j} \gg N$ and e.g. only one transition to other i'' with $T_{i''j} = 1$). At the same time there are also transition elements G_{ij} with large values whose fraction decays in an algebraic law $N_g \approx AN/g^{\nu-1}$ with some constant A and an exponent ν . The fit of numerical data in the

range $-5.5 < \log_{10} g < -0.5$ of algebraic decay gives for $m = 6$: $\nu = 2.46 \pm 0.025$ (BT), 2.57 ± 0.025 (CF), 2.67 ± 0.022 (LA), 2.48 ± 0.024 (HS), 2.22 ± 0.04 (DR). For HS case we find $\nu = 2.68 \pm 0.038$ at $m = 5$ and $\nu = 2.43 \pm 0.02$ at $m = 7$ with the average $A \approx 0.003$ for $m = 5, 6, 7$. There are visible oscillations in the algebraic decay of N_g with g but in global we see that on average all species are well described by a universal decay law with the exponent $\nu \approx 2.5$. For comparison we also show the distribution N_g for the WWW networks of University of Cambridge and Oxford in year 2006. We see that in these cases the distribution N_g has a very short range in which the decay is at least approximately algebraic ($-5.5 < \log_{10}(N_g/N^2) < -6$). In contrast to that for the DNA sequences we have a large range of algebraic decay.

Since in each column we have the sum of all elements equal to unity we can say that the differential fraction $dN_g/dg \propto 1/g^\nu$ gives the distribution of outgoing matrix elements which is similar to the distribution of outgoing links extensively studied for the WWW networks. Indeed, for the WWW networks all links in a column are considered to have the same weight so that these matrix elements are given by an inverse number of outgoing links with the decay exponent $\nu \approx 2.7$. Thus, the obtained data show that the distribution of DNA matrix elements is similar to the distribution of outgoing links in the WWW networks. Indeed, for outgoing links of Cambridge and Oxford networks the fit of numerical data gives the exponents $\nu = 2.80 \pm 0.06$ (Cambridge) and 2.51 ± 0.04 (Oxford).

As discussed above, on average the probability of PageRank vector is proportional to the number of ingoing links that works satisfactory for sparse G matrices. For DNA we have a situation where the Google matrix is almost full and zero matrix elements are practically absent. In such a case an analogue of number of ingoing links is the sum of ingoing matrix elements $g_s = \sum_{j=1}^N G_{ij}$. The integrated distribution of ingoing matrix elements with the dependence of N_s on g_s is shown in Fig. 53. Here N_s is defined as the number of nodes with the sum of ingoing matrix elements being larger than g_s . A significant part of this dependence, corresponding to large values of g_s and determining the PageRank probability decay, is well described by a power law $N_s \approx BN/g_s^{\mu-1}$. The fit of data at $m = 6$ gives $\mu = 5.59 \pm 0.15$ (BT), 4.90 ± 0.08 (CF), 5.37 ± 0.07 (LA), 5.11 ± 0.12 (HS), 4.04 ± 0.06 (DR). For HS case at $m = 5, 7$ we find respectively $\mu = 5.86 \pm 0.14$ and 4.48 ± 0.08 . For HS and other species we have an average $B \approx 1$.

For WWW one usually have $\mu \approx 2.1$. Indeed, for the ingoing matrix elements of Cambridge and Oxford networks we find respectively the exponents $\mu = 2.12 \pm 0.03$ and 2.06 ± 0.02 (see curves in Fig. 53). For ingoing links distribution of Cambridge and Oxford networks we obtain respectively $\mu = 2.29 \pm 0.02$ and $\mu = 2.27 \pm 0.02$ which are close to the usual WWW value $\mu \approx 2.1$. In contrast the exponent μ for DNA Google matrix elements

gets significantly larger value $\mu \approx 5$. This feature marks a significant difference between DNA and WWW networks.

The PageRank vector can be obtained by a direct diagonalization. The dependence of probability P on index K is shown in Fig. 54 for various species and different word length m . The probability $P(K)$ describes the steady state of random walks on the Markov chain and thus it gives the frequency of appearance of various words of length m in the whole sequence L . The frequencies or probabilities of words appearance in the sequences have been obtained in (Mantegna *et al.*, 1995) by a direct counting of words along the sequence (the available sequences L were shorted at that times). Both methods are mathematically equivalent and indeed our distributions $P(K)$ are in good agreement with those found in (Mantegna *et al.*, 1995) even if now we have a significantly better statistics.

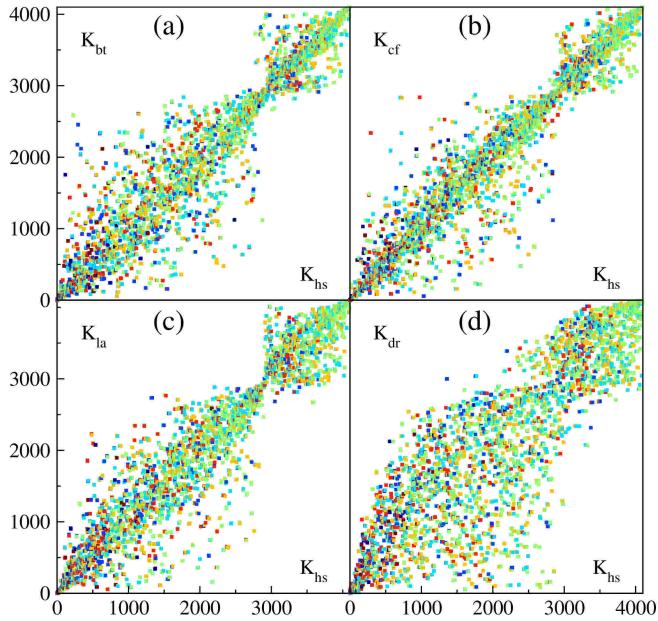


FIG. 55 (Color online) PageRank proximity $K - K$ plane diagrams for different species in comparison with *Homo sapiens*: (a) x -axis shows PageRank index $K_{hs}(i)$ of a word i and y -axis shows PageRank index of the same word i with $K_{bt}(i)$ of bull, (b) $K_{cf}(i)$ of dog, (c) $K_{la}(i)$ of elephant and (d) $K_{dr}(i)$ of zebrafish; here the word length is $m = 6$. The colors of symbols marks the purine content in a word i (fractions of letters A or G in any order); the color varies from red/gray at maximal content, via brown, yellow, green, light blue, to blue/black at minimal zero content. After (Kandiah and Shepelyansky , 2013).

The decay of P with K can be approximately described by a power law $P \sim 1/K^\beta$. Thus for example for HS sequence at $m = 7$ we find $\beta = 0.357 \pm 0.003$ for the fit range $1.5 \leq \log_{10} K \leq 3.7$ that is rather close to the exponent found in (Mantegna *et al.*, 1995). Since on average the PageRank probability is proportional to the number of ingoing links, or the sum of ingoing matrix

elements of G , one has the relation between the exponent of PageRank β and exponent of ingoing links (or matrix elements): $\beta = 1/(\mu - 1)$. Indeed, for the HS DNA case at $m = 7$ we have $\mu = 4.48$ that gives $\beta = 0.29$ being close to the above value of $\beta = 0.357$ obtained from the direct fit of $P(K)$ dependence. The agreement is not so perfect since there is a visible curvature in the log-log plot of N_s vs g_s and also since a small value of β gives a moderate variation of P that produces a reduction of accuracy of numerical fit procedure. In spite of this only approximate agreement we conclude that in global the relation between β and μ works correctly.

It is interesting to plot a PageRank index $K_s(i)$ of a given species s versus the index $K_{hs}(i)$ of HS for the same word i . For identical sequences one should have all points on diagonal, while the deviations from diagonal characterize the differences between species. The examples of such PageRank proximity $K - K$ diagrams are shown in Fig. 55 for words at $m = 6$. A visual impression is that CF case has less deviations from HS rank compared to BT and LA. The non-mammalian DR case has most strong deviations from HS rank.

The fraction of purine letters A or G in a word of $m = 6$ letters is shown by color in Fig. 55 for all words ranked by PageRank index K . We see that these letters are approximately homogeneously distributed over the whole range of K values. To determine the proximity between different species or different HS individuals we compute the average dispersion

$$\sigma(s_1, s_2) = \sqrt{\frac{1}{N} \sum_{i=1}^N (K_{s_1}(i) - K_{s_2}(i))^2} \quad (16)$$

between two species (individuals) s_1 and s_2 . Comparing the words with length $m = 5, 6, 7$ we find that the scaling $\sigma \propto N$ works with a good accuracy (about 10% when N is increased by a factor 16). To represent the result in a form independent of m we compare the values of σ with the corresponding random model value σ_{rnd} . This value is computed assuming a random distribution of N points in a square $N \times N$ when only one point appears in each column and each line (e.g. at $m = 6$ we have $\sigma_{rnd} \approx 1673$ and $\sigma_{rnd} \propto N$). The dimensionless dispersion is then given by $\zeta(s_1, s_2) = \sigma(s_1, s_2)/\sigma_{rnd}$. From the ranking of different species we obtain the following values at $m = 6$: $\zeta(CF, BT) = 0.308$; $\zeta(LA, BT) = 0.324$, $\zeta(LA, CF) = 0.303$; $\zeta(HS, BT) = 0.246$, $\zeta(HS, CF) = 0.206$, $\zeta(HS, LA) = 0.238$; $\zeta(DR, BT) = 0.425$, $\zeta(DR, CF) = 0.414$, $\zeta(DR, LA) = 0.422$, $\zeta(DR, HS) = 0.375$ (other m have similar values). According to this statistical analysis of PageRank proximity between species we find that ζ value is minimal between CF and HS showing that these are two most similar species among those considered here. The comparison of two HS individuals gives the value $\zeta(HS1, HS2) = 0.031$ being significantly smaller than the proximity correlator between different species (Kandiah and Shepelyansky , 2012).

The spectrum of G is analyzed in detail in (Kandiah

and Shepelyansky , 2012). It is shown that it has a relatively large gap due to which there is a relatively rapid relaxation of probability of a random surfer to the PageRank values.

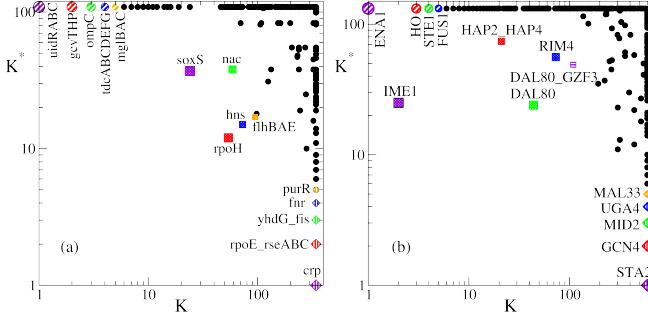


FIG. 56 (Color online) Distribution of nodes in the PageRank-CheiRank plane (K, K^*) for *Escherichia Coli* v1.1 (a), and Yeast (b) gene transcription networks (network data are taken from (Milo *et al.*, 2002; Shen-Orr *et al.*, 2002) and (Alon, 2014)). The nodes with five top probability values of PageRank, CheiRank and 2DRank are labeled by their corresponding operon (node) names; they correspond to 5 lowest values of indexes K, K_2, K^* . After (Ermann *et al.*, 2012a).

C. Gene regulation networks

At present the analysis of gene transcription regulation networks and recovery of their control biological functions becomes an active research field of bioinformatics (see e.g. (Milo *et al.*, 2002)). Here, following (Ermann *et al.*, 2012a), we provide two simple examples of 2DRanking analysis for gene transcriptional regulation networks of *Escherichia Coli* ($N = 423$, $N_\ell = 519$ (Shen-Orr *et al.*, 2002)) and Yeast ($N = 690$, $N_\ell = 1079$ (Milo *et al.*, 2002)). In the construction of G matrix the outgoing links to all nodes in each column are taken with the same weight, $\alpha = 0.85$.

The distribution of nodes in PageRank-CheiRank plane is shown in Fig. 56. The top 5 nodes, with their operon names, are given there for indexes of PageRank K , CheiRank K^* and 2DRank K_2 . This ranking selects operons with most high functionality in communication (K^*), popularity (K) and those that combines these both features (K_2). For these networks the correlator κ is close to zero ($\kappa = -0.0645$ for *Escherichia Coli* and $\kappa = -0.0497$ for Yeast, see Fig. 6)) that indicates the statistical independence between outgoing and ingoing links being quite similarly to the case of the PCN for the Linux Kernel. This may indicate that a slightly negative correlator κ is a generic property for the data flow network of control and regulation systems. A similar situation appears for networks of business process management and brain neural networks. Thus it is possible that the networks performing control functions are characterized in general by small correlator κ values. We expect that 2DRanking will find further useful applications for large scale gene regulation networks.

D. Networks of game go

The complexity of the well-known game go is such that no computer program has been able to beat a good player, in contrast with chess where world champions have been bested by game simulators. It is partly due to the fact that the total number of possible allowed positions in go is about 10^{171} , compared to e.g. only 10^{50} for chess (Tromp and Farnebäck , 2007).

It has been argued that the complex network analysis can give useful insights for a better understanding of this game. With this aim a network, modeling the game of go, has been defined by a statistical analysis of the data bases of several important historical professional and amateur Japanese go tournaments (Georgeot and Giraud , 2012). In this approach moves/nodes are defined as all possible patterns in 3×3 plaquettes on a go board of 19×19 intersections. Taking into account all possible obvious symmetry operations the number of non-equivalent moves is reduced to $N = 1107$. Moves which are close in space (typically a maximal distance of 4 intersections) are assumed to belong to the same tactical fight generating transitions on the network.

Using the historical data of many games, the transition probabilities between the nodes may be determined leading to a directed network with a finite size Perron-Frobenius operator which can be analyzed by tools of PageRank, CheiRank, complex eigenvalue spectrum, properties of certain selected eigenvectors and also certain other quantities (Georgeot and Giraud , 2012; Kandiah *et al.*, 2014b). The studies are done for plaquettes of different sizes with the corresponding network size changing from $N = 1107$ for plaquettes squares with 3×3 intersections up to maximal $N = 193995$ for diamond-shape plaquettes with 3×3 intersections plus the four at distance two from the center in the four directions left, right, top, down. It is shown that the PageRank leads to a frequency distribution of moves which obeys a Zipf law with exponents close to unity but this exponent may slightly vary if the network is constructed with shorter or longer sequences of successive moves. The important nodes in certain eigenvectors may correspond to certain strategies, such as protecting a stone and eigenvectors are also different between amateur and professional games. It is also found that the different phases of the game go are characterized by a different spectrum of the G matrix. The obtained results show that with the help of the Google matrix analysis it is possible to extract communities of moves which share some common properties.

The authors of these studies (Georgeot and Giraud , 2012; Kandiah *et al.*, 2014b) argue that the Google matrix analysis can find a number of interesting applications in the theory of games and the human decision-making processes.

E. Opinion formation on directed networks

Understanding the nature and origins of mass opinion formation is an outstanding challenge of democratic societies (Zaller, 1999). In the last few years the enormous development of such social networks as LiveJournal, Facebook, Twitter, and VKONTAKTE, with up to hundreds of millions of users, has demonstrated the growing influence of these networks on social and political life. The small-world scale-free structure of the social networks, combined with their rapid communication facilities, leads to a very fast information propagation over networks of electors, consumers, and citizens, making them very active on instantaneous social events. This invokes the need for new theoretical models which would allow one to understand the opinion formation process in modern society in the 21st century.

The important steps in the analysis of opinion formation have been done with the development of various voter models, described in great detail in (Castellano *et al.*, 2009; Krapivsky *et al.*, 2010). This research field became known as sociophysics (Galam, 1986, 2008). Here, following (Kandiah and Shepelyansky, 2012), we analyze the opinion formation process introducing several new aspects which take into account the generic features of social networks. First, we analyze the opinion formation on real directed networks such as WWW of Universities of Cambridge and Oxford (2006), Twitter (2009) and LiveJournal. This allows us to incorporate the correct scale-free network structure instead of unrealistic regular lattice networks, often considered in voter models. Second, we assume that the opinion at a given node is formed by the opinions of its linked neighbors weighted with the PageRank probability of these network nodes. The introduction of such a weight represents the reality of social networks where network nodes are characterized by the PageRank vector which provides a natural ranking of node importance, or elector or society member importance. In a certain sense, the top nodes of PageRank correspond to a political elite of the social network whose opinion influences the opinions of other members of the society (Zaller, 1999). Thus the proposed PageRank opinion formation (PROF) model takes into account the situation in which an opinion of an influential friend from high ranks of the society counts more than an opinion of a friend from a lower society level. We argue that the PageRank probability is the most natural form of ranking of society members. Indeed, the efficiency of PageRank rating had been well demonstrated for various types of scale-free networks.

The PROF model is defined in the following way. In agreement with the standard PageRank algorithm we determine the probability $P(K_i)$ for each node ordered by PageRank index K_i (using $\alpha = 0.85$). In addition, a network node i is characterized by an Ising spin variable σ_i which can take values $+1$ or -1 , coded also by red or blue color, respectively. The sign of a node i is determined by its direct neighbors j , which have PageRank probabilities

P_j . For that we compute the sum Σ_i over all directly linked neighbors j of node i :

$$\Sigma_i = a \sum_j (P_{j,\text{in}}^+ - P_{j,\text{in}}^-) + b \sum_j (P_{j,\text{out}}^+ - P_{j,\text{out}}^-), \quad a + b = 1, \quad (17)$$

where $P_{j,\text{in}}$ and $P_{j,\text{out}}$ denote the PageRank probability P_j of a node j pointing to node i (ingoing link) and a node j to which node i points to (outgoing link), respectively. Here, the two parameters a and b are used to tune the importance of ingoing and outgoing links with the imposed relation $a + b = 1$ ($0 \leq a, b \leq 1$). The values P^+ and P^- correspond to red and blue nodes, and the spin σ_i takes the value 1 or -1 , respectively, for $\Sigma_i > 0$ or $\Sigma_i < 0$. In a certain sense we can say that a large value of parameter b corresponds to a conformist society in which an elector i takes an opinion of other electors to which he/she points. In contrast, a large value of a corresponds to a tenacious society in which an elector i takes mainly the opinion of those electors who point to him/her. A standard random number generator is used to create an initial random distribution of spins σ_i on a given network. The time evolution then is determined by the relation (17) applied to each spin one by one. When all N spins are turned following (17) a time unit t is changed to $t \rightarrow t + 1$. Up to $N_r = 10^4$ random initial generations of spins are used to obtain statistically stable results. We present results for the number of red nodes since other nodes are blue.

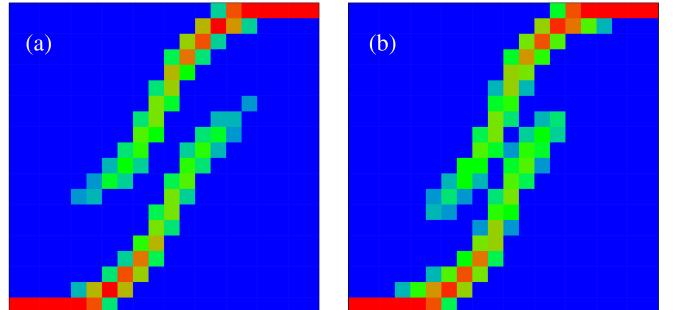


FIG. 57 (Color online) Density plot of probability W_f to find a final red fraction f_f , shown in y -axis, in dependence on an initial red fraction f_i , shown in x -axis; data are shown inside the unit square $0 \leq f_i, f_f \leq 1$. The values of W_f are defined as a relative number of realisations found inside each of 20×20 cells which cover the whole unit square. Here $N_r = 10^4$ realizations of randomly distributed colors are used to obtain W_f values; for each realization the time evolution is followed up to the convergence time with up to $t = 20$ iterations; (a) Cambridge network; (b) Oxford network at $a = 0.1$. The probability W_f is proportional to color changing from zero (blue/black) to unity (red/gray). After (Kandiah and Shepelyansky, 2012).

The main part of studies is done for the WWW of Cambridge and Oxford discussed above. We start with a random realization of a given fraction of red nodes $f_i = f(t = 0)$ which evolution in time converges to a

steady state with a final fraction of red nodes f_f approximated after time $t_c \approx 10$. However, different initial realisations with the same f_i value evolve to different final fractions f_f clearly showing a bistability phenomenon. To analyze how the final fraction of red nodes f_f depends on its initial fraction f_i , we study the time evolution $f(t)$ for a large number N_r of initial random realizations of colors following it up to the convergence time for each realization. We find that the final red nodes are homogeneously distributed in PageRank index K . Thus there is no specific preference for top society levels for an initial random distribution. The probability distribution W_f of final fractions f_f is shown in Fig. 57 as a function of initial fraction f_i at $a = 0.1$. The results show two main features of the model: a small fraction of red opinion is completely suppressed if $f_i < f_c$ and its larger fraction dominates completely for $f_i > 1 - f_c$; there is a bistability phase for the initial opinion range $f_b \leq f_i \leq 1 - f_b$. Of course, there is a symmetry in respect to exchange of red and blue colors. For the small value $a = 0.1$ we have $f_b \approx f_c$ with $f_c \approx 0.25$. For the larger value $a = 0.9$ we have $f_c \approx 0.35$, $f_b \approx 0.45$ (Kandiah and Shepelyansky , 2012).

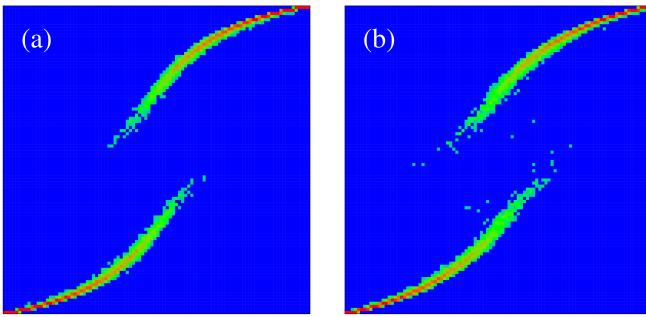


FIG. 58 (Color online) PROF-Sznajd model, option 1: density plot of probability W_f to find a final red fraction f_f , shown in y -axis, in dependence on an initial red fraction f_i , shown in x -axis; data are shown inside the unit square $0 \leq f_i, f_f \leq 1$. The values of W_f are defined as a relative number of realizations found inside each of 100×100 cells which cover the whole unit square. Here $N_r = 10^4$ realizations of randomly distributed colors are used to obtain W_f values; for each realization the time evolution is followed up the convergence time with up to $\tau = 10^7$ steps. (a) Cambridge network; (b) Oxford network; here $N_g = 8$. The probability W_f is proportional to color changing from zero (blue/black) to unity (red/gray). After (Kandiah and Shepelyansky , 2012).

Our interpretation of these results is the following. For small values of $a \ll 1$ the opinion of a given society member is determined mainly by the PageRank of neighbors to whom he/she points (outgoing links). The PageRank probability P of nodes to which many nodes point is usually high, since P is proportional to the number of ingoing links. Thus at $a \ll 1$ the society is composed of members who form their opinion by listening to an elite opinion. In such a society its elite with one color opinion can impose this opinion on a large fraction of the society.

Indeed, the direct analysis of the case, where the top $N_{top} = 2000$ nodes of PageRank index have the same red color, shows that this 1% of the society elite can impose its opinion to about 50% of the whole society at small a values (conformist society) while at large a values (tenacious society) this fraction drops significantly (see Fig.4 in (Kandiah and Shepelyansky , 2012)). We attribute this to the fact that in Fig. 57 we start with a randomly distributed opinion, since the opinion of the elite has two fractions of two colors this creates a bistable situation when the two fractions of society follow the opinions of this divided elite, which makes the situation bistable on a larger interval of f_i compared to the case of a tenacious society at $a \rightarrow 1$. When we replace in (17) P by 1 then the bistability disappears.

However, the detailed understanding of the opinion formation on directed networks still waits its development. Indeed, the results of PROF model for the LiveJournal and Twitted networks show that the bistability in these networks practically disappears. Also e.g. for the Twitter network studied in Sec. X.A, the elite of $N_{top} = 35000$ (about 0.1% of the whole society) can impose its opinion to 80% of the society at small $a < 0.15$ and to about 30% for $a > 0.15$ (Kandiah and Shepelyansky , 2012). It is possible that a large number of links between top PageRank nodes in Twitter creates a stronger tendency to a totalitarian opinion formation comparing to the case of University networks. At the same time the studies of opinion formation with the PROF model on the Ulam networks (Chakhmakhchyan and Shepelyansky , 2013), which have not very large number of links, show practically no bistability in opinion formation. It is expected that a small number of loops is at the origin of such a difference in respect to university networks.

Finally we discuss a more generic version of opinion formation called the PROF-Sznajd model (Kandiah and Shepelyansky , 2012). Indeed, we see that in the PROF model on university network opinions of small groups of red nodes with $f_i < f_c$ are completely suppressed that seems to be not very realistic. In fact, the Sznajd model (Sznajd-Weron and Sznajd , 2000) features the idea of resistant groups of a society and thus incorporates a well-known trade union principle “United we stand, divided we fall”. Usually the Sznajd model is studied on regular lattices. Its generalization for directed networks is done on the basis of the notion of group of nodes N_g at each discrete time step τ .

The evolution of group is defined by the following rules:

(a) we pick in the network by random a node i and consider the polarization of $N_g - 1$ highest PageRank nodes pointing to it;

(b) if node i and all other $N_g - 1$ nodes have the same color (same polarization), then these N_g nodes form a group whose effective PageRank value is the sum of all the member values $P_g = \sum_{j=1}^{N_g} P_j$;

(c) consider all the nodes pointing to any member of the group and check all these nodes n directly linked to the group: if an individual node PageRank value P_n is

less than the defined above P_g , the node joins the group by taking the same color (polarization) as the group nodes and increase P_g by the value of P_n ; if it is not the case, a node is left unchanged.

The above time step is repeated many times during time τ , counting the number of steps and choosing a random node i on each next step.

The time evolution of this PROF-Sznajd model converges to a steady state approximately after $\tau \approx 10N$ steps. This is compatible with the results obtained for the PROF model. However, the statistical fluctuations in the steady-state regime are present keeping the color distribution only on average. The dependence of the final fraction of red nodes f_f on its initial value f_i is shown by the density plot of probability W_f in Fig. 58 for the university networks. The probability W_f is obtained from many initial random realizations in a similar way to the case of Fig. 57. We see that there is a significant difference compared to the PROF model: now even at small values of f_i we find small but finite values of f_f , while in the PROF model the red color disappears at $f_i < f_c$. This feature is related to the essence of the Sznajd model: here, even small groups can resist against the totalitarian opinion. Other features of Fig. 58 are similar to those found for the PROF model: we again observe bistability of opinion formation. The number of nodes N_g , which form the group, does not significantly affect the distribution W_f (for studied $3 \leq N_g \leq 13$).

The above studies of opinion formation models on scale-free networks show that the society elite, corresponding to the top PageRank nodes, can impose its opinion on a significant fraction of the society. However, for a homogeneous distribution of two opinions, there exists a bistability range of opinions which depends on a conformist parameter characterizing the opinion formation. The proposed PROF-Sznajd model shows that totalitarian opinions can be escaped from by small subcommunities. The enormous development of social networks in the last few years definitely shows that the analysis of opinion formation on such networks requires further investigations.

XV. DISCUSSION

Above we considered many examples of real directed networks where the Google matrix analysis finds useful applications. The examples belong to various sciences varying from WWW, social and Wikipedia networks, software architecture to world trade, games, DNA sequences and Ulam networks. It is clear that the concept of Markov chains and Google matrix represents now the mathematical foundation of directed network analysis.

For Hermitian and unitary matrices there are now many universal concepts, developed in theoretical physics, so that the main properties of such matrices are well understood. Indeed, such characteristics as level spacing statistics, localization and delocalization prop-

erties of eigenstates, Anderson transition (Anderson, 1958), quantum chaos features can be now well handled by various theoretical methods (see e.g. (Akemann *et al.*, 2011; Evers and Mirlin, 2008; Guhr *et al.*, 1998; Haake, 2010; Mehta, 2004)). A number of generic models has been developed in this area allowing to understand the main effects via numerical simulations and analytical tools.

In contrast to the above case of Hermitian or unitary matrices, the studies of matrices of Markov chains of directed networks are now only at their initial stage. In this review, on examples of real networks we illustrated certain typical properties of such matrices. Among them there is the fractal Weyl law, which has certain traces in the field of quantum chaotic scattering, but the main part of features are new ones. In fact, the spectral properties of Markov chains had not been investigated on a large scale. We try here to provide an introduction to the properties of such matrices which contain all information about large scale directed networks. The Google matrix is like *The Library of Babel* (Borges, 1962), which contains everything. Unfortunately, we are still not able to find generic Markov matrix models which reproduce the main features of the real networks. Among them there is the possible spectral degeneracy at damping $\alpha = 1$, absence of spectral gap, algebraic decay of eigenvectors. Due to absence of such generic models it is still difficult to capture the main properties of real directed networks and to understand or predict their variations with a change of network parameters. At the moment the main part of real networks have an algebraic decay of PageRank vector with an exponent $\beta \approx 0.5 - 1$. However, certain examples of Ulam networks (see Figs. 13, 14) show that a delocalization of PageRank probability over the whole network can take place. Such a phenomenon looks to be similar to the Anderson transition for electrons in disordered solids. It is clear that if an Anderson delocalization of PageRank would took place, as a result of further developments of the WWW, the search engines based on the PageRank would loose their efficiency since the ranking would become very sensitive to various fluctuations. In a sense the whole world would go blind the day such a delocalization takes place. Due to that a better understanding of the fundamental properties of Google matrices and their dependencies on various system parameters have a high practical significance. We believe that the theoretical research in this direction should be actively continued. In many respects, as *the Library of Babel*, the Google matrix still keeps its secrets to be discovered by researchers from various fields of science. We hope that a further research will allow “*to formulate a general theory of the Library and solve satisfactorily the problem which no conjecture had deciphered: the formless and chaotic nature of almost all the books.*” (Borges, 1962)

XVI. ACKNOWLEDGMENTS

We are grateful to our colleagues M. Abel, A. D. Chepelianskii, Y.-H. Eom, B. Georgeot, O. Giraud, V. Kandiah, O. V. Zhirov for fruitful collaborations on the topics included in this review. We also thank our partners of the EC FET Open project NADINE A. Benczúr, N. Litvak, S. Vigna and colleague A. Kaltenbrunner for illuminating discussions. Our special thanks go to Debora Donato for her insights at our initial stage of this research.

Our research presented here is supported in part by the EC FET Open project “New tools and algorithms for directed network analysis” (NADINE No 288956). This work was granted access to the HPC resources of CALMIP (Toulouse) under the allocation 2012-P0110. We also thank the United Nations Statistics Division for provided help and friendly access to the UN COMTRADE database.

References

- Abel, M. W., and D. L. Shepelyansky, 2011, Eur. Phys. J. B **84**, 493.
- Akemann, G., J. Baik, and Ph. Di Francesco 2011, *The Oxford Handbook of Random Matrix Theory* (Oxford University Press, Oxford).
- Albert, R., and A.-L. Barabási, 2000, Phys. Rev. Lett. **85**, 5234.
- Albert, R., and A.-L. Barabási, 2002, Rev. Mod. Phys. **74**, 47.
- Alon, U., 2014, *U.Alon web site* <http://www.weizmann.ac.il/mcb/UriAlon/>.
- Altun, Z.F., L.A. Herndon, C. Crocker, R. Lints, and D.H. Hall (Eds.), 2012, WormAtlas <http://www.wormatlas.org>.
- Anderson, P. W., 1958, Phys. Rev. **109**, 1492.
- Aragón, P., D. Laniado, A. Kaltenbrunner, and Y. Volkovich, 2012, Proc. 8th WikiSym2012, ACM, New York **19**, arXiv:1204.3799v2[cs.SI].
- Arnoldi, W. E., 1951, Quart. Appl. Math. **9**, 17.
- M. Barigozzi, G. Fagiolo, and D. Garlaschelli, 2010, Phys. Rev. E **81**, 046104.
- Bascompte, J., P. Jordano, C.J. Melian, and J.M. Olesen, 2003, Proc. Nat. Acad. Sci. USA **100**, 9383.
- Bastolla, U., M. A. Pascual-Garcia, A. Ferrera, B. Luque, and J. Bascompte, 2009, Nature (London) **458**, 1018.
- Blank, M., G. Keller, and J. Liverani, 2002, Nonlinearity **15**, 1905.
- Bohigas, O., M.-J. Giannoni, and C. Schmit, 1984, Phys. Rev. Lett. **52**, 1.
- Borges, J.L., 1962, *The Library of Babel (Ficciones)* (Grove Press, N.Y.).
- Brin, S., and L. Page, 1998, Comp. Networks ISDN Syst. **30**, 107.
- Brin, M., and G. Stuck 2002, *Introduction to Dynamical Systems* (Cambridge University Press, Cambridge UK).
- Bruzda, W., M. Smaczyński, V. Cappellini, H.-J. Sommers, and K. Zyczkowski, 2010, Phys. Rev. E **81**, 066209.
- Bullmore, E., and O. Sporns, 2009, Nat. Rev. Neurosci. **10**, 312.
- Burgos, E., H. Ceva, R.P.J. Perazzo, M. Devoto, D. Medan, M. Zimmermann, and A.M. Delbue, 2007, J. Theor. Biol. **249**, 307.
- Burgos, E., H. Ceva, L. Hernández, R.P.J. Perazzo, M. Devoto, and D. Medan, 2008, Phys. Rev. E **78**, 046113.
- Caldarelli, G., 2003, *Scale-free networks* (Oxford Univ. Press, Oxford).
- Capocci, A., V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, 2006, Phys. Rev. E **74**, 036116.
- Castellano, C., S. Fortunato, and V. Loreto, 2009, Rev. Mod. Phys. **81**, 591.
- Chakhmakhchyan, L., and D. L. Shepelyansky, 2013, Phys. Lett. A **377**, 3119.
- Chepelianskii, A.D., and D. L. Shepelyansky, 2001, <http://www.quantware.ups-tlse.fr/talks-posters/chepelianskii2001.pdf>.
- Chepelianskii, A. D., 2010, arXiv:1003.5455[cs.SE].
- Chirikov, B. V., 1979, Phys. Rep. **52**, 263.
- Chirikov, B.V., and D. Shepelyansky, 2008, Scholarpedia **3(3)**, 3550.
- Central Intelligence Agency, 2009, *The CIA World Factbook 2010* (Skyhorse Publ. Inc.).
- Cornfeld, I.P., Fomin, S.V., and Y.G. Sinai 1982, *Ergodic Theory* (Springer, New York).
- Craig, B., and G. von Peter 2010, *Interbank tiering and money center bank* (Discussion paper N 12/2010, Deutsche Bundesbank)
- De Benedictis, L., and L. Tajoli 2011, The World Economy **34**, 8, 1417–1454.
- Dijkstra, E.W., 1982, *Selected Writing on Computing: a Personal Perspective* (Springer-Verlag, New York).
- Dimassi, M., and J. Sjöstrand 1999, *Spectral Asymptotics in the Semicalssical Limit* (Cambridge University Press, Cambridge).
- Donato, D., L. Laura, S. Leonardi, and S. Millozzi, 2004, Eur. Phys. J. B **38**, 239.
- Dorogovtsev, S. N., A. V. Goltsev, and J. F. F. Mendes, 2008, Rev. Mod. Phys. **80**, 1275.
- Dorogovtsev, S. 2010, *Lectures on Complex Networks* (Oxford University Press, Oxford).
- Ensemble Genome database, 2011, *Ensemble Genome database* <http://www.ensembl.org/>.
- Eom, Y.-H., and D. L. Shepelyansky, 2013a, PLoS ONE **8(10)**, e74554.
- Eom, Y.-H., K. M. Frahm, A. Benczúr, and D. L. Shepelyansky, 2013b, Eur. Phys. J. B **86**, 492.
- Eom, Y.-H., P. Aragón, D. Laniado, A. Kaltenbrunner, S. Vigna, and D. L. Shepelyansky, 2014, arXiv:1405.7183 [cs.SI] (submitted to PLoS ONE).
- Ermann, L., and D. L. Shepelyansky, 2010a, Phys. Rev. E **81**, 032621.
- Ermann, L., and D. L. Shepelyansky, 2010b, Eur. Phys. J. B **75**, 299.
- Ermann, L., A. D. Chepelianskii, and D. L. Shepelyansky, 2011a, Eur. Phys. J. B **79**, 115.
- Ermann, L., and D. L. Shepelyansky, 2011b, Acta Phys. Polonica A **120(6A)**, A158; <http://www.quantware.ups-tlse.fr/QWLIB/tradecheirank/>.
- Ermann, L., A. D. Chepelianskii, and D. L. Shepelyansky, 2012a, J. Phys. A: Math. Theor. **45**, 275101; <http://www.quantware.ups-tlse.fr/QWLIB/dvvadi/>.
- Ermann, L., and D. L. Shepelyansky, 2012b, Physica D **241**, 514.

- Ermann, L., and D. L. Shepelyansky, 2013a, Phys. Lett. A **377**, 250.
- Ermann, L., K. M. Frahm, and D. L. Shepelyansky, 2013b, Eur. Phys. J. B **86**, 193.
- Evers, F., and A. D. Mirlin, 2008, Rev. Mod. Phys. **80**, 1355.
- Felleman, D.J., and D. C. van Essen, 1991, Celeb. Cortex **1**, 1.
- FETNADINE database, 2014, *Quantware group*, <http://www.quantware.ups-tlse.fr/FETNADINE/datasets.htm>
- Fogaras, D., 2003, Lect. Not. Comp. Sci. **2877**, 65.
- Fortunato, S., 2010, Phys. Rep. **486**, 75.
- Frahm, K. M., and D. L. Shepelyansky, 2009, Phys. Rev. E **80**, 016210.
- Frahm, K. M., and D. L. Shepelyansky, 2010, Eur. Phys. J. B **76**, 57.
- Frahm, K. M., B. Georgeot, and D. L. Shepelyansky, 2011, J. Phys A: Math. Theor. **44**, 465101.
- Frahm, K. M., A. D. Chepelianskii, and D. L. Shepelyansky, 2012a, J. Phys A: Math. Theor. **45**, 405101.
- Frahm, K. M., and D. L. Shepelyansky, 2012b, Eur. Phys. J. B **85**, 355.
- Frahm, K. M., and D. L. Shepelyansky, 2012c, Phys. Rev. E **85**, 016214.
- Frahm, K. M., and D. L. Shepelyansky, 2013, Eur. Phys. J. B **86**, 322.
- Frahm, K. M., and D. L. Shepelyansky, 2014a, Eur. Phys. J. B **87**, 93.
- Frahm, K. M., Y.-H. Eom, and D. L. Shepelyansky, 2014b, Phys. Rev. E **89**, 052814.
- Franceschet, M., 2011, Communications of the ACM **54**(6), 92.
- Froyland, G., and K. Padberg 2009, Physica D **238**, 1507.
- Galam, S., 1986, J. Math. Psych. **30**, 426.
- Galam, S., 2008, Int. J. Mod. Phys. C **19**, 409.
- Gamow, G. A., 1928, Z. für Phys **51**, 204.
- Garlaschelli, D., and M. I. Loffredo 2005, Physica A: Stat. Mech. Appl. **355**, 138.
- Garratt, R.J., Mahadeva, L., and K. Svirydzenka 2011, *Mapping systemic risk in the international banking network* (Working paper N 413, Bank of England)
- Gaspard, P., 1998, *Chaos, Scattering and Statistical Mechanics* (Cambridge Univ. Press, Cambridge).
- Gaspard, P., 2014, Scholarpedia **9**(6), 9806.
- Georgeot, B., O. Giraud, and D. L. Shepelyansky, 2010, Phys. Rev. E **81**, 056109.
- Georgeot, B., and O. Giraud, 2012, Eurphys. Lett. **97**, 68002.
- Giraud, O., B. Georgeot, and D. L. Shepelyansky, 2005, Phys. Rev. E **72**, 036203.
- Giraud, O., B. Georgeot, and D. L. Shepelyansky, 2009, Phys. Rev. E **80**, 026107.
- Goldsheid, I.Y., and B.A. Khoruzhenko, 1998, Phys. Rev. Lett. **80**, 2897.
- Golub, G. H., and C. Greif, 2006, BIT Num. Math. **46**, 759.
- Gühr, T., A. Mueller-Groeling, and H. A. Weidenmueller, 1998, Phys. Rep. **299**, 189.
- Haake, F., 2010, *Quantum Signatures of Chaos* (Springer-Verlag, Berlin).
- Hart, M.H., 1992, *The 100: ranking of the most influential persons in history* (Citadel Press, N.Y.).
- He, J. and M. W. Deem 2010, Phys. Rev. Lett. **105**, 198701.
- Hrisitidis, V., H. Hwang, and Y. Papakonstantinou, 2008, ACM Trans. Database Syst. **33**, 1.
- Izhikevich, E.M., and G. M. Edelman, 2008, Proc. Nat. Acad. Sci. **105**, 3593.
- Kandiah, V., and D. L. Shepelyansky, 2012, Physica A **391**, 5779.
- Kandiah, V., and D. L. Shepelyansky, 2013, PLoS ONE **8**(5), e61519.
- Kandiah, V., and D. L. Shepelyansky, 2014a, Phys. Lett. A **378**, 1932.
- Kandiah, V., B. Georgeot, and O. Giraud, 2014b, arXiv:1405.6077 [physics.soc-ph].
- Kernighan, B.W., and D.M. Ritchie 1978, *The C Programming Language* (Englewood Cliffs, NJ Prentice Hall).
- Kleinberg, J.M., 1999, Jour. of the ACM **46**(5), 604.
- Krapivsky, P.L., S. Redner, and E. Ben-Naim 2010, *A Kinetic View of Statistical Physics* (Cambridge University Press, Cambridge UK).
- Krugman, P. R., M. Obstfeld, and M. Melitz, *International economics: theory & policy*, 2011, Prentic Hall, New Jersey.
- Landau, L.D., and E.M. Lifshitz 1989, *Quantum Mechanics* (Nauka, Moscow).
- Langville, A.M., and C.D. Meyer 2006, *Google's PageRank and Beyond: The Science of Search Engine Rankings* (Princeton University Press, Princeton).
- Li, T.-Y., 1976, J. Approx. Theory **17**, 177.
- Lichtenberg, A. J., and M.A. Lieberman 1992, *Regular and Chaotic Dynamics* (Springer, Berlin).
- Linux Kernel releases are downloaded from, 2010, <http://www.kernel.org/>.
- Litvak, N., W. R. W. Scheinhardt, and Y. Volkovich, 2008, Lect. Not. Comp. Sci. (Springer) **4936**, 72.
- Lu, W.T., S. Sridhar, and M. Zworski, 2003, Phys. Rev. Lett. **91**, 154101.
- Mantegna, R.N., S.V. Buldyrev, A.L. Goldberg, S. Havlin, C.-K. Peng, M. Simons, and H.E. Stanley, 1995, Phys. Rev. E **52**, 2939.
- Markov, A. A., 1906, Izvestiya Fiziko- matematicheskogo obshchestva pri Kazanskom universitete, 2-ya seriya (in Russian) **15**, 135.
- May, R.M., 2001, *Stability and Complexity in Model Ecosystems* (Princeton Univ. Press, New Jersey, USA).
- Mehta, M. L., 2004, *Random matrices* (Elsevier-Academic Press, Amsterdam).
- Memmott, J., N.M. Waser, and M.V. Price, 2004, Proc. R. Soc. Lond. B **271**, 2605.
- Meusel, R., S. Vigna, O. Lehmerberg, and C. Bizer, 2014, Proc. WWW14 Companion, <http://dx.doi.org/10.1145/2567948.2576928>.
- Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, 2002, Science **298**, 824.
- Muchnik, L., R. Itzhack, S. Solomon, and Y. Louzoun, 2007, Phys. Rev. E **76**, 016106.
- von Neumann, J., 1958, *The Computer and The Brain* (Yale Univ. Press, New Haven CT).
- Newman, M. E. J., 2001, Proc. Natl. Acad. Sci. USA **98**, 404.
- Newman, M. E. J., 2003, SIAM Review **45**, 167.
- Newman, M. E. J., 2010, *Networks: An Introduction* (Oxford University Press, Oxford UK).
- Nonnenmacher, S., and M. Zworski, 2007, Commun. Math. Phys. **269**, 311.
- Nonnenmacher, S., J. Sjostrand, and M. Zworski, 2014, Ann. Math. **179**, 179.
- Olesen, J.M., J. Bascompte, Y.L. Dupont, and P. Jordano, 2007, Proc. Natl. Acad. Sci. USA **104**, 19891.
- Pandurangan, G., P. Raghavan, and E. Upfal, 2005, Internet Math. **3**, 1.
- Pantheon MIT project, 2014, *Pantheon MIT project*

- http://pantheon.media.mit.edu .
- Perra, N., V. Zlatic, A. Chessa, C. Conti, D. Donato, and G. Caldarelli, 2009, *Europhys. Lett.* **88**, 48002.
- Radicchi, F., S. Fortunato, B. Markines, and A. Vespignani, 2009, *Phys. Rev. E* **80**, 056103.
- Redner, S., 1998, *Eur. Phys. J. B* **4**, 131.
- Redner, S., 2005, *Phys. Today* **58(6)**, 49.
- Rezende, E.L., J.E. Lavabre, P.R. Guimaraes, P. Jordano, and J. Bascompte, 2007, *Nature (London)* **448**, 925.
- Rodríguez-Gironés, M.A., and L. Santamaría, 2006, *J. Biogeogr.* **33**, 924.
- Saverda, S., D.B. Stouffer, B. Uzzi, and J. Bascompte, 2011, *Nature (London)* **478**, 233.
- Serra-Capizzano, S., 2005, *SIAM J. Matrix Anal. Appl.* **27**, 305.
- Serrano, M. A., M. Boguna, and A. Vespignani, 2007, *J. Econ. Interac. Coor.* **2**, 111.
- Shanghai ranking, 2010, Academic ranking of world universities <http://www.shanghairanking.com/> .
- Shen-Orr, A., R. Milo, S. Mangan, and U. Alon, 2002, *Nature Genetics* **31(1)**, 64.
- Shepelyansky, D. L., 2001, *Phys. Scripta* **T90**, 112.
- Shepelyansky, D. L., 2008, *Phys. Rev. E* **77**, 015202(R).
- Shepelyansky, D. L., and O. V. Zhirov, 2010a, *Phys. Rev. E* **81**, 036213.
- Shepelyansky, D. L., and O. V. Zhirov, 2010b, *Phys. Lett. A* **374**, 3206.
- Sjöstrand, J., 1990, *Duke Math. J.* **60**, 1.
- SJR, 2007, SCImago. (2007). SJR SCImago Journal & Country Rank <http://www.scimagojr.com> .
- Skiena, S., and C.B. Ward 2014, *Who's bigger?: where historical figures really rank* (Cambridge University Press, New York); <http://www.whoisbigger.com/>.
- Soramäki, K., M. L. Bech, J. Arnold, R. J. Glass, and W. E. Beyeler, 2005, *Physica A* **379**, 317.
- Song, C., S. Havlin, and H.A. Makse, 2005, *Nature (London)* **433**, 392.
- Sporns, O., 2007, *Scholarpedia* **2(10)**, 4695.
- Stewart, G. W., 2001, *Matrix Algorithms Vol. II: Eigensystems* (SIAM, Philadelphia PA).
- Sznajd-Weron, K., and J. Sznajd, 2000, *Int. J. Mod. Phys. C* **11**, 1157.
- Towlson, E.K., P. E. Vértes, S.E. Ahnert, W.R. Schafer, and E.T. Bullmore, 2013, *J. Neurosci.* **33(15)**, 6380.
- Tromp, J., and G. Farnebäck, 2007, *Lect. Notes Comp. Sci. (Springer)* **4630**, 84.
- UK universities, 2011, *Academic Web Link Database* <http://cybermetrics.wlv.ac.uk/database/> .
- Ulam, S., 1960, *A Collection of Mathematical Problems* (Interscience Tracts in Pure and Applied Mathematics, Interscience, New York).
- UN COMTRADE, 2011, *United Nations Commodity Trade Statistics Database* <http://comtrade.un.org/db/> .
- Vázquez, D.P., and M. A. Aizen, 2004, *Ecology* **85**, 1251.
- Vigna, S., 2013, arXiv:0912.0238v13[cs.IR] .
- Watts, D. J., and S. H. Strogatz, 1998, *Nature (London)* **393**, 440.
- Weyl, H., 1912, *Math. Ann.* **141**, 441.
- Wikipedia top 100 article, 2014, *Top 100 historical figures of Wikipedia* <http://www.wikipedia.org/en>
- Zaller, J.R., 1999, *The Nature and Origins of Mass Opinion* (Cambridge University Press, Cambridge UK).
- Zhirov, A. O., O. V. Zhirov, and D. L. Shepelyansky, 2010, *Eur. Phys. J. B* **77**, 523; <http://www.quantware.ups-tlse.fr/QWLIB/2drankwikipedia/> .
- Zipf, G. K., 1949, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Boston).
- Zlatic, V., M. Bozicevic, H. Stefancic, and M. Domazet, 2006, *Phys. Rev. E* **74**, 016115.
- Zuo, X.-N., R. Ehmke, M. Mennes, D. Imperati, F.X. Castellanos, O. Sporns, and M.P. Milham, 2012, *Cereb. Cortex* **22**, 1862.
- Zworski, M., 1999, *Not. Am. Math. Soc.* **46**, 319.
- Zyczkowski, K., M. Kus, W. Slomczynski, and H.-J. Sommers, 2003, *J. Phys. A: Math. Gen.* **36**, 3425.



OPEN

Generalized friendship paradox in complex networks: The case of scientific collaboration

SUBJECT AREAS:
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
COMPLEX NETWORKS
APPLIED MATHEMATICS

Received
8 January 2014

Accepted
19 March 2014

Published
8 April 2014

Correspondence and requests for materials should be addressed to
Y.-H.E. (thinking22@gmail.com)

¹Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, F-31062 Toulouse, France, ²BECS, Aalto University School of Science, P.O. Box 12200, Espoo, Finland.

The friendship paradox states that your friends have on average more friends than you have. Does the paradox “hold” for other individual characteristics like income or happiness? To address this question, we generalize the friendship paradox for arbitrary node characteristics in complex networks. By analyzing two coauthorship networks of Physical Review journals and Google Scholar profiles, we find that the generalized friendship paradox (GFP) holds at the individual and network levels for various characteristics, including the number of coauthors, the number of citations, and the number of publications. The origin of the GFP is shown to be rooted in positive correlations between degree and characteristics. As a fruitful application of the GFP, we suggest effective and efficient sampling methods for identifying high characteristic nodes in large-scale networks. Our study on the GFP can shed lights on understanding the interplay between network structure and node characteristics in complex networks.

People live in social networks. Various behaviors of individuals are significantly influenced by their positions in such networks, whether they are offline or online^{1–3}. Through the interaction and communication among individuals, information, behaviors, and diseases spread^{4–10}. Thus understanding the structure of social networks could enable us to understand, predict, and even control social collective behaviors taking place on or via those networks. Social networks have been known to be heterogeneous, characterized by broad distributions of the number of neighbors or degree¹¹, assortative mixing¹², and community structure¹³ to name a few.

One of interesting phenomena due to the structural heterogeneity in social networks is the friendship paradox¹⁴. The friendship paradox (FP) can be formulated at individual and network levels, respectively. At the individual level, the paradox holds for a node if the node has smaller degree than the average degree of its neighbors. It has been shown that the paradox holds for most of nodes in both offline and online social networks^{14–16}. However, most people believe that they have more friends than their friends have¹⁷. The paradox holds for a network if the average degree of nodes in the network is smaller than the average degree of their neighbors¹⁴. The paradox can be understood as a sampling bias in which individuals having more friends are more likely to be observed by their friends. This bias has important implications for the dynamical processes on social networks, especially when it is crucial for the process to identify individuals having many neighbors, or high degree nodes. For example, let us consider the spreading process on networks. It turns out that sampling neighbors of random individuals is more effective and efficient than sampling random individuals for the early detection of epidemic spreading in large-scale social networks^{18,19}, and for developing efficient immunization strategies in computer networks²⁰. Recently, the information overwhelming or spam in social networking services like Twitter¹⁶ has been also explained in terms of the friendship paradox.

The friendship paradox has been considered only as the topological structure of social networks, mainly by focusing on the number of neighbors, among many other node characteristics. Each individual could be described by his/her cultural background, gender, age, job, personal interests, and genetic information^{21,22}. This is also the case for other kinds of networks: Web pages have their own fitness in World Wide Web²³, and scientific papers have intrinsic attractiveness in a citation network²⁴. These characteristics play significant roles in dynamical processes on complex networks^{21–25}. Hence, one can ask the question: Can the friendship paradox be applied to node characteristics other than degree?

To address this question, we generalize the friendship paradox for arbitrary node characteristics including degree. Similarly to the FP, our generalized friendship paradox (GFP) can be formulated at individual and network levels. The GFP holds for a node if the node has lower characteristic than the average characteristic



of its neighbors. The GFP holds for a network if the average characteristic of nodes in the network is smaller than the average characteristic of their neighbors. When the degree is considered as the node characteristic, the GFP reduces to the FP. In this paper, by analyzing two coauthorship networks of physicists and of network scientists, we show that your coauthors have more coauthors, more citations, and more publications than you have. This indicates that the friendship paradox holds not only for degree but also for other node characteristics. We also provide a simple analysis to show that the origin of the GFP is rooted in the positive correlation between degree and node characteristics. As applications of the GFP, two sampling methods are suggested for sampling nodes with high characteristics. We show that these methods are simple yet effective and efficient in large-scale social networks.

Results

Generalized friendship paradox in complex networks. We consider two coauthorship networks constructed from the bibliographic information of Physical Review (PR) journals and Google Scholar (GS) profile dataset of network scientists (See Method Section). Each node of a network denotes an author of papers and a link is established between two authors if they wrote a paper together. The number of nodes, denoted by N , is 242592 for the PR network and 29968 for the GS network. For the node characteristics in the PR network, we consider the number of coauthors, the number of citations, the number of publications, and the average number of citations per publication. As for the GS network, the number of coauthors and the number of citations are considered. The characteristic of node i will be denoted by x_i , and for the degree we denote it by k_i .

The generalized friendship paradox (GFP) can be studied at two different levels: (i) Individual level and (ii) network level.

(i) *Individual level.* The GFP holds for a node i if the following condition is satisfied:

$$x_i < \frac{\sum_{j \in \Lambda_i} x_j}{k_i}, \quad (1)$$

where Λ_i denotes the set of neighbors of node i . Note that setting $x_i = k_i$ reduces the GFP to the FP. We define the paradox holding probability $h(k, x)$ that a node with degree k and characteristic x satisfies the condition in Eq. (1). Figure 1 shows the empirical results of $h(k, x)$ for PR and GS networks. It is found that for fixed degree k , $h(k, x)$ decreases with increasing x for any characteristic x other than k (Fig. 1 (b-d,f)). The same decreasing tendency has been observed for $x = k$ (Fig. 1 (a,e)). In Eq. (1), the larger value of x_i is expected to lower the probability $h(k, x)$ if the characteristics of node i 's neighbors remain the same. As a limiting case, the node with minimum value of x , i.e., x_{\min} , is most likely to have friends with higher values of x , leading to $h(k, x_{\min}) = 1$. On the other hand, for the node with maximum value of x , we get $h(k, x_{\max}) = 0$.

Next, the dependence of $h(k, x)$ on the degree k can be classified as either increasing or being constant. Here the case of x denoting the degree is disregarded for both networks. The increasing behavior is observed mainly for the number of citations and the number of publications in the PR network in Fig. 1 (b,c), while the constant behavior is observed for the average number of citations per publication in the PR network and for the number of citations in the GS network, shown in Fig. 1 (d,f), respectively. In order to understand such difference, we calculate the Pearson correlation coefficient between k and x as

$$\rho_{kx} = \frac{1}{N} \sum_{i=1}^N \frac{(k_i - \langle k \rangle)(x_i - \langle x \rangle)}{\sigma_k \sigma_x}, \quad (2)$$

where $\langle x \rangle$ and σ_x denote the average and standard deviation of x . We also obtain the characteristic assortativity for each characteristic x , adopted from¹²:

$$r_{xx} = \frac{L \sum_l x_l x'_l - [\sum_l \frac{1}{2} (x_l + x'_l)]^2}{L \sum_l \frac{1}{2} (x_l^2 + x'^2_l) - [\sum_l \frac{1}{2} (x_l + x'_l)]^2}, \quad (3)$$

where x_l and x'_l denote characteristics of nodes of the l th link, with $l = 1, \dots, L$ and L is the total number of links in the network. The value of r_{xx} ranges from -1 to 1 , and it increases according to the tendency of high characteristic nodes to be connected to other high characteristic nodes. The values of these quantities are summarized in Table I. From now on, we denote the degree assortativity as r_{kk} .

The k -dependent behavior of $h(k, x)$ can be understood mainly as the combined effect of r_{kk} and ρ_{kx} . Since $r_{kk} \approx 0.47$ in the PR network, for a node i with fixed x_i , the larger k_i implies the larger k_j of its friend j . This may lead to the higher x_j , e.g., due to $\rho_{kx} \approx 0.79$ for the number of publications, leading to the increasing behavior of $h(k, x)$. However, for the average number of citations per publication showing $\rho_{kx} \approx 0.07$, the larger k_j does not imply the higher x_j , which leads to the constant behavior of $h(k, x)$. For the number of citations in the GS network, the almost neutral degree correlation by $r_{kk} \approx -0.02$ inhibits any correlated behavior between characteristics, thus we again observe the constant behavior of $h(k, x)$. We note that the neutral degree correlation in the GS network is unlike many other coauthorship networks, mainly due to incomplete information available from GS profiles, and due to the snowball sampling method we employed²⁶.

Now we define the average paradox holding probability as $H = \sum_k \int dx h(k, x) P(k, x)$, where $P(k, x)$ denotes the probability distribution function of node with degree k and characteristic x . As shown in Table I, the value of H is larger than 0.7 for every considered characteristic, implying that the GFP holds at the individual level to a large extent.

(ii) *Network level.* In order to investigate the GFP at the network level, we define the average characteristic of neighbors $\langle x \rangle_{nn}$ for comparing it to the average characteristic $\langle x \rangle$:

$$\langle x \rangle_{nn} = \frac{\sum_{i=1}^N k_i x_i}{\sum_{i=1}^N k_i}. \quad (4)$$

Here a node i with degree k_i has been considered as a neighbor k_i times. The GFP holds at the network level if the following condition is satisfied:

$$\langle x \rangle < \langle x \rangle_{nn}. \quad (5)$$

Note that setting $x_i = k_i$ reduces the GFP to the FP. As shown in Table I, the GFP holds for all characteristics considered. In other words, your coauthors have on average more coauthors, more citations, and more publications than you have.

In summary, our results indicate that the generalized friendship paradox holds at both individual and network levels for many node characteristics of networks.

Origin of the GFP. The prevalence of the GFP for most nodes in networks regardless of node characteristics implies that there might be a universal origin of the GFP. For the original friendship paradox, the existence of hub nodes and the variance of degree have been suggested for the origin of the paradox¹⁴. In order to investigate the origin of the GFP at the network level, we define a function $F = \langle x \rangle_{nn} - \langle x \rangle$, and straightforwardly obtain the following equation:

$$F = \langle x \rangle_{nn} - \langle x \rangle = \frac{\rho_{kx} \sigma_k \sigma_x}{\langle k \rangle}. \quad (6)$$

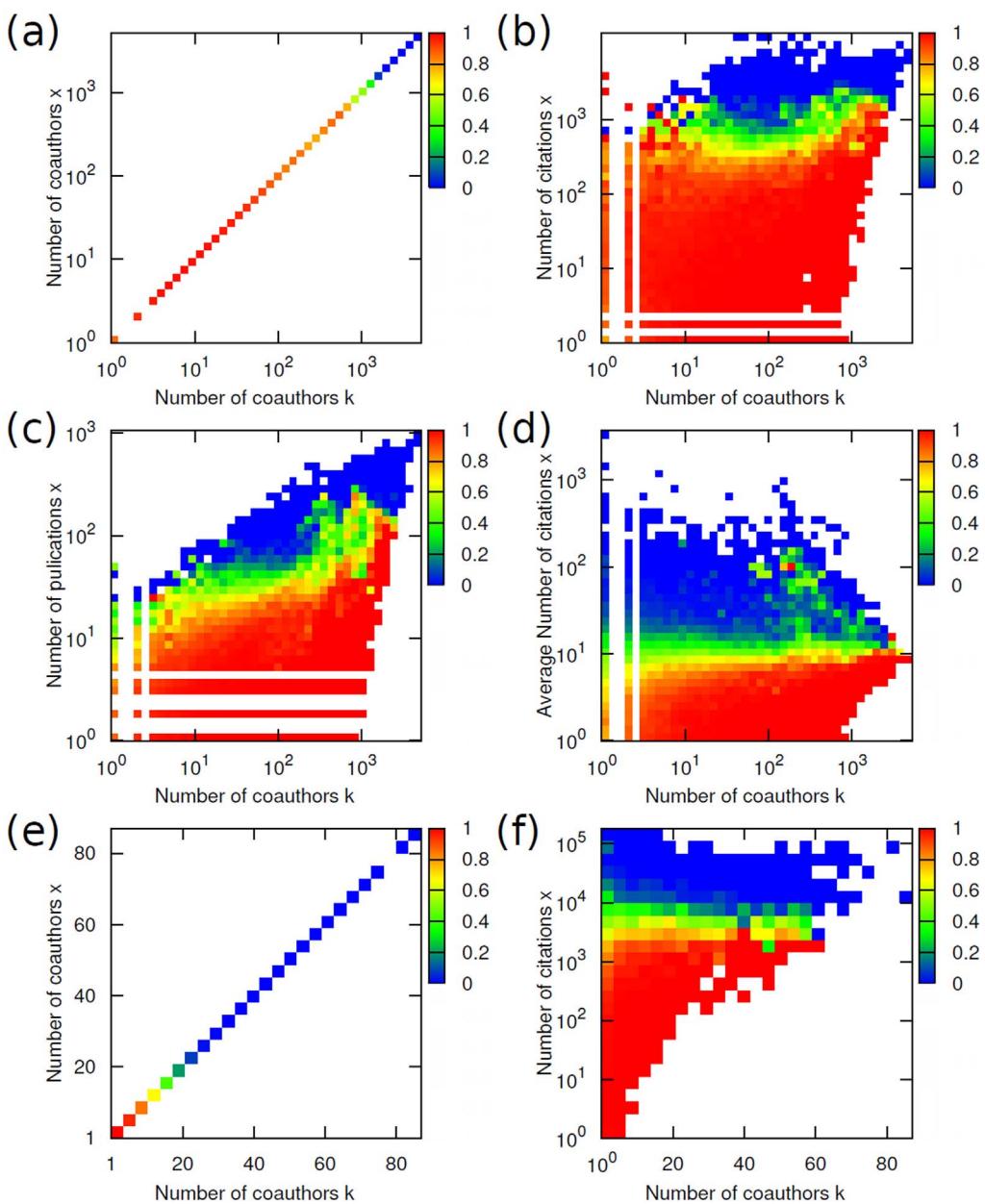


Figure 1 | The paradox holding probability $h(k, x)$ as a function of degree k and node characteristic x . For the Physical Review (PR) coauthorship network, we use (a) the number of coauthors, i.e., $x = k$, (b) the number of citations, (c) the number of publication, and (d) the average number of citations per publication, while for the Google Scholar (GS) coauthorship network, we use (e) the number of coauthors, i.e., $x = k$, and (f) the number of citations.

Table I | Empirical results for the generalized friendship paradox in two coauthorship networks from Physical Review (PR) journals and from Google Scholar (GS) profiles. For each node characteristic x , we measure the Pearson correlation coefficient with degree ρ_{kx} , the characteristic assortativity r_{xx} , the average paradox holding probability H , and average characteristics of nodes $\langle x \rangle$ and their neighbors $\langle x \rangle_{nn}$

characteristic x	ρ_{kx}	r_{xx}	H	$\langle x \rangle$	$\langle x \rangle_{nn}$
The number of coauthors (PR)	1.00	0.47	0.934	58.3	< 771.7
The number of citations (PR)	0.69	0.21	0.921	110.1	< 1135.7
The number of publications (PR)	0.79	0.25	0.912	10.2	< 102.1
The average number of citations per publication (PR)	0.07	0.34	0.720	7.8	< 12.4
The number of coauthors (GS)	1.00	-0.02	0.863	6.9	< 16.1
The number of citations (GS)	0.44	0.14	0.792	3089.8	< 5401.0

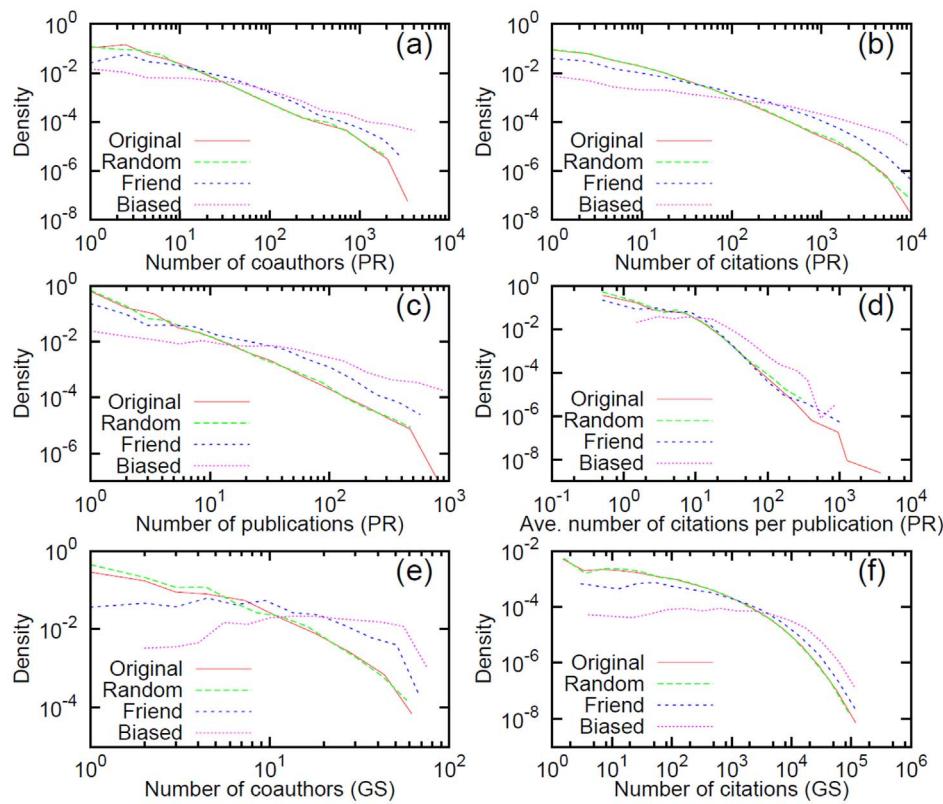


Figure 2 | Characteristic distributions for control group, friend group, and biased group, for each of which 5000 nodes are sampled. The original full distributions are also plotted for comparison. We use (a) the number of coauthors (PR), (b) the number of citations (PR), (c) the number of publications (PR), (d) the average number of citations per publication (PR), (e) the number of coauthors (GS), and (f) the number of citations (GS).

One can say that the GFP holds if $F > 0$. Since standard deviations σ_k and σ_x are positive in any non-trivial cases, the GFP holds if $\rho_{kx} > 0$. Thus the degree-characteristic correlation ρ_{kx} is the key element for the generalized friendship paradox. Note that in case when $x_i = k_i$, i.e., $\rho_{kk} = 1$, the FP holds in any non-trivial cases.

The origin of the GFP can help us to better understand the dynamical processes on networks when the characteristic x is considered to be a node activity such as communication frequency or traffic. The positive correlation between degree and node activity has been observed in mobile phone call patterns²⁷ and the air-transportation network²⁸, enabling the application of the GFP to those phenomena. In case of protein interaction networks, the degrees of proteins are positively correlated with their lethality^{29,30}, while they are negatively correlated with their rates of evolution³¹. The negative degree-characteristic correlations, i.e., $\rho_{kx} < 0$, can lead to the opposite behavior of the GFP, which can be called anti-GFP.

Sampling high characteristic nodes using GFP in complex networks. Identifying important or central nodes in a network is crucial for understanding the structure of complex networks and dynamical processes on those networks. The recent advance of information-communication technology (ICT) has opened up access to the data on large-scale social networks. However, complete mapping of social networks is not feasible, partially due to privacy issues. Thus it is still important to devise proper sampling methods that exploit local network structure. In this sense, the original friendship paradox has been used to sample high degree nodes in empirical networks. It was found that the set of neighbors of randomly chosen nodes can have the predictive power of epidemic spreading on both offline social networks¹⁸ and online social networks¹⁹.

We suggest two simple sampling methods using the GFP to identify high characteristic nodes in a network: (i) Friend sampling

and (ii) biased sampling. These methods are then compared to the random sampling method to test whether our methods are more efficient to sample high characteristic nodes. We first choose random nodes to make a control group. For each node in the control group, one of its neighbors is randomly chosen. These chosen nodes compose a friend group. Finally, for each node in the control group, we choose its neighbor having the highest characteristic to make a biased group. For the biased sampling, we have assumed that each node has the full information about characteristics of its neighbors.

Figure 2 shows the characteristic distributions of sampled nodes from PR and GS networks by different sampling methods. Heavier tails of distributions imply better sampling for identifying high characteristic nodes. The performance of biased sampling is the best in all cases because this sampling utilizes more information about neighbors than the friend sampling. The friend sampling shows better performance than the random sampling (control group) for most characteristics as it is expected by large values of ρ_{kx} . One exceptional case is for the average number of citations per publication in the PR network, shown in Fig. 2 (d). Here the friend sampling does not better than the random sampling due to the very small degree-characteristic correlation, $\rho_{kx} \approx 0.07$, while the result by biased sampling is still better than those by other sampling methods.

Next, in order to investigate the effect of degree-characteristic correlation on the performance of sampling methods, we consider an auxiliary characteristic X based on the method of Cholesky decomposition³². To each node i with degree k_i in the PR network, we assign a characteristic X_i given by

$$X_i = \rho k_i + \sqrt{1 - \rho^2} y_i, \quad (7)$$

where y_i denotes the i th element of the shuffled set of $\{k_i\}$. Since $\rho = \rho_{kX}$ (See Method Section), the correlation can be easily controlled by ρ . Then we apply the same sampling methods to identify nodes with

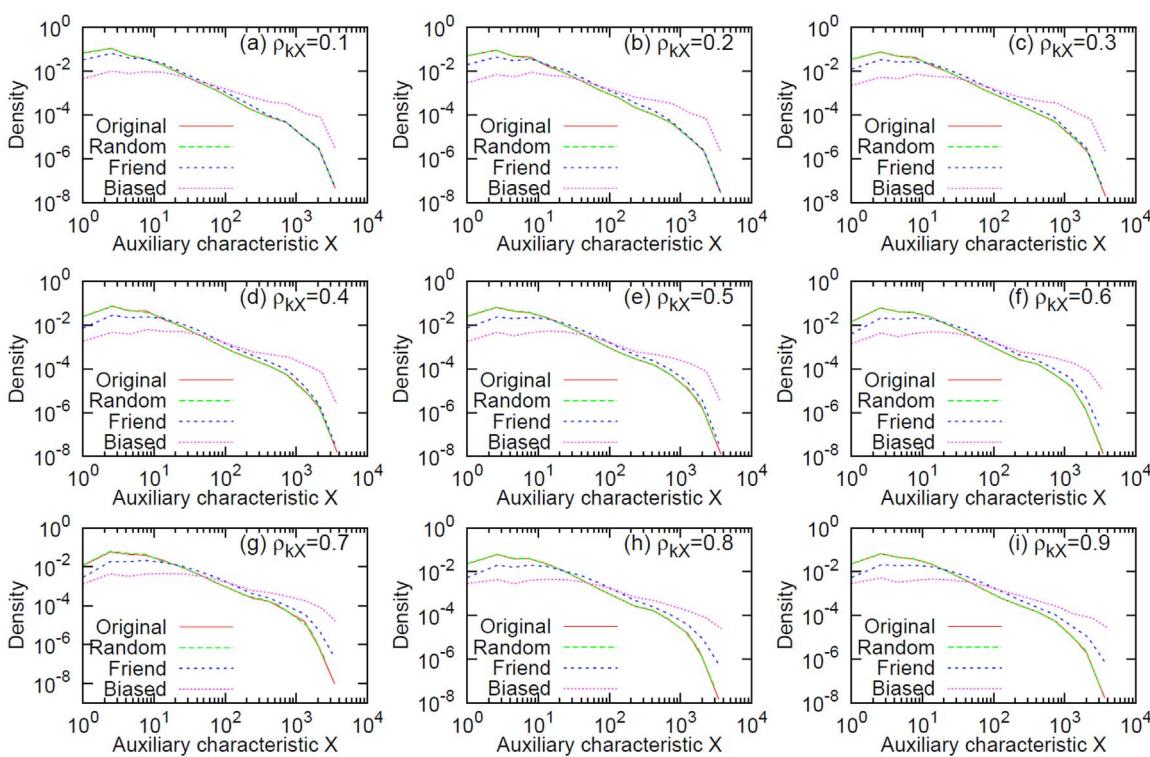


Figure 3 | Performance comparison of control group, friend group, and biased group for auxiliary characteristics X with various values of correlation with degree: $\rho_{kX} = 0.1, \dots, 0.9$. For each value of ρ_{kX} , 1000 random configurations are generated, for each of which 5000 nodes are sampled as in Fig. 2.

high X , and compare their performances for different values of ρ_{kX} . Figure 3 shows that the biased sampling performs significantly better than any other sampling methods, independent of ρ_{kX} . The friend sampling performs better than the random sampling, while the difference in performance increases with the value of ρ_{kX} .

The sampling results suggest that the biased sampling can be very efficient and effective to detect a group of high characteristic nodes when the information about characteristics of neighbors is available. Otherwise the friend sampling still performs better than the random sampling.

Discussion

Node characteristics have profound influence on the evolution of networks^{23,24} and dynamical processes on such networks like spreading^{18,19,25}. By taking into account various node characteristics, we have generalized the friendship paradox in complex networks. The generalized friendship paradox (GFP) states that your friends have on average higher characteristics than you have. By analyzing two coauthorship networks of Physical Review (PR) journals and of Google Scholar (GS) profiles, we have found that the GFP holds at both individual and network levels for various node characteristics, such as the number of coauthors, the number of citations, the number of publications, and the average number of citations per publication. It is also shown that the origin of the GFP at the network level is rooted in the positive correlation between degree and characteristic. Thus the GFP is expected to hold for any characteristic showing the positive correlation with degree. Here the characteristic can be also purely topological like various node centralities as they show significant positive correlations with degree, such as PageRank³³.

Despite the access to the data on large-scale social networks, complete mapping of social networks is not feasible. Thus it is still important to devise effective and efficient sampling methods that exploit local network structure. We have suggested two simple sampling methods for identifying high characteristic nodes using the

GFP. It is empirically found that a control group of randomly chosen nodes has the smaller number of high characteristic nodes than a friend group that consists of random neighbors of nodes in the control group. Moreover, provided that nodes have full information about characteristics of their neighbors, a biased group of the highest characteristic neighbors of nodes in the control group has the largest number of high characteristic nodes than other groups. This turns out to be the case even when the degree-characteristic correlation is negligible.

Our sampling methods propose an explanation about how our perception can be affected by our friends. People's perception of the world and themselves depends on the status of their friends, colleagues, and peers¹⁷. When we compare our characteristics like popularity, income, reputation, or happiness to those of our friends, our perception of ourselves might be distorted as expected by the GFP. Comparing to the average friend, i.e., the friend sampling, is biased due to the positive degree-characteristic correlation. Furthermore, comparing to the "better" friend, i.e., the biased sampling, is much more biased towards the "worse" perception of ourselves. This might be the reason why active online social networking service users are not happy³⁴, in which it is much easier to compare to other people in online social media.

Another interesting application of the GFP can be found in multiplex networks^{35,36}. If degrees of one layer are positively correlated with those of other layers, our sampling methods can be used to identify high degree nodes in other layers. Indeed, the degrees of each node are positively correlated across layers in a player network of an online game³⁷ and in a multiplex transportation network³⁸.

Nodes are not only embedded in the topological structure, but they also have many other characteristics relevant to the structure and evolution of complex networks. However, the role of these non-topological characteristics is far from being fully understood. Our work on the generalized friendship paradox will help us consider the interplay between network structure and node characteristics for deeper understanding of complex networks.



Methods

Data description. We describe how the data for coauthorship networks have been collected and prepared. For the Physical Review (PR) network, the bibliographic data containing all papers published in Physical Review journals from 1893 to 2009 was downloaded from American Physical Society. The number of papers is 463348, and each paper has the title, the list of authors, the date of publication, and citation information. By using author identification algorithm proposed by³⁹, we identified each author by his/her last name and initials of first and middle names if available. The number of identified authors is 242592. Combined with the numbers of citations and the list of authors of papers, we obtained for each author the number of coauthors, the number of citations, the number of publications, and the average number of citations per publication.

Google Scholar (GS) service (scholar.google.com) provides profiles of academic authors. Each profile of the author contains information of the total number of citations and coauthor list of the author. Using snowball sampling²⁶ starting from “Albert-László Barabási” (one of the leading network scientists), the coauthor relations and their citation information are collected. The number of authors in the dataset is 29968. Here we note that not all scientists have profile in the GS and not all coauthor relations are accessible.

Generating random node characteristics of arbitrary correlation with degree.

Consider two independent random variables $Y = (y_1, y_2, \dots, y_N)$ and $Z = (z_1, z_2, \dots, z_N)$ with the same standard deviation, i.e., $\sigma_Y = \sigma_Z$. We generate a random sequence $X = (x_1, x_2, \dots, x_N)$ from the following equation:

$$X = \rho Y + \sqrt{1 - \rho^2} Z. \quad (8)$$

The correlation ρ_{XY} between X and Y is given by

$$\rho_{XY} = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y}, \quad (9)$$

where $E(X)$ denotes the expectation of X . Using the independence of Y and Z , i.e., $E(YZ) = E(Y)E(Z)$, we get

$$\rho_{XY} = \rho \frac{\sigma_Y}{\sigma_X}. \quad (10)$$

Then, from $\sigma_X^2 = E(X^2) - E(X)^2$, we obtain $\sigma_X = \sigma_Y$, leading to

$$\rho_{XY} = \rho. \quad (11)$$

1. Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).
2. Castello, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646 (2009).
3. Lazer, D. *et al.* Computational social science. *Science* **323**, 721–723 (2009).
4. Vespignani, A. Modelling dynamical processes in complex socio-technical systems. *Nat. Phys.* **8**, 32–39 (2011).
5. Centola, D. The spread of behavior in an online social network experiment. *Science* **329**, 1194–1197 (2010).
6. Bakshy, E., Rosenn, I., Marlow, C. & Adamic, L. The role of social networks in information diffusion. In *WWW’12: Proc. 21st Intl. Conf. on World Wide Web* Lyon, France. New York, NY, USA: ACM. (2012, April 16–20).
7. Christakis, N. A. & Fowler, J. H. The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.* **357**, 370 (2007).
8. Pastor-Satorras, R. & Vespignani, A. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200–3203 (2001).
9. Weng, L., Menczer, F. & Ahn, Y.-Y. Virality prediction and community structure in social networks. *Sci. Rep.* **3**, 2522 (2013).
10. Marvel, S. A., Martin, T., Doering, C. R., Lusseau, D. & Newman, M. E. J. The small-world effect is a modern phenomenon. *arXiv:1310.2636* (2013).
11. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1998).
12. Newman, M. E. J. Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).
13. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).
14. Feld, S. L. Why Your Friends Have More Friends Than You Do. *Am. J. of Sociol.* **96**, 1464–1477 (1991).
15. Ugander, J., Karrer, B., Backstrom, L. & Marlow, C. The anatomy of the Facebook social graph. *arXiv:1111.4503* (2011).
16. Hodas, N. O., Kooti, F. & Lerman, K. Friendship paradox redux: Your friends are more interesting than you. In *ICWSM’13: Proc 7th Int. AAAI Conf. on Weblogs and Social Media*, Cambridge, MA, USA. Palo Alto, CA, USA: The AAAI press (2013, July 8–10).
17. Zuckerman, E. & Jost, J. What makes you think you’re so popular? Self-evaluation maintenance and the subjective side of the “friendship paradox”. *Soc. Psychol. Q.* **64**, 207–223 (2001).

18. Christakis, N. A. & Fowler, J. H. Social network sensors for early detection of contagious outbreaks. *PLoS ONE* **5**, e12948 (2010).
19. Garcia-Herranz, M., Moro, E., Cebrian, M., Christakis, N. A. & Fowler, J. H. Using friends as sensors to detect global-scale contagious outbreaks. *arXiv:1211.6512* (2012).
20. Cohen, R., Havlin, S. & ben-Avraham, D. Efficient immunization strategies for computer networks and populations. *Phys. Rev. Lett.* **91**, 247901 (2003).
21. Park, J. & Barabási, A.-L. Distribution of node characteristics in complex networks. *Proc. Natl. Acad. Sci. USA* **104**, 17916–17920 (2007).
22. Fowler, J. H., Dawes, C. T. & Christakis, N. A. Model of genetic variation in human social networks. *Proc. Natl. Acad. Sci. USA* **106**, 1720–1724 (2008).
23. Kong, J. S., Sarshar, N. & Roychowdhury, V. P. Experience versus talent shapes the structure of the Web. *Proc. Natl. Acad. Sci. USA* **105**, 13724–13729 (2008).
24. Eom, Y.-H. & Fortunato, S. Characterizing and modeling citation dynamics. *PLoS ONE* **6**, e24926 (2011).
25. Aral, S., Muchnik, L. & Sundararajan, A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. USA* **106**, 21544–21549 (2009).
26. Lee, S. H., Kim, P.-J. & Jeong, H. Statistical properties of sampled networks. *Phys. Rev. E* **73**, 016102 (2006).
27. Onnela, J. *et al.* Analysis of a large-scale weighted network of one-to-one human communication. *New J. Phys.* **9**, 179 (2007).
28. Barrat, A., Barthelemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **101**, 3747–3752 (2004).
29. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
30. Zotenko, E., Mestre, J., O’Leary, D. P. & Przytycka, T. M. Why do hubs in the Yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.* **4**, e1000140 (2008).
31. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. Evolutionary rate in the protein interaction network. *Science* **296**, 750–752 (2002).
32. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes in C: The Art of Scientific Computing - Second Edition*. (Cambridge University Press, Cambridge 1992).
33. Fortunato, S., Boguña, M., Flammini, A. & Menczer, F. [Approximating PageRank from indegree] *Algorithms and Models for the Web-Graph* [59–71] (Springer Berlin Heidelberg, Germany, 2008).
34. Kross, E. *et al.* Facebook Use Predicts Declines in Subjective Well-Being in Young Adults. *PLoS ONE* **8**, e69841 (2013).
35. Kivelä, M. *et al.* Multilayer networks. *arXiv:1309.7233* (2013).
36. Jo, H.-H., Baek, S. K. & Moon, H.-T. Immunization dynamics on a two-layer network model. *Physica A* **361**, 534–542 (2006).
37. Szell, M., Lambiotte, R. & Thurner, S. Multirelational organization of large-scale social networks in an online world. *Proc. Natl. Acad. Sci. USA* **107**, 13636 (2010).
38. Parshani, R., Rozenblat, C., Ietri, D., Ducruet, C. & Havlin, S. Inter-similarity between coupled networks. *Europhys. Lett.* **92**, 68002 (2010).
39. Radicchi, F., Fortunato, S., Markines, B. & Vespignani, A. Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E* **80**, 056103 (2009).

Acknowledgments

The authors thank Daniel Kim and Hawoong Jeong for providing Google Scholar data and American Physical Society for providing Physical Review bibliographic data. Y.-H.E. acknowledges support from the EC FET Open project “New tools and algorithms for directed network analysis” (NADINE number 288956). H.-H.J. acknowledges financial support by the Aalto University postdoctoral programme.

Author contributions

Y.-H.E. and H.-H.J. designed research, wrote, reviewed, and approved the manuscript. Y.-H.E. performed data collection and analysis.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Eom, Y.-H. & Jo, H.-H. Generalized friendship paradox in complex networks: The case of scientific collaboration. *Sci. Rep.* **4**, 4603; DOI:10.1038/srep04603 (2014).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The images in this article are included in the article’s Creative Commons license, unless indicated otherwise in the image credit; if the image is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the image. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/>

Generalized friendship paradox in networks with tunable degree-attribute correlation

Hang-Hyun Jo^{1,*} and Young-Ho Eom²

¹BECS, Aalto University School of Science, P.O. Box 12200, Espoo, Finland

²Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, F-31062 Toulouse, France

(Received 6 May 2014; published 21 August 2014)

One of the interesting phenomena due to topological heterogeneities in complex networks is the friendship paradox: Your friends have on average more friends than you do. Recently, this paradox has been generalized for arbitrary node attributes, called the generalized friendship paradox (GFP). The origin of GFP at the network level has been shown to be rooted in positive correlations between degrees and attributes. However, how the GFP holds for individual nodes needs to be understood in more detail. For this, we first analyze a solvable model to characterize the paradox holding probability of nodes for the uncorrelated case. Then we numerically study the correlated model of networks with tunable degree-degree and degree-attribute correlations. In contrast to the network level, we find at the individual level that the relevance of degree-attribute correlation to the paradox holding probability may depend on whether the network is assortative or disassortative. These findings help us to understand the interplay between topological structure and node attributes in complex networks.

DOI: [10.1103/PhysRevE.90.022809](https://doi.org/10.1103/PhysRevE.90.022809)

PACS number(s): 89.75.-k, 89.65.-s

I. INTRODUCTION

Human societies have been successfully described within the framework of complex networks, where nodes and links denote individuals and their dyadic relationships, respectively [1–5]. As individuals are embedded in social networks, their positions in such networks strongly influence their behaviors [3] as well as self-evaluations [6] and subjective well-being [7]. In particular, the comparison to friends, colleagues, and peers enables individuals to adopt and transmit opinion, information, and technologies [2,8,9], e.g., for competitiveness [10]. Thus understanding positional differences between individuals is crucial to understanding the emergent collective dynamics at the community or societal level [11].

Topological structures of social networks have been known to be heterogeneous, characterized by broad distributions of the number of neighbors or degree [12], assortative mixing [13], and community structure [14]. One of the interesting phenomena due to topological heterogeneities is the friendship paradox (FP). The FP states that your friends have on average more friends than you do [15]. The paradox has been shown to hold in both offline and online social networks [15–20]. Examples include friendship networks of middle and high school students [15,20] and of university students [6], scientific collaboration networks [18], and Facebook and Twitter user networks [16,17,19]. The paradox can be understood as a sampling bias in which individuals having more friends are more likely to be observed by their friends. This bias has important implications for the dynamical processes on social networks, e.g., for efficient immunization [21] and for early detection of contagious outbreaks [22,23] or of natural disasters [24]. The paradox implies that your friends and neighbors tend to occupy more important or central positions in social networks than you do.

The importance or centrality of individuals is not determined only by their topological positions in networks, but is also influenced by their attributes. Individuals can be described by various attributes like gender, age, cultural preferences, and genetic information [25,26]. This requires us to study the interplay between topological structure and node attributes of social networks. The friendship paradox has been also considered for arbitrary node attributes [17–19], which is called the generalized friendship paradox (GFP) [18]. Note that if the degree of node is considered as the attribute, the GFP reduces to the FP.

The GFP can be formulated at the individual and network levels. The GFP holds for a network if the average attribute of nodes in the network is smaller than the average attribute of their neighbors. The GFP holds for a node if the node has a lower attribute than the average attribute of its neighbors. The GFP at both levels has been observed in the coauthorship networks [18]. While the GFP at the network level accounts for the average behavior of the network, the GFP at the individual level can provide more detailed understanding of the centrality of individuals, and of their subjective evaluations of attributes. For example, consider a star network, where one hub node is connected to all other nodes. The network level analysis cannot tell the positional and attribute-related difference between the hub node and all other nodes. Thus it is obvious that these individual properties cannot be fully revealed in the network level analysis, especially when the individuals are heterogeneous in terms of broad distributions of degree and attribute.

The origin of the GFP at the network level has been clearly shown to be rooted in positive degree-attribute correlations [18]. In other words, high-attribute individuals are more likely to be observed by their friends as high-attribute individuals have more friends. However, the role of degree-attribute correlations at the individual level is far from being fully understood. In order to investigate the role of various correlations for the GFP at the individual level, we first analyze a solvable model to characterize the paradox holding probability of nodes for the uncorrelated case. Then we numerically study the correlated model of networks with

*Present address: BK21plus Physics Division and Department of Physics, Pohang University of Science and Technology, Pohang 790-784, Republic of Korea.

tunable degree-degree and degree-attribute correlations. By calculating the paradox holding probabilities for the entire range of correlations, we show that the relevance of degree-attribute correlation to the paradox holding probability may depend on whether the network is assortative or disassortative. This result is compared to the GFP at the network level. Finally, we conclude the paper by summarizing the results.

II. GENERALIZED FRIENDSHIP PARADOX

A. Network level

The generalized friendship paradox (GFP) holds for a network if the average attribute of nodes in the network is smaller than the average attribute of their neighbors. For a network of N nodes, let us denote a degree and an attribute of node i as k_i and x_i , respectively. The average degree and average attribute are $\langle k \rangle = N^{-1} \sum_{i=1}^N k_i$ and $\langle x \rangle = N^{-1} \sum_{i=1}^N x_i$. The average attribute of neighbors $\langle x \rangle_{nn}$ is obtained as

$$\langle x \rangle_{nn} = \frac{\sum_{i=1}^N k_i x_i}{\sum_{i=1}^N k_i}, \quad (1)$$

where a node i with degree k_i has been counted k_i times by its neighbors. Then the GFP holds for a network if the following condition is satisfied:

$$\langle x \rangle < \langle x \rangle_{nn}. \quad (2)$$

By the straightforward calculation, one gets

$$\langle x \rangle_{nn} - \langle x \rangle = \frac{\rho_{kx} \sigma_k \sigma_x}{\langle k \rangle}, \quad (3)$$

where the degree-attribute correlation is given by

$$\rho_{kx} = \frac{1}{N} \sum_{i=1}^N \frac{(k_i - \langle k \rangle)(x_i - \langle x \rangle)}{\sigma_k \sigma_x}. \quad (4)$$

Since standard deviations of degree and attribute, i.e., σ_k and σ_x , are positive in any nontrivial cases, the positive ρ_{kx} leads to the GFP at the network level. Thus, the origin of GFP at the network level is rooted in positive correlation between degree and attribute [18]. The GFP at the network level has been observed in the coauthorship networks of Physical Review journals (PR) and of Google Scholar profiles (GS) for several attributes such as the number of publications by each author [18]. In addition, the negative ρ_{kx} can lead to the opposite tendency, implying that your friends have on average a lower attribute than you do. This can be called anti-GFP.

B. Individual level: Uncorrelated solvable model

In order to investigate the GFP at the individual level, we study an uncorrelated solvable model. The GFP holds for a node i if the node has a lower attribute than the average attribute of its neighbors, precisely if the following condition is satisfied:

$$x_i < \frac{1}{k_i} \sum_{j \in \Lambda_i} x_j, \quad (5)$$

where Λ_i denotes the set of i 's neighbors. The probability of satisfying Eq. (5) or *paradox holding probability* may be interpreted as the degree of self-evaluation of the node when

compared to its neighbors. We assume no correlation between attributes of neighboring nodes, implying that the degrees of neighbors are entirely irrelevant to the probability. Then one gets the paradox holding probability of a node with degree k and attribute x as

$$h_k(x) \equiv \Pr \left(\frac{1}{k} \sum_{j=1}^k x_j > x \right) \quad (6)$$

$$= \prod_{j=1}^k \int_0^\infty dx_j P(x_j) \theta \left(\frac{1}{k} \sum_{j=1}^k x_j - x \right), \quad (7)$$

where $\theta(\cdot)$ is a Heaviside step function. The distribution of x has been denoted by $P(x)$ with $x \geq 0$. In general x can have negative values, which will be considered in Sec. II C. By taking the Laplace transform with respect to x , we get

$$\tilde{h}_k(s) = \frac{1}{s} \left[1 - \tilde{P} \left(\frac{s}{k} \right)^k \right], \quad (8)$$

where $\tilde{P}(s)$ is the Laplace transform of $P(x)$. Then, the paradox holding probability $h_k(x)$ can be obtained by taking the inverse Laplace transform of $\tilde{h}_k(s)$ analytically or numerically if necessary.

For the solvable yet broadly distributed case, we consider the Gamma distribution for x , i.e.,

$$P(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}, \quad (9)$$

where $\alpha, \beta > 0$ and the mean of x is $\langle x \rangle = \alpha\beta$. Since $\tilde{P}(s) = (\beta s + 1)^{-\alpha}$, one gets

$$h_k(x) = \frac{\Gamma(\alpha k, \alpha k \frac{x}{\langle x \rangle})}{\Gamma(\alpha k)}. \quad (10)$$

Here $\Gamma(s, z) = \int_z^\infty t^{s-1} e^{-t} dt$ denotes the upper incomplete Gamma function. The heat map of $h_k(x)$ as a function of αk and $x/\langle x \rangle$ is depicted in Fig. 1(a).

For any given k , it is obvious that $h_k(0) = 1$ and $h_k(\infty) = 0$, and that $h_k(x)$ is a decreasing function of x . For a given x , one can study the k -dependent behavior of $h_k(x)$. In case of $k = 1$,

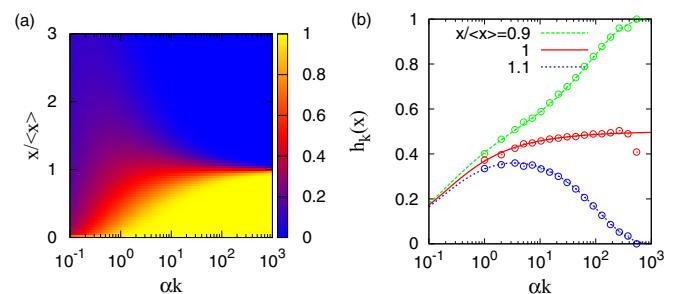


FIG. 1. (Color online) Analytic results of the uncorrelated model with Gamma distributions for x and k in Eq. (9). (a) Heat map of the paradox holding probability $h_k(x)$ in Eq. (10) as a function of αk and $x/\langle x \rangle$. (b) $h_k(x)$ as a function of αk for values of $x/\langle x \rangle = 0.9$, 1, and 1.1 (curves), which are compared to the numerical results (circles) from the uncorrelated network of size $N = 10^5$ and of $\langle x \rangle = 50$ using the same Gamma distribution in Eq. (9).

$h_1(x)$ is the probability of drawing one number larger than x from $P(x)$, which we denote $f_x \equiv \int_x^\infty P(x')dx'$. The value of $h_2(x)$ is upper bounded by the probability that when two numbers are drawn from $P(x)$, both numbers are not smaller than x , i.e., $h_2(x) \leq 1 - (1 - f_x)^2$. Even when one neighbor has an attribute less than x and the other has an attribute more than x , it is likely that the average of them exceeds x due to the broadness of $P(x)$. Thus, we approximate as $h_2(x) \approx 1 - (1 - f_x)^2$, which is then generalized to $h_k(x) \approx 1 - (1 - f_x)^k$. This argument accounts for the k -dependent increasing behavior for small αk in the solution of Eq. (10). It could imply that having more friends may lead to the lower self-evaluation to some extent. However, for sufficiently large k , the average of attributes of neighbors converges to $\langle x \rangle$. Hence, when the given x is smaller (larger) than $\langle x \rangle$, $h_k(x)$ approaches 1 (0) as k increases. In case of $x = \langle x \rangle$, $h_k(x)$ approaches $1/2$ as k increases. Note that only when $x > \langle x \rangle$, $h_k(x)$ increases and then decreases according to k . Such nontrivial behavior emerges even in the uncorrelated case.

Next, in order to study the FP in the uncorrelated setup, one needs to solve the following equation:

$$h_k^{\text{FP}} \equiv \Pr \left(\frac{1}{k} \sum_{j=1}^k k_j > k \right) \quad (11)$$

$$= \sum_{\{k_j\}} \prod_{j=1}^k P(k_j) \theta \left(\frac{1}{k} \sum_{j=1}^k k_j - k \right), \quad (12)$$

where $P(k)$ denotes the degree distribution. As there is no general solution to our knowledge, the FP will be numerically studied in Sec. II C.

C. Individual level: Correlated network model

We numerically study more general cases, including the uncorrelated model, by generating networks with tunable degree-degree and degree-attribute correlations. Following the configuration model [27], we generate the degree sequence, $\{k_i\}$ for nodes $i = 1, \dots, N$, where each degree is independently drawn from $P(k)$ with minimum degree as $k_{\min} = 1$. Each node has k_i stubs or half links. A pair of nodes are randomly selected and a link is established between them if both nodes have residual stubs and if there is no link between them. This process is repeated until when no stubs remain. In principle, the generated network has no degree-degree correlations. Degree-degree correlations can be fully measured in terms of the joint degree distribution $P(k, k')$ with k and k' denoting degrees of neighboring nodes. However, for tractability of the model, we adopt the assortativity coefficient [13]

$$r_{kk} = \frac{L \sum_l k_l k'_l - [\sum_l \frac{1}{2}(k_l + k'_l)]^2}{L \sum_l \frac{1}{2}(k_l^2 + k'^2_l) - [\sum_l \frac{1}{2}(k_l + k'_l)]^2}, \quad (13)$$

where k_l and k'_l denote degrees of nodes of the l th link with $l = 1, \dots, L$, and L is the total number of links in the network. Indeed, r_{kk} is the normalized quantity of the first order moment of $P(k, k')$, i.e., $\langle k k' P(k, k') \rangle$. The value of r_{kk} ranges from -1 to 1 , and it quantifies the tendency of large degree nodes being connected to other large degree nodes. A network with

the maximal r_{kk} can be implemented, e.g., by constructing k cliques or complete subgraph with k nodes. The minimal r_{kk} can be found in the starlike network structure, where hubs are connected to dangling nodes.

For preparing the network with a desired value of r_{kk} , we rewire links as following [28]: Two links are randomly selected, e.g., a link between nodes i and j and a link between nodes i' and j' . These nodes are rewired to links between i and i' and between j and j' , only when the value of r_{kk} gets closer to the desired value. This rewiring is repeated until the desired value of r_{kk} is reached. In order to investigate if these generated networks result in the desired degree-degree correlations, we measure $P(k, k')$ (not shown here) implying that the generated networks are fully random to any other respect than the correlation by the assortativity coefficient. Thus, r_{kk} will also be used as an indicator for the degree-degree correlations.

For the tunable degree-attribute correlation, denoted by ρ_{kx} , we adopt the method used in Ref. [18]. For a given degree sequence, the attribute of a node i is assigned as

$$x_i = \rho k_i + \sqrt{1 - \rho^2} k_j, \quad (14)$$

where the node index j is randomly chosen from $\{1, \dots, N\}$. It is straightforward to prove that $\rho = \rho_{kx}$ [18]. ρ can have a value in $[-1, 1]$. The attribute has the average $\langle x \rangle = (\rho + \sqrt{1 - \rho^2}) \langle k \rangle$, while its standard deviation is the same as that of degrees, i.e., $\sigma_x = \sigma_k$, independent of ρ . From the generated attribute sequence, one can measure the attribute-attribute correlation r_{xx} using Eq. (13) but with k replaced by x . r_{xx} can be interpreted as the degree of attribute homophily [29]. For comparison to the analytic solution in Eq. (10), we assume the Gamma distribution for the degree as in Eq. (9). Since the analytic results are not sensitive to the variation of α , we use $\alpha = 1$ for simplicity.

Let us first consider the uncorrelated case, i.e., $r_{kk} = \rho_{kx} = 0$. We generate an uncorrelated network of size $N = 10^5$ and of $\langle k \rangle = \langle x \rangle = 50$. Then we measure the paradox holding probability $h_k(x)$ to find that the numerical result in Fig. 2(e) supports our analytic solution of Eq. (10), also depicted in Fig. 1(a). The values of $h_k(x)$ for $x/\langle x \rangle = 0.9, 1$, and 1.1 are plotted in Fig. 1(b) for the precise comparison to the analytic solution. In all cases, $h_k(x)$ has been averaged over 100 different assignments of attributes using Eq. (14).

In general, the paradox holding probability is expected to be affected by the combined effect of two correlations, i.e., r_{kk} and ρ_{kx} . As shown in Figs. 2(d)–2(f), when $\rho_{kx} = 0$, the overall behavior of $h_k(x)$ is the same as the uncorrelated case in Fig. 1(a), irrespective of r_{kk} . It is because attributes of neighboring nodes are fully uncorrelated, supported by the observation of $r_{xx} \approx 0$. By the same argument, the similar pattern is observed for $r_{kk} = 0$ and $\rho_{kx} \neq 0$. This is evidenced by the fact that the border x_k , defined by the condition $h_k(x = x_k) = 1/2$, is mostly flat for a wide range of k . However, such borders show some deviations from $x = \langle x \rangle$, depicted by blue horizontal lines in Fig. 2, possibly due to finite size effects.

When both r_{kk} and ρ_{kx} are positive [Fig. 2(c)], the effect of attribute homophily by $r_{xx} > 0$ becomes pervasive. The GFP holds for high-attribute nodes due to their neighbors of even higher attributes, while low-attribute nodes have lower paradox

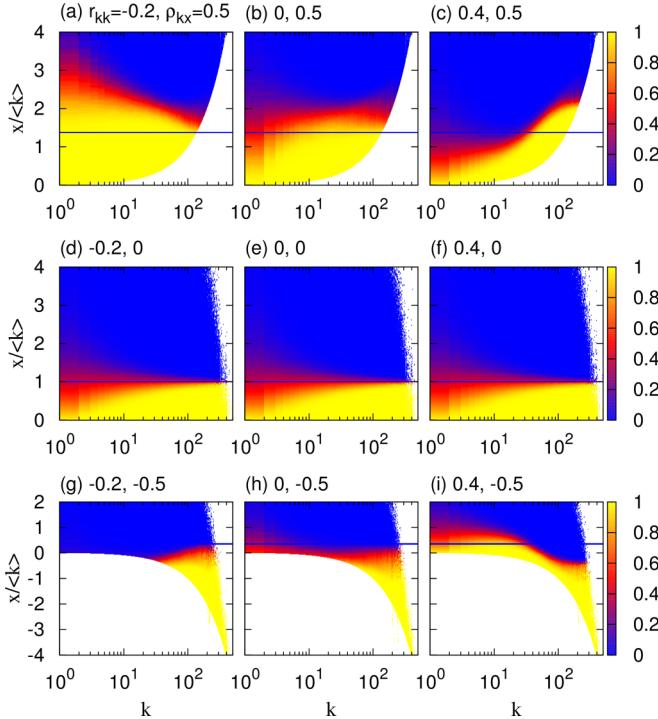


FIG. 2. (Color online) Paradox holding probability $h_k(x)$ of the correlated networks of size $N = 10^5$ for values of $r_{kk} = -0.2, 0$, and 0.4 (from left to right) and of $\rho_{kx} = -0.5, 0$, and 0.5 (from bottom to top). Degrees k follow the Gamma distribution in Eq. (9) with $\alpha = 1$ and $\beta = 50$, i.e., $\langle k \rangle = 50$, and attributes x are assigned to nodes using Eq. (14). For comparison to the uncorrelated case, x has been regularized by $\langle k \rangle$ that has the same value as $\langle x \rangle$ for $\rho_{kx} = 0$. Blue horizontal lines correspond to $\langle x \rangle / \langle k \rangle$ for each case.

holding probability, compared to the uncorrelated case. The opposite behavior is observed for the dissorative networks [Fig. 2(a)]. Hub nodes of high attribute tend to be connected with dangling nodes of low attribute, leading to smaller $h_k(x)$ for the former and larger $h_k(x)$ for the latter. It also means the negative attribute-attribute correlation ($r_{xx} < 0$). Let us now consider when degrees and attributes are negatively correlated ($\rho_{kx} < 0$). In the assortative networks [Fig. 2(i)], the GFP holds even for some high attribute nodes but with small degrees, which is comparable to the case of $r_{kk}, \rho_{kx} > 0$. In the dissorative networks [Fig. 2(g)], hub nodes of low attribute tend to be connected to dangling nodes of high attribute, leading to larger $h_k(x)$ for the former and smaller $h_k(x)$ for the latter. This is in contrast to the case of $r_{kk} < 0$ and $\rho_{kx} > 0$. It is notable that the results for $r_{xx} \approx 0$ and for $r_{kk}, \rho_{kx} > 0$ are comparable to empirical results for coauthorship networks of PR and GS in Figs. 1(d), 1(f) and Figs. 1(a), 1(c) of Ref. [18], respectively.

Now we calculate the average paradox holding probability $H(r_{kk}, \rho_{kx})$, which is defined as the fraction of nodes satisfying Eq. (5). The result is shown in Fig. 3(a). As a reference, we define $H_0 \equiv H(0, 0) \approx 0.62$ for the uncorrelated case. If $r_{kk} \lesssim 0.4$, it is found that $H > H_0$ ($H < H_0$) for $\rho_{kx} > 0$ ($\rho_{kx} < 0$). Otherwise, if $r_{kk} > 0.4$, $H \approx H_0$ is observed for almost entire range of ρ_{kx} . We first note that most nodes in the network have small degrees from the Gamma distribution, and

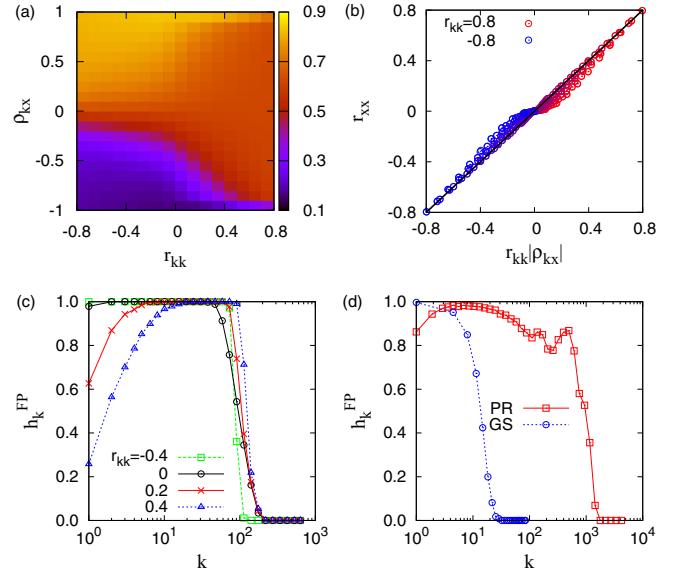


FIG. 3. (Color online) Numerical results for correlated networks of size $N = 10^5$ and of $\langle k \rangle = 50$ with the Gamma distribution for degrees (a)–(c): (a) Average paradox holding probability H as a function of r_{kk} and ρ_{kx} . (b) Scatter plot showing r_{xx} and $r_{kk}|\rho_{kx}|$ for $-0.8 \leq r_{kk} \leq 0.8$. The solid line corresponds to $r_{xx} = r_{kk}|\rho_{kx}|$. (c) Paradox holding probability of the FP for various values of degree-degree correlations. (d) Empirical paradox holding probability of the FP for coauthorship networks of Physical Review (PR) journals and Google Scholar (GS) profiles from Ref. [18].

they have low attributes if $\rho_{kx} \geq 0$ or high attributes but around 0 for $\rho_{kx} < 0$. These nodes dominate the population, hence the behavior of H . Next, the paradox holding probability of such dominant nodes needs to be understood. In the dissorative networks ($r_{kk} < 0$), large degree nodes tend to be connected to small degree nodes, leading to a starlike structure. If hub nodes have high attributes and peripheral nodes have low attributes ($\rho_{kx} > 0$), the dominant nodes, i.e., peripheral nodes in this case, have large paradox holding probability, resulting in $H > H_0$. Otherwise, if $\rho_{kx} < 0$, since the dominant nodes have high attribute, we find $H < H_0$. Here the attributes of neighboring nodes are negatively correlated ($r_{xx} < 0$) irrespective of the sign of ρ_{kx} . In the assortative networks ($r_{kk} > 0$), nodes of similar degrees tend to be connected to each other. The attributes of neighboring nodes are similar ($r_{xx} > 0$) whether high (low) degree nodes have high (low) attributes ($\rho_{kx} > 0$) or vice versa ($\rho_{kx} < 0$). In either case, the dominant nodes have neighbors of similar attribute, implying that the behavior of H is robust against the variation and sign of ρ_{kx} . Conclusively, the sign of ρ_{kx} is relevant to H in the dissorative network with $r_{kk} < 0$, while it is irrelevant to H in the assortative network with $r_{kk} > 0$. This can be compared to the GFP at the network level, which is determined by the sign of ρ_{kx} as shown in Eq. (3). We also numerically find that $r_{xx} \approx r_{kk}|\rho_{kx}|$ in Fig. 3(b), implying that the behavior of H cannot be explained only in terms of r_{xx} .

Finally, using the above generated networks, we calculate the probability of holding the FP, denoted by h_k^{FP} . As shown in Fig. 3(c), for $r_{kk} \leq 0$, h_k^{FP} stays close to 1 until k reaches ≈ 100 , and decays quickly to 0. It is because small-degree

nodes tend to be connected to large-degree nodes. However, in the assortative networks with $r_{kk} > 0$, h_k^{FP} begins with small values, increases according to k , and eventually decays to 0. It implies that the FP holds most strongly for nodes of average degree, or so-called middle class, not for nodes of the smallest degree. These variations at the individual level are observed only due to different effects of assortativity coefficient, r_{kk} . In contrast, the FP at the network level is influenced only by the shape of degree distribution, irrespective of r_{kk} . These results enable us to understand the empirical finding of h_k^{FP} from coauthorship networks [18], replotted in Fig. 3(d). The increasing behavior of h_k^{FP} for $k < 10$ in the coauthorship network of PR is due to $r_{kk} \approx 0.47$, while such increasing behavior is not observed in the coauthorship network of GS showing no degree-degree correlation, i.e., $r_{kk} \approx -0.02$.

D. Case with scale-free networks

In order to study the GFP in a more realistic setup such as scale-free networks, we generate the correlated networks using the power-law distribution of degrees and attributes. In case of power-law degree distribution, the degree-degree correlation r_{kk} is strongly limited by various factors, such as the system size and the power-law exponent of degree distribution, as studied in Ref. [30]. For the realistic consideration, we choose

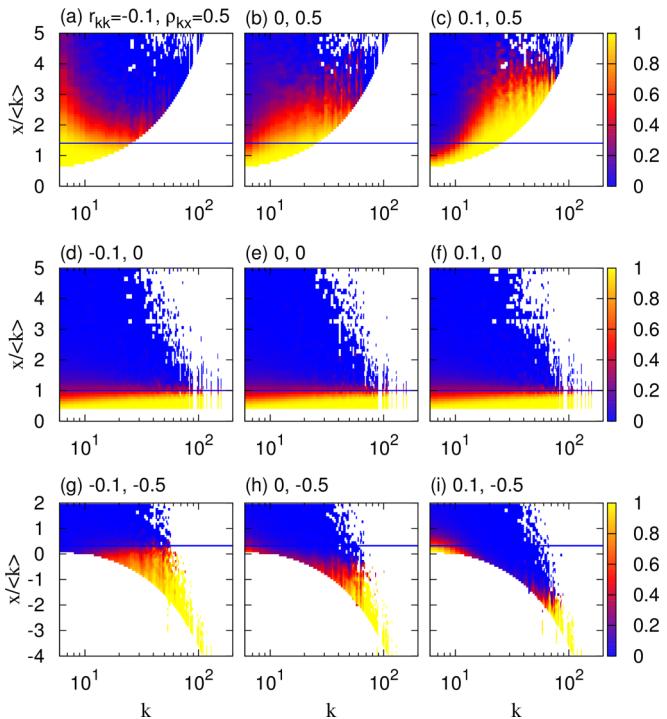


FIG. 4. (Color online) Paradox holding probability $h_k(x)$ of the correlated networks of size $N = 10^4$ for values of $r_{kk} = -0.1, 0$, and 0.1 (from left to right) and of $\rho_{kx} = -0.5, 0$, and 0.5 (from bottom to top). Degrees k follow the power-law distribution in Eq. (15) with $\gamma = 2.7$ and $k_{\min} = 6$, and attributes x are assigned to nodes using Eq. (14). For comparison to the uncorrelated case, x has been regularized by $\langle k \rangle$ that has the same value as $\langle x \rangle$ for $\rho_{kx} = 0$. Blue horizontal lines correspond to $\langle x \rangle / \langle k \rangle$ for each case.

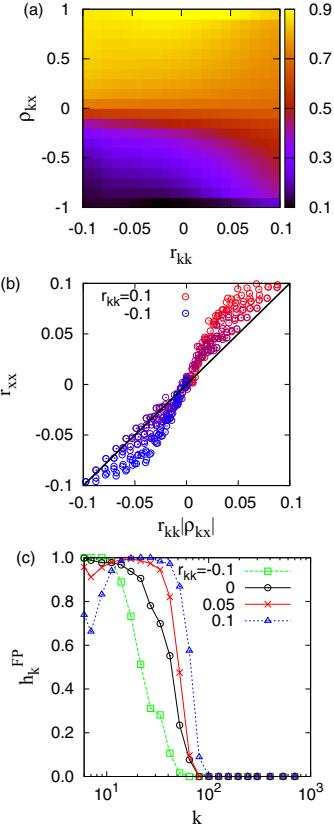


FIG. 5. (Color online) Numerical results for correlated networks of size $N = 10^4$ with the power-law distribution for degrees: (a) Average paradox holding probability H as a function of r_{kk} and ρ_{kx} . (b) Scatter plot showing r_{xx} and $r_{kk}|\rho_{kx}|$ for $-0.1 \leq r_{kk} \leq 0.1$. The solid line corresponds to $r_{xx} = r_{kk}|\rho_{kx}|$. (c) Paradox holding probability of the FP for various values of degree-degree correlations.

the following distribution

$$P(k) \propto k^{-\gamma} \quad \text{for } k \geq k_{\min}, \quad (15)$$

with $\gamma = 2.7$ and $k_{\min} = 6$. For these values of parameters, one can generate the network in the range of $-0.1 \leq r_{kk} \leq 0.1$ for $N = 10^4$. Then, we calculate the paradox holding probability $h_k(x)$ to find that its overall behavior is qualitatively similar to those in the case of Gamma distribution, as shown in Fig. 4. We also find similar behaviors for the average paradox holding probability $H(r_{kk}, \rho_{kx})$, for the linear relationship between r_{xx} and $r_{kk}|\rho_{kx}|$ but with larger deviations due to the relatively narrow range of r_{kk} , and for the probability of holding the FP for various values of degree-degree correlation. The results are summarized in Fig. 5.

III. CONCLUSIONS

As an interplay between topological heterogeneities and node attributes in complex networks, the generalized friendship paradox (GFP) has been recently suggested, implying that your friends have on average higher attribute than you do [18]. While the GFP at the network level was clearly explained in terms of the positive degree-attribute correlations, the GFP at the individual level has been far from being fully understood. In order to understand the role of degree-attribute correlations for

the GFP at the individual level in more detail, we analyze the uncorrelated solvable model, which already shows nontrivial behavior especially for high-attribute nodes. For the general case, we numerically study the correlated network model with tunable degree-degree and degree-attribute correlations, denoted by r_{kk} and ρ_{kx} , respectively. We obtain the detailed patterns of the paradox holding probability of individuals depending on their degrees and attributes, for the entire range of correlations of r_{kk} and ρ_{kx} . Similarly to the GFP at the network level, the average paradox holding probability is strongly affected by the sign of ρ_{kx} only in the assortative networks with $r_{kk} < 0$. On the other hand, the results for the assortative networks with $r_{kk} > 0$ are robust against the variation and sign of ρ_{kx} .

In our study, we have ignored other topological heterogeneities of networks, such as community structure [14], and assumed that node attributes are fixed and do not change. In future works, it would be interesting to study the GFP in more realistic network topology and/or in cases where the attributes

can change in time, such as the attractiveness of scientific papers [31], or they evolve according to the individual decisions, e.g., within the framework of evolutionary game theory [32].

Finally, we like to remark that successful applications of statistical physics to social phenomena necessitate the detailed understanding of both objective and subjective sides of individual behaviors. In this sense, our study of the GFP can provide insights for the subjective self-evaluation of individuals compared to their neighbors [6,7], which shapes the way they interact with others. This is crucial to understand the emergent collective dynamics at the community or societal level.

ACKNOWLEDGMENTS

We gratefully acknowledge the Aalto University postdoctoral program (H.-H.J.) and the EC FET Open project “New tools and algorithms for directed network analysis,” NADINE number 288956 (Y.-H.E.) for financial support.

-
- [1] R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002).
- [2] C. Castellano, S. Fortunato, and V. Loreto, *Rev. Mod. Phys.* **81**, 591 (2009).
- [3] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, *Science* **323**, 892 (2009).
- [4] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann *et al.*, *Science* **323**, 721 (2009).
- [5] P. Holme and J. Saramäki, *Phys. Rep.* **519**, 97 (2011).
- [6] E. W. Zuckerman and J. T. Jost, *Soc. Psychol. Q.* **64**, 207 (2001).
- [7] E. Kross, P. Verduyn, E. Demiralp, J. Park, D. S. Lee, N. Lin, H. Shablock, J. Jonides, and O. Ybarra, *PLoS ONE* **8**, e69841 (2013).
- [8] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* (Cambridge University Press, Cambridge, 2010).
- [9] Flavio L. Pinheiro, Marta D. Santos, Francisco C. Santos, and Jorge M. Pacheco, *Phys. Rev. Lett.* **112**, 098702 (2014).
- [10] S. M. Garcia, A. Tor, and T. M. Schiff, *Perspect. Psychol. Sci.* **8**, 634 (2013).
- [11] M. Buchanan, *The Social Atom: Why the Rich Get Richer, Cheaters Get Caught, and Your Neighbor Usually Looks Like You* (Bloomsbury Publishing, London, 2008).
- [12] A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).
- [13] M. E. J. Newman, *Phys. Rev. Lett.* **89**, 208701 (2002).
- [14] S. Fortunato, *Phys. Rep.* **486**, 75 (2010).
- [15] S. L. Feld, *Am. J. Sociol.* **96**, 1464 (1991).
- [16] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, *arXiv:1111.4503*.
- [17] N. O. Hudas, F. Kooti, and K. Lerman, in *Proceedings of 7th International AAAI Conference on Weblogs and Social Media, MA, USA, ICWSM’13* (The AAAI Press, Palo Alto, CA, USA, 2013).
- [18] Y.-H. Eom and H.-H. Jo, *Sci. Rep.* **4**, 4603 (2014).
- [19] F. Kooti, N. O. Hudas, and K. Lerman, in *Proceedings of 8th International AAAI Conference on Weblogs and Social Media, Cambridge, MA, USA* (The AAAI Press, Palo Alto, CA, USA, 2014).
- [20] T. Grund, *Sociol. Sci.* **1**, 128 (2014).
- [21] R. Cohen, S. Havlin, and D. ben-Avraham, *Phys. Rev. Lett.* **91**, 247901 (2003).
- [22] N. A. Christakis and J. H. Fowler, *PLoS ONE* **5**, e12948 (2010).
- [23] M. Garcia-Herranz, E. Moro, M. Cebrian, N. A. Christakis, and J. H. Fowler, *PLoS ONE* **9**, e92413 (2014).
- [24] Y. Kryvasheyeu, H. Chen, E. Moro, P. Van Hentenryck, and M. Cebrian, *arXiv:1402.2482*.
- [25] J. Park and A.-L. Barabási, *Proc. Nat. Acad. Sci. USA* **104**, 17916 (2007).
- [26] J. H. Fowler, C. T. Dawes, and N. A. Christakis, *Proc. Nat. Acad. Sci. USA* **106**, 1720 (2008).
- [27] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, *Phys. Rev. E* **71**, 027103 (2005).
- [28] S. Maslov and K. Sneppen, *Science* **296**, 910 (2002).
- [29] M. McPherson, L. Smith-Lovin, and J. M. Cook, *Ann. Rev. Sociol.* **27**, 415 (2001).
- [30] J. Menche, A. Valleriani, and R. Lipowsky, *Phys. Rev. E* **81**, 046103 (2010).
- [31] Y.-H. Eom and S. Fortunato, *PLoS ONE* **6**, e24926 (2011).
- [32] C. Hauert and G. Szabó, *Am. J. Phys.* **73**, 405 (2005).

Move ordering and communities in complex networks describing the game of go

Vivek Kandiah^{1,2}, Bertrand Georgeot^{1,2}, and Olivier Giraud^{3,a}

¹ Université de Toulouse, UPS, Laboratoire de Physique Théorique (IRSAMC), 31062 Toulouse, France

² CNRS, LPT (IRSAMC), 31062 Toulouse, France

³ LPTMS, CNRS and Université Paris-Sud, UMR 8626, Bât. 100, 91405 Orsay, France

Received 23 July 2014 / Received in final form 5 September 2014

Published online 22 October 2014 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2014

Abstract. We analyze the game of go from the point of view of complex networks. We construct three different directed networks of increasing complexity, defining nodes as local patterns on plaquettes of increasing sizes, and links as actual successions of these patterns in databases of real games. We discuss the peculiarities of these networks compared to other types of networks. We explore the ranking vectors and community structure of the networks and show that this approach enables to extract groups of moves with common strategic properties. We also investigate different networks built from games with players of different levels or from different phases of the game. We discuss how the study of the community structure of these networks may help to improve the computer simulations of the game. More generally, we believe such studies may help to improve the understanding of human decision process.

1 Introduction

The study of complex networks has become more and more important in the recent past. In particular, communication and information networks have become ubiquitous in everyday life. New tools have been created to understand the mechanisms of growth of such networks and their generic properties. On the other hand, it has been realized that other phenomena can also be modeled by such tools, e.g. in social sciences, linguistics, and so on [1–3].

However, the tools of complex networks were never applied to the study of human games. Nevertheless, games represent one of the oldest human activities, and may give insight into the human decision-making processes. In reference [4], a network was built that describes the game of go, one of the oldest and most famous board games. The complexity of the game is such that no computer program has been able to beat a good player, in contrast with chess where world champions have been bested by game simulators. It is partly due to the fact that the total number of possible allowed positions in go is about 10^{171} , compared to e.g. only 10^{50} for chess [5]. In fact, among traditional board games it has by far the largest state space complexity [6]. Part of the complexity of the game of go comes from this large number of different board states, due to the fact that it is played on a board (the goban) composed of 19 vertical lines and 19 horizontal lines, implying 361 possible positions, against 64 in chess. Also, it is very

hard for a computer to evaluate the positional advantages in the course of the game, while in chess the capture of different pieces can be easily compared.

Due to that, the study of computer go has become an important subfield of computer science. Its main challenge is to estimate a value function of moves, that is, a function which assigns a value to each move, given a certain state of the goban. Traditional approaches evaluate the value function by using huge databases of patterns, from initial patterns to life-and-death situations, and can learn to predict the value of moves by reinforcement learning (see e.g. [7]). By contrast, the recently introduced Monte-Carlo go does not rest primarily on expert knowledge. Its basic principle is to evaluate the value of a move by playing at random, according to the rules of go, from a given state, until the end, so that a value can be assigned to the move. Playing thousands of games allows to estimate the value function for each move. This approach has proved way more efficient than the classical approaches [8,9].

Many improvements have since then been included in Monte-Carlo go. In particular, Monte-Carlo tree search, implemented in computer programs such as Crazy Stone [10] or MoGo [11], is based on the construction of a tree of goban states, where new states are added iteratively as they are met in a simulation. The value function is updated depending on the outcome of each randomly played game. Random moves are chosen according to some playing policy which can itself be biased towards certain moves (for instance, capture whenever possible), and in such a way that most promising moves are more carefully

^a e-mail: olivier.giraud@lptms.u-psud.fr

explored, but with an incentive to visit moves with a large uncertainty on their actual value. Recent improvements allow to improve the exploration of the tree [12]. To get faster estimates of the value of a move the RAVE (Rapid Action Value Estimation) algorithm, or its Monte-Carlo version, attributes to a move in a given state s the average outcome of all games where that move is played *after* state s has been encountered [13]. In order to account for rarely visited states, a heuristic prior knowledge can be fed into the algorithm to attribute an a priori value to a move, such as e.g. the value of its grand-father, or a value depending on local patterns [14].

Although global features, such as chain connections, or the influence of stones over domains of the goban, are crucial in the game of go, local features can be used at many places in the algorithms of computer go, for instance to improve the heuristic value function which initializes the value of each move, or to get a faster estimate of the exact value [15,16].

There is therefore a clear interest in having a better understanding of local features in the game of go. In reference [4], two of us introduced a small network based on local positional patterns and showed that it can be used to extract information on the tactical sequences used in real games. However, the small size of the plaquettes made it difficult to disambiguate many strategically different moves. In the present paper, we construct three networks based on positional patterns of different sizes, and study their properties. The network size varies by a factor one hundred, and the largest one enables to specify more precise features that were difficult to disambiguate in reference [4]. In particular, the community structure is much easier to characterize and discuss. After presenting the details of the construction of the networks (Sect. 2) we study their global properties such as ranking vectors and spectra of the Google matrix, contrast them to other types of networks, and relate them to specific features of the game (Sect. 3). In Section 4, we study in detail the characterization of communities of nodes in the networks, a well-known subject in network theory, which in our case enables to regroup tactical moves with common features. In Section 5 we propose the construction of different networks corresponding to specific phases of the game or to different levels of players.

2 The go networks

The game of go is played on a board (goban) of 19×19 intersections of vertical and horizontal lines. Each player alternately places a stone of his/her color (black or white) at an empty intersection. Empty intersections next to a group of connected stones of the same color are called “liberties”. If only one liberty remains, the group of stones is said to be in atari. When the last liberty is occupied and the group is entirely surrounded by the opponent, its stones must be removed. The aim of the game is to surround large territories and to secure their possession. Good players follow general strategies through a series of

local tactical fights. We construct the networks representing the game by connecting local moves played in the same neighbourhood (note the similarity with some language networks [17–19] which are also based on local features). We describe a move by identifying the empty intersection (h, v) (with $1 \leq h, v \leq 19$) where the new stone is placed.

The vertices of our networks are based on what we call “plaquettes”, i.e. a part of the goban with a given shape and size which depends on the network. Each plaque corresponds to a certain pattern of white and black stones with an empty intersection at its center, on which black will put a stone. We identify plaquettes which are related by translation on the goban or by a symmetry of the square, and additionally those with colors swapped.

The first network we consider (Network I) is made as in reference [4] by taking as plaquettes squares of 3×3 intersections, which are subparts of the goban of the form $\{(h + r, v + s), -1 \leq r, s \leq 1\}$ (edges and corners of the board can be accounted for by imagining additional dummy lines outside the board). Once borders and symmetries are taken into account, we obtain as vertices of network I a total of 1107 nonequivalent plaquettes (with empty centers).

Network II is made by also taking squares of 3×3 intersections and identifying plaquettes related by symmetry, but we also include the atari status of the four nearest-neighbour points from the center. Atari status assesses if the chain of stones to which a given stone belongs has only one liberty (one empty intersection connected to it). Removing the last liberty of a chain in atari entails the capture of the whole group. In this case, many seemingly possible configurations are not legal since they would contradict the atari status. This leaves 2051 legal nonequivalent plaquettes with empty centers (the same figure was found in Ref. [20]).

Network III is based on diamond-shape plaquettes: the 3×3 plaquettes discussed above plus the four at distance two from the center in the four directions left, right, top, down. We still identify plaquettes related by symmetry, but do not take into account the atari status. This gives us 193 995 nonequivalent plaquettes with empty centers, which are the vertices of network III (96 771 are so rare that they are actually never used in our database of games).

We have identified the occurrence of these different plaquettes in games from a database available at [21]. This database contains the sequence of moves of 135 663 different games corresponding to players of diverse levels (the level of the players is marked by a number of dans, from 1 to 9). The games recorded have been played online, and the dans have been mutually assessed according to the results of these plays. The frequency of the different plaquettes is shown in Figure 1. It can be compared to Zipf’s law, an empirical law seen in many natural distributions (word frequency, city sizes, chess openings...) [22–25]. For items ranked according to their frequency, it corresponds to a power-law decay of the frequency versus the rank. The data presented in Figure 1 show that the three different network choices all give rise to a distribution following

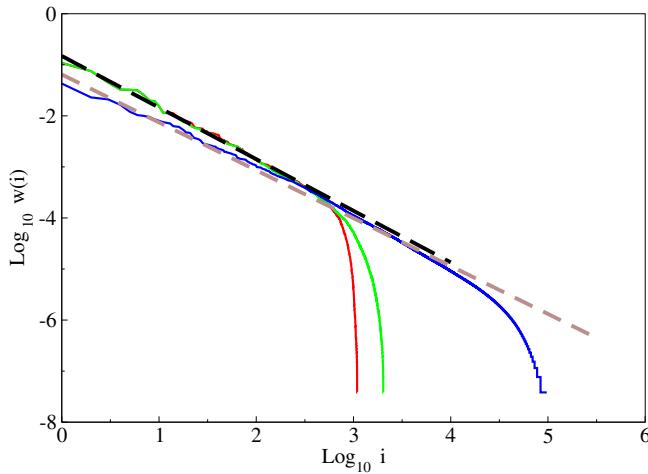


Fig. 1. Distribution of frequency of occurrences $w(i)$ of different plaquettes for the three different networks (full lines), from left to right at the bottom: red: square plaquettes (network I), green: square plaquettes with atari status (network II), blue: diamond plaquettes (network III) (see text) (data from networks I and II are indistinguishable over parts of the curves). The dashed straight lines are power law fits with slopes -1.02 (black upper line, fit of network II) and -0.94 (brown lower line, fit of network III).

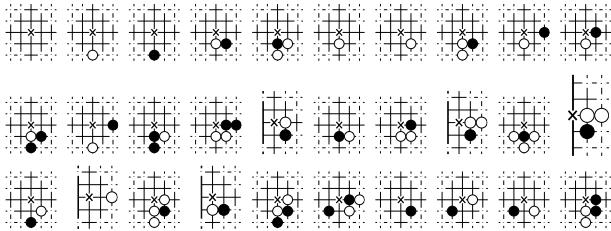


Fig. 2. Top 30 plaquettes in frequency of occurrences for the network III (diamond plaquettes). Black plays at the black cross. Dotted intersections are outside the diamond plaquette and their status is unknown.

Zipf's law, although the slope varies from ≈ -1 (networks I and II) to a slightly slower decay for the largest network (network III).

We display in Figure 2 the top 30 moves in order of decreasing frequency of occurrences for network III. The most common correspond to few stones on the plaquettes, which is natural since these ones are present at the beginning of almost all local fights, while the subsequent moves differ from games to games.

To define links of our three networks, we connect vertices corresponding to moves a and b played at (h_a, v_a) and (h_b, v_b) on the board if b follows a in a game of the database and $\max\{|h_b - h_a|, |v_b - v_a|\} \leq d$, where d is some distance. Here contrary to [4] we put a link only between a and the first move following a in the specified zone. Each integer d corresponds to a different network. It specifies the distance beyond which two moves are considered unrelated. In reference [4], different values of d were considered and it was shown that the value $d = 4$ was the most relevant,

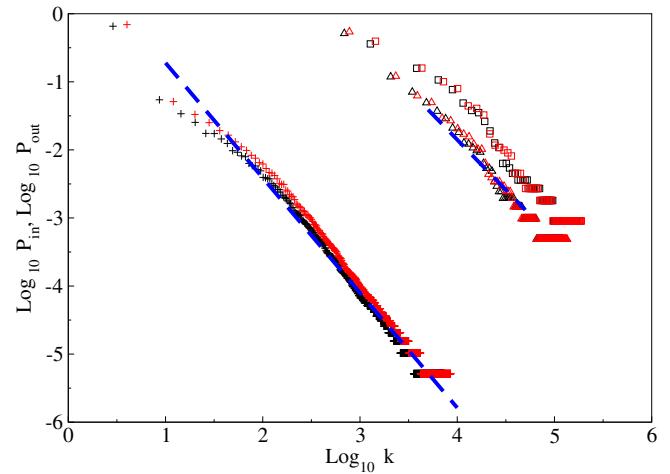


Fig. 3. Distribution of incoming links P_{in} (black) and outgoing links P_{out} (red/grey) for the three different networks; square plaquettes (network I) (squares), square plaquettes with atari (network II) (triangles), diamond plaquettes (network III) (crosses). The dashed lines are power law fits with slopes -1.47 (right) and -1.69 (left).

allowing a correct hierarchization of moves: related local fights are kept while far away tactical moves are not taken into account. In the following we will thus retain this value $d = 4$. Two vertices are thus connected by a number of directed links given by the number of times the two corresponding moves follow each other in the same neighbourhood of the goban in the games of the database.

With this definition, the three networks are now defined, with vertices connected by directed links. The total number of links including degeneracies is 26 116 006 links. The numbers without degeneracies are, respectively, 558 190 (network I), 852 578 (network II) and 7 405 395 (network III). The link distributions are shown in Figure 3; it is close to a power-law. This implies that the networks present the scale-free property [1–3]. One can notice a symmetry between ingoing and outgoing links, which is a peculiarity of this problem, and is not seen in e.g. the World Wide web, where the exponent for P_{out} (≈ -2.7) is different from the one for P_{in} (≈ -2.1) [26,27]. Here exponents are similar and close to 1.5, intermediate between these two values. Our results indicate the presence of a symmetry (at least at a statistical level) between moves that follow many different others and moves which have many possible followers. This symmetry is natural, since in many cases (i.e. in the course of a local fight) the occurrence of a plaquette in the database implies the presence of both an ingoing and an outgoing link.

3 Ranking vectors and spectra of Google matrices

We have presented up to now the construction of our networks for the game of go, and their global statistical properties. To get more insight into the organization of the

game, we use tools developed in the framework of network theory, in order to hierarchize vertices of a network. Such tools are routinely used by search engines to decide in which order answers to queries are presented. The general strategy is to build a ranking vector, whose value on each vertex will measure its importance. A famous vector of this type is the PageRank [28,29], which has been at the basis of the Google search engine. It can be obtained from the Google matrix G , defined as $G_{ij} = \alpha S_{ij} + (1 - \alpha)^t ee/N$, where $e = (1, \dots, 1)$, N is the size of the network, α is a parameter such that $0 < \alpha \leq 1$ (we chose $\alpha = 1$ in the computations in this paper), and S is the weighted adjacency matrix. The latter starts from the adjacency matrix where the value of the entry (i, j) corresponds to the number of links from vertex j to vertex i ; then one replaces any column of 0 by a column of 1, and one normalizes the sum of each column to 1. This ensures that the matrix G has the mathematical property of stochasticity. The PageRank vector is defined as the right eigenvector of the matrix G associated with the largest eigenvalue $\lambda = 1$. It singles out as important vertices the ones with many incoming links from other important nodes. Equivalently, it can be seen as giving the average time a random surfer on the network will spend on each vertex. Indeed, the process of iterating G can be seen as the action of a random surfer choosing randomly at each node to follow a link to another node. The largest eigenvalue corresponds to the equilibrium distribution of the surfer, and gives the average time spent on each node. Other ranking vectors which can be built from the graph include the CheiRank vector [30,31], and the Hubs and Authorities of the HITS algorithm [32]. While PageRanks and Hubs attribute importance to vertices depending on their incoming links, CheiRanks and Authorities stem from outgoing links. In particular, CheiRank can be defined as the PageRank of the “dual” network where all links are inverted. We denote the Google matrix of this dual network by G^* .

In Figure 4 the distributions of PageRank and CheiRank are shown for the three networks, showing that ranking vectors follow an algebraic law, with a slightly different exponent for the largest network. Similarly as for the link distribution, one sees a symmetry between distributions of ranking vectors based on ingoing links and outgoing links, again an original feature which can be related to the statistical symmetry between ingoing and outgoing links and the fact that at lowest approximation ranking vectors can be approximated by in- or outgoing links [33].

In order to check to what extent this symmetry affects the ranking vectors, we plot in Figure 5 the CheiRank K^* as a function of the PageRank K . It indeed shows that the two quantities are not independent, and strong correlations between PageRank and CheiRank do exist. This symmetry is not visible in general for other networks (see e.g. [34] where similar plots are shown in the context of world trade, displaying much less correlation). Nevertheless, the symmetry is clearly not exact, especially for the largest network (a perfect correlation will produce points only on the diagonal); the plots are not even symmetric with respect to the diagonal. Thus PageRank and

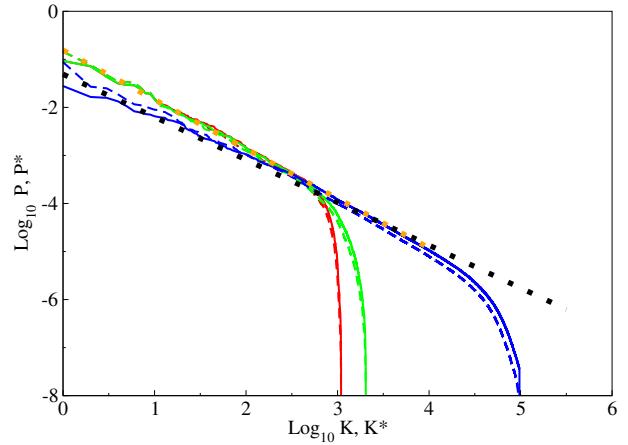


Fig. 4. Distribution of ranking vectors (normalized by $\sum_K P(K) = \sum_{K^*} P^*(K^*) = 1$) for the three different networks: PageRank $P(K)$ (solid lines) and CheiRank $P^*(K^*)$ (dashed lines), same color code for the networks as in Figure 1 (data from networks I and II are indistinguishable over parts of the curves). The dotted lines are power law fits with slopes -1.03 (orange upper line, fit of network II) and -0.89 (black lower line, fit of network III).

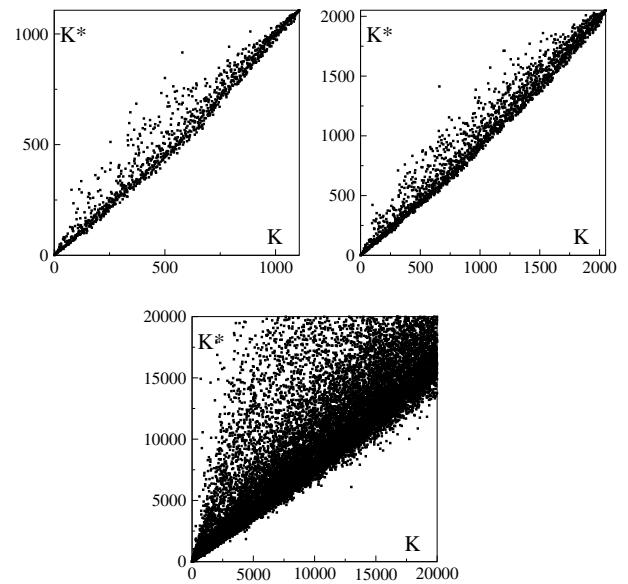


Fig. 5. PageRank-CheiRank correlation plot of the three different networks: square plaquettes (network I)(top left), square plaquettes with atari status (network II)(top right) and diamond plaquettes (network III)(bottom). PageRank K is given in x -axis and CheiRank K^* in y -axis, the plot of network III is a zoom on the top 20 000 moves in both K and K^* .

CheiRank produce genuinely different information on the network.

Figure 6 shows the first 30 plaquettes in decreasing importance in the PageRank and CheiRank vectors. The correlation between the two sequences is clearly visible, although it is again not perfect. We note that these sequences are also very similar to the one obtained by just

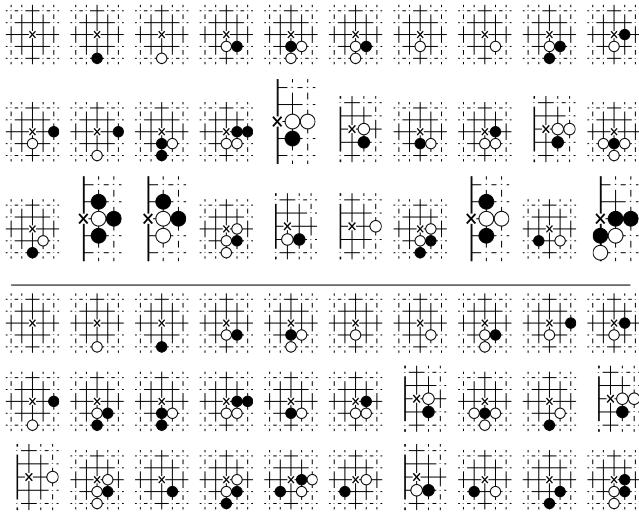


Fig. 6. Top 30 plaquettes for first eigenvector of G (PageRank) (top) and G^* (CheiRank) (bottom) of the network III.

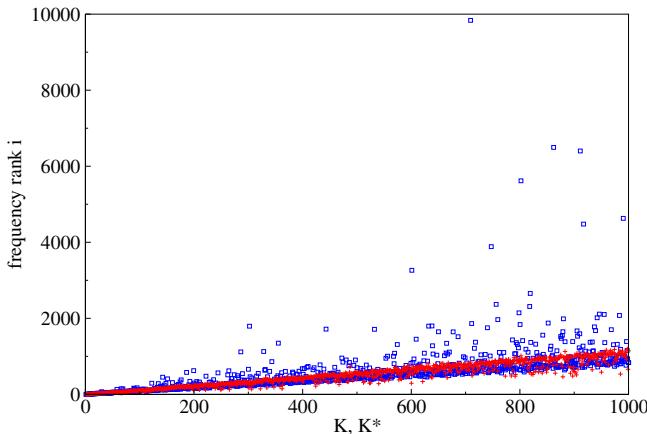


Fig. 7. Correlation plot of PageRank-CheiRank vs. frequency of moves for network III (diamond plaquettes) (only first 1000 moves in K are shown); blue squares: PageRank K , red crosses: CheiRank K^* .

counting the move frequency (as in Zipf's law): most frequent moves tend to dominate the ranking vectors.

However, as Figure 7 shows, the correlation between ranking vectors and frequency ordering is far from perfect, especially for the PageRank, which can be extremely different from the rank obtained by frequency. This shows that the ranking vectors present an information obtained from the network construction, which differs from the mere frequency count of moves in the database. Indeed, as explained above the frequency count is related to the link distribution due to the construction process of the network. It is known in general that the PageRank has some relation with the distribution of ingoing links, but with the significant difference that it highlights nodes whose ingoing links come from (recursively defined) other important nodes. This was the basis of the fortune of Google and in our case means that highlighted moves correspond to plaquettes with ingoing links coming from other important plaquettes. Thus the PageRank underlines moves to

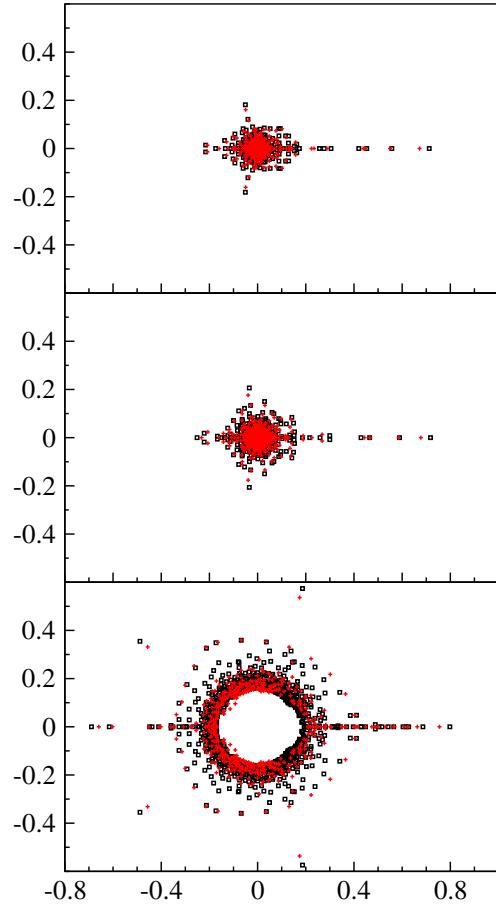


Fig. 8. Spectrum in the complex plane of G (black squares) and G^* (red/grey crosses) for the three different networks: I (top), II (middle) and III (bottom).

which converge many well-trodden paths of history in the different games of the database. The CheiRank does the same in the reverse direction, highlighting moves which open many such paths.

The ranking vectors discussed above are just one eigenvector of the matrices associated with a given network. However, other eigenvalues and their associated eigenvectors also contain information about the network. We have computed the spectrum of the Google matrix for the three networks; they are shown in Figure 8. For square plaquettes (network I) and square plaquettes plus atari status (network II) all eigenvalues are computed. In the case of the largest network, standard diagonalization techniques could not be used and therefore we used an Arnoldi-type algorithm to compute the largest few thousands eigenvalues in the complex plane. For the G matrix of the diamond network (network III), about 1000 eigenvectors were computed. For G^* matrix of diamond network, about 500 eigenvectors were computed.

Stochasticity of G and G^* implies that their spectra are necessarily inside the unit disk. For the World Wide Web the spectrum is spread inside the unit circle [35,36], with no gap between the largest eigenvalue and the bulk. For networks I and II, Figure 8 shows a huge gap between

the first and the other eigenvalues. For the third network, there is still a gap between the first eigenvalue and next ones, but it is smaller. While the distribution of the ranking vectors shown in Figure 4 reflects the distribution of links, the gap in the spectrum is related to the connectivity of the network and the presence of large isolated communities [35,36]. The presence of a large gap indicates a large connectivity, which is reasonable for the smaller networks. The presence of a smaller gap for network III indicates that there is more structure in the networks with larger plaquettes which disambiguate the different game paths and makes more visible the communities of moves. However, the gap being still present shows that even at the level of diamond-shaped plaquettes, the moves can belong to many different communities: this underlines one of the specificities of the game of go, which makes a given position part of many different strategic processes, and makes it so difficult to simulate by a computer.

The results in this section show that the tools of complex networks such as ranking vectors associated to the largest eigenvalue already give new information which clearly go beyond the mere frequency count of the moves. This could be used to make more efficient the Monte Carlo algorithms of computer go. Nevertheless, other eigenvalues also carry valuable information, that we will study in the next section.

4 Eigenvectors and communities

In the preceding section, we displayed the spectra of the networks constructed from the game of go. We have already discussed the ranking vectors associated to the largest eigenvalue. The other eigenvectors give a different information. In Figure 9 we display the intensities of the first 200 eigenvectors of the three different networks. It is clear that eigenvectors have specific features, not being spread out uniformly or localized around a single specific location. Correlations are also clearly visible between different eigenvectors, materialized by the vertical lines where several eigenvectors have similar intensities on the same node. Correlations are less visible on the largest network, but it is also due to the much largest size of the vectors which decreases the individual projections on each node. It is interesting to note that these correlations are not necessarily related to the PageRank values or the frequency of moves: vertical lines tend to be more visible on the left of the figure corresponding to high PageRank, but they are present all over the interval: certain sequences of eigenvectors have correlated peaks at locations with relatively low PageRank.

In order to quantify these effects, we first look at the spreading of eigenvectors: for a given vector, how many sites have significant projections? This can be measured for a vector ψ through the Inverse Participation Ratio (IPR): $\sum_i |\psi_i|^4 / (\sum_i |\psi_i|^2)^2$. For a vector uniformly spread over P vertices it would be equal to P . A random vector thus has an IPR proportional to the size of the system. The data of Figure 10 for the eigenvectors corresponding to the largest eigenvalues show that these vectors are not

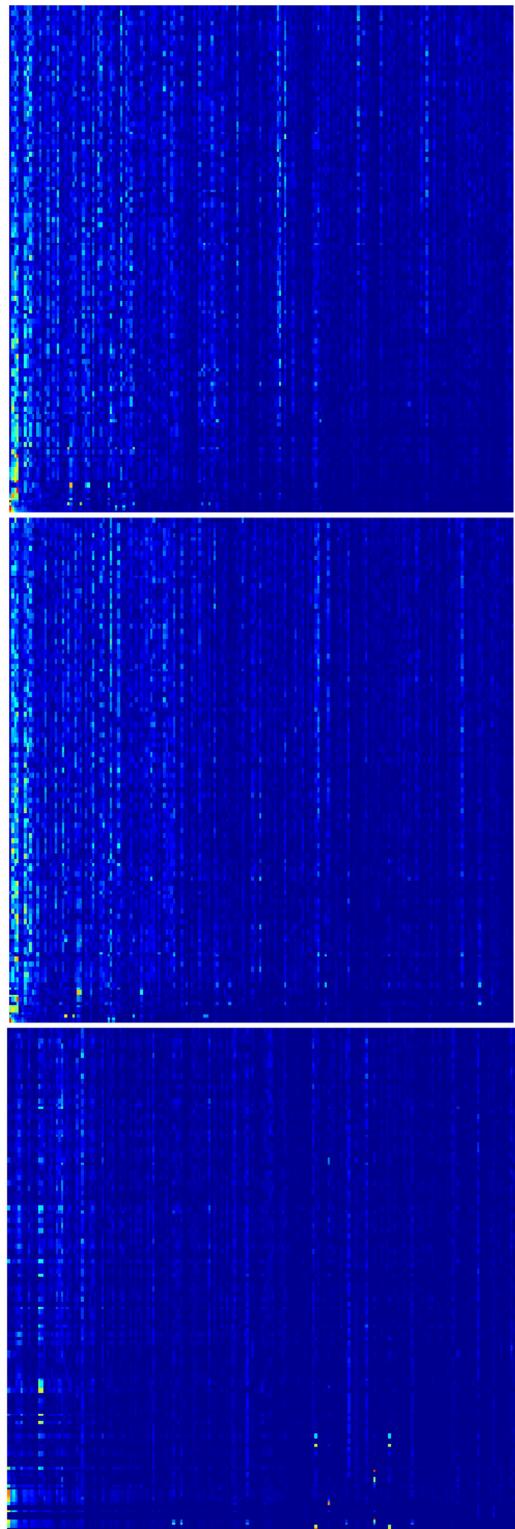


Fig. 9. Eigenvector correlation map of the matrix G for the three different networks: I (top), II (middle) and III (bottom). Top 200 eigenvectors in order of decreasing eigenvalue modulus are plotted horizontally from bottom to top. Only the first 200 components are shown in the PageRank basis. The colors are proportional to the modulus of components (the normalization of an eigenstate ψ is $\sum_i |\psi_i|^2 = 1$), from blue/dark grey (minimal) to red/light grey (maximal).

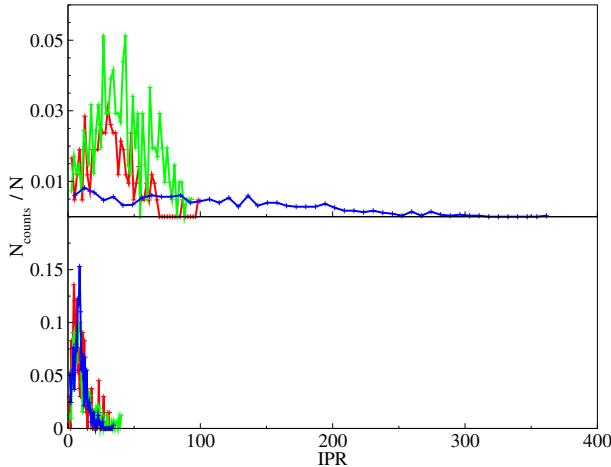


Fig. 10. Histogram of IPR values (see text) for network I (red/dark grey), network II (green/light grey) and network III (blue/black). Top panel shows the values computed for eigenvectors of G and bottom panel shows the same for G^* . Data correspond to the top 221 eigenvalues (network I), top 410 eigenvalues (network II) and top 999 eigenvalues (network III).

random or uniformly spread. On the contrary, their IPR is quite small, even for the largest network: in this case only a few dozen sites contribute to a given eigenvector, among almost 200 000 possible nodes. Figure 10 also shows that there is a relatively large dispersion of the IPR around the mean value. We provide the distributions for the Google matrices G and G^* . Qualitatively the features are similar, but there is both a lower mean value and a lower dispersion for G^* , indicating that the statistical symmetry found previously between incoming and outgoing links is indeed only approximate.

What is the meaning of these eigenvectors? If one interprets the Google matrix as describing a random walk among the nodes of the network as in the original paper [28], eigenvectors of G correspond to parts of the network where the random surfer gets stopped for some time before going elsewhere in the network. In other words, they are localized on sets of moves which are more linked together than with the rest of the network. This corresponds to so-called communities of nodes which share certain common properties (see e.g. [37]). In social networks, the importance of communities has been stressed several times and they are the subject of a large number of studies (see e.g. the review [38]). The use of the eigenvectors of G to extract the communities is one of the many available methods, which has been used already in the different context of the World Wide Web [39]. As already mentioned, eigenvectors with largest eigenvalues tend to be localized on groups of nodes where the probability is trapped for some time. This approach will thus detect communities of nodes from where it is difficult to escape, i.e. with few links leading to the outside. In parallel, the eigenvectors of G^* tend to be localized on groups of nodes with few incoming links from the outside. Figure 10 shows that this latter type of community, obtained from G^* , tends to be smaller

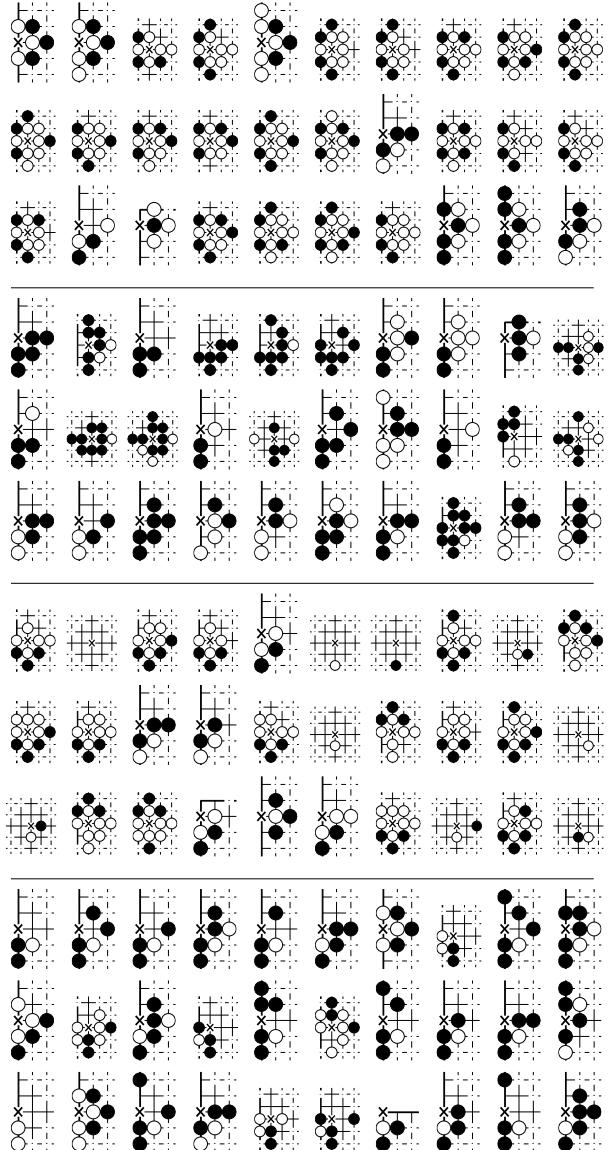


Fig. 11. Examples of the top 30 nodes where eigenvectors of G localize themselves for diamond network, from top to bottom $\lambda_7 = -0.618$, $\lambda_{11} = 0.185 - 0.5739i$, $\lambda_{13} = 0.5651$, $\lambda_{21} = -0.4380$.

on average for the go game than the former type, obtained from G . These different communities should reflect different strategic groupings of moves during the course of the game.

The concept of community being intrinsically ambiguous, one can assign a subjective meaning to the definition of the community related to a chosen method. In our case, it is a difficult task to establish clear characteristics regarding what moves should be considered belonging to which community, however in the spirit of “moves that are more played together” or “similar moves” we can observe that a single eigenvector may contain a mixing of several communities. This could explain why in Figure 9 one can see similar patterns appearing in different eigenvectors. These considerations are confirmed by Figures 11 and 12

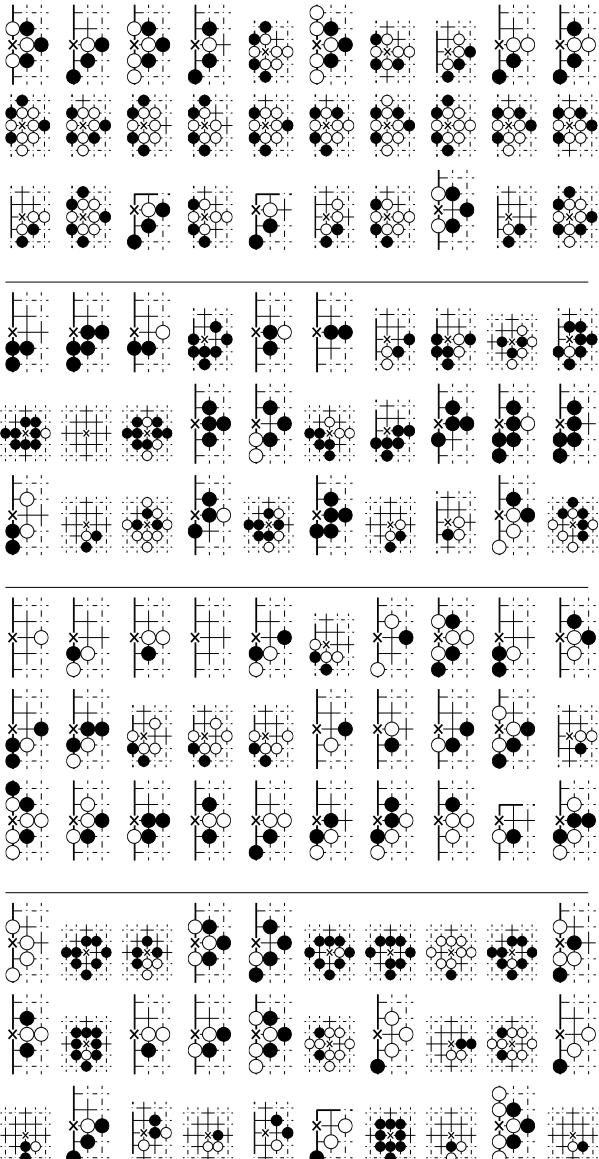


Fig. 12. Examples of the top 30 nodes where eigenvectors of G^* localize themselves for diamond network, from top to bottom $\lambda_7 = -0.6023$, $\lambda_{11} = 0.1743 - 0.5365i$, $\lambda_{18} = -0.4511$, $\lambda_{21} = -0.4021$.

where the first 30 moves of representative eigenvectors of G and G^* are displayed, ranked by decreasing component modulus. While some common features appear, one gets the impression that groups of moves corresponding to different strategic processes are mixed and should be disentangled; for instance the last example of Figure 11 seems to mix moves where black captures a white stone and moves where black connects a chain.

In principle one could use correlations as the ones shown in Figure 9 directly to identify communities, but we chose a different strategy. We propose here different basic methods that can be a first step into separating the communities within a given eigenvector. The simplest and most straightforward method consists in filtering out the

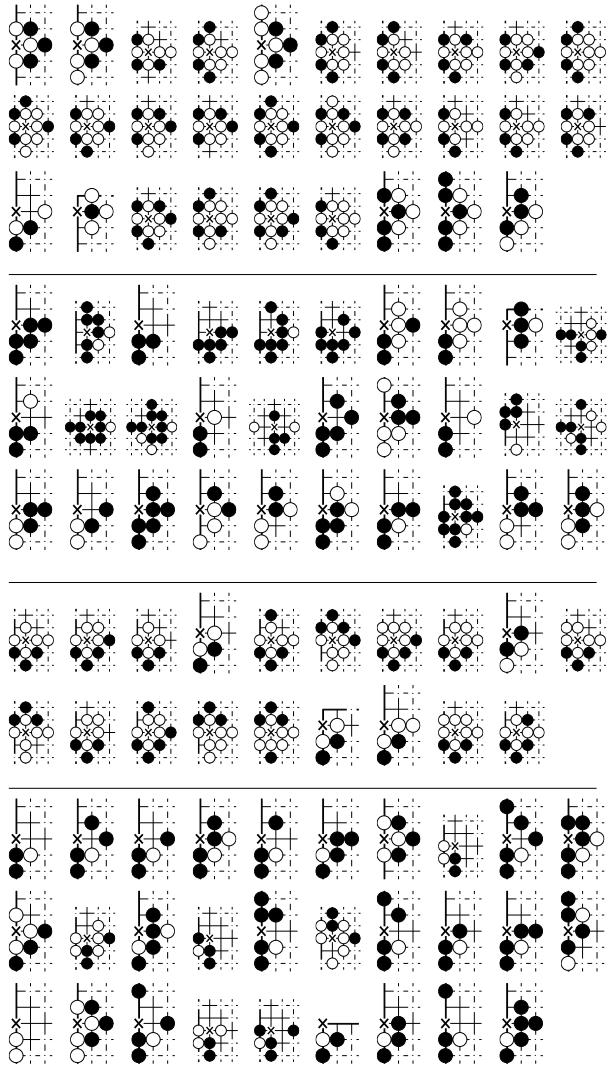


Fig. 13. Same eigenvectors as in Figure 11 treated by filtering out the top 30 PageRank moves.

effects of the most common and important moves by removing the top moves given by PageRank and CheiRank vectors. An example is shown in Figures 13 and 14 where the remaining moves in the given eigenvectors corresponds to a specific set of moves. Very common moves (such as empty or almost empty plaquettes) have been deleted, leaving more focused groups of moves. For example, the third eigenvector in Figure 13 is much more focused on various moves containing situations of Ko or of imminent capture (Ko or “eternity” is a famous type of fights with alternate captures of opponent’s stones).

A more systematic method that we propose is to consider the ancestors of each move and determine if they share a significant number of preceding moves. As the Google matrix describes a Markovian transition model it would be natural to look for incoming flows of two moves to decide whether they belong to the same community. We implement it as follows: We choose two moves m_1 and m_2 , with, respectively, N_1 and N_2 incoming links. We denote the origin of these incoming links pointing to m_1

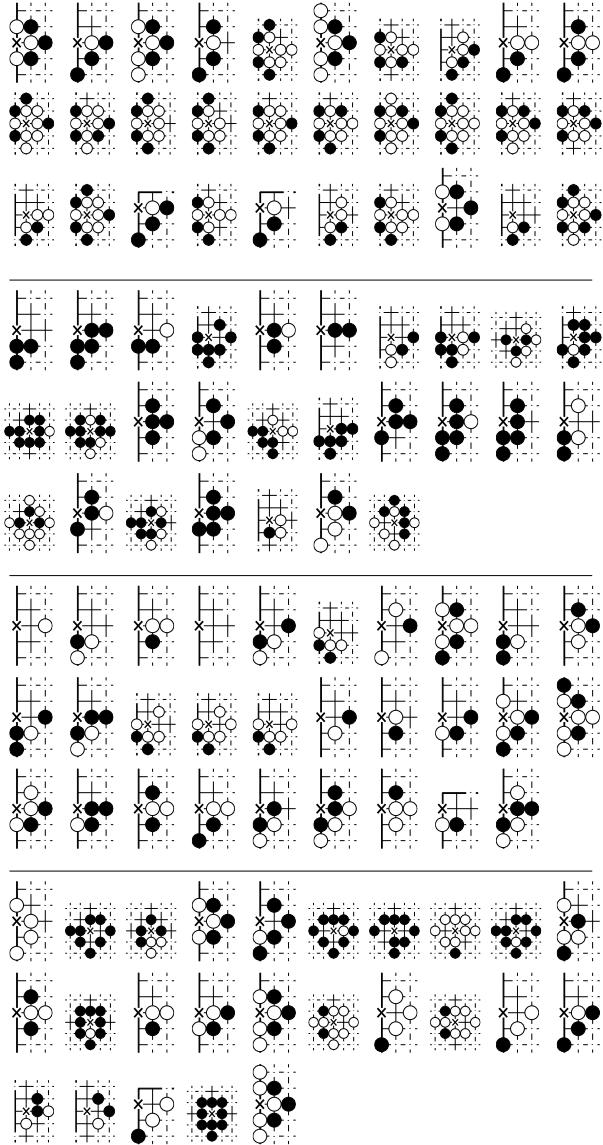


Fig. 14. Same eigenvectors as in Figure 12 treated by filtering out the top 30 CheiRank moves.

and m_2 as sets of moves S_1 and S_2 . If both moves share at least a certain fraction ϵ of common ancestors, that is if $\epsilon \min(N_1, N_2) < \text{card}(S_1 \cap S_2)$, we assign both moves to the same community. This process is iterated until no more new moves are added to this community. This extracting process is of course empirical, but helps us nevertheless to sort out some subgroups of moves that are different from those extracted with previous methods, provided that the parameter ϵ is carefully tuned. Indeed a too low value of ϵ does not help much in extracting a group as in most cases moves share naturally a certain amount of preceding moves but a too high value of ϵ will not capture anything for a sparse matrix. In our network III we thus used the range of values $0.3 < \epsilon < 0.7$. Unfortunately there is no typical behaviour of how the size of a community varies with respect to ϵ : this size depends highly on the initial move and on the number of components

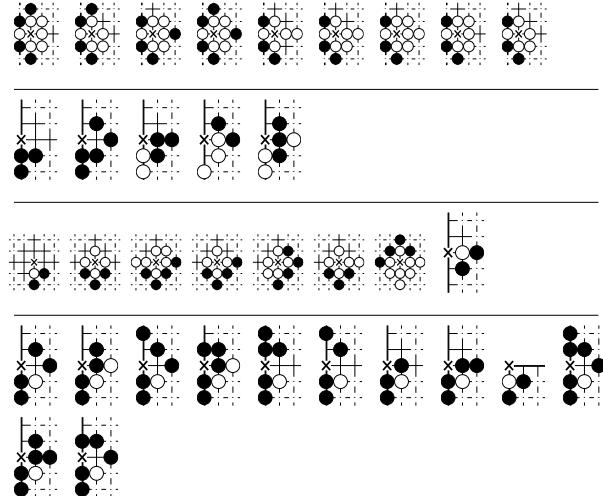


Fig. 15. Example of set of moves extracted from data of Figure 11 by considering common ancestry of moves with threshold level $\epsilon = 0.3$ (see text) applied to λ_7, λ_{11} and λ_{21} , and threshold level $\epsilon = 0.5$ applied to λ_{13} .

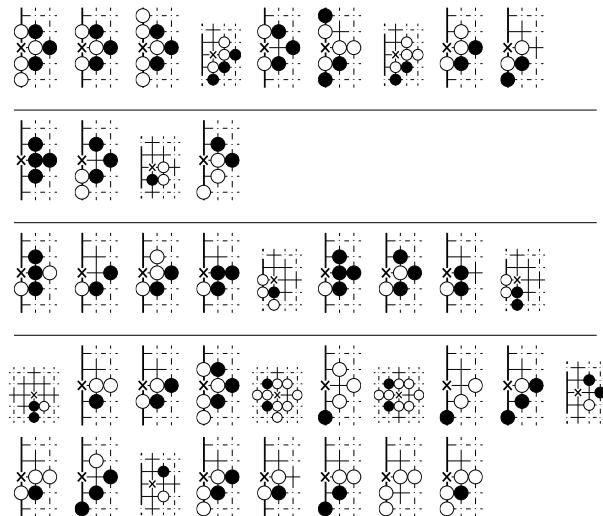


Fig. 16. Example of set of moves extracted from data of Figure 12 by considering common ancestry of moves with threshold level $\epsilon = 0.3$ (see text) applied to $\lambda_7, \lambda_{11}, \lambda_{18}$ and λ_{21} .

of an eigenvector on which one is allowed to explore the ancestries.

We have applied this extracting process on eigenvectors. We thus identify communities in two steps, the first being to select eigenvectors corresponding to the largest eigenvalues of G or G^* , and the second step to follow this ancestry technique. As mentioned earlier an eigenvector corresponding to a large eigenvalue modulus is more likely to be localized on a small number of nodes, therefore one can truncate a given eigenvector to retain its top nodes and apply this method by choosing one of the top nodes as the starting move and constructing the community by successively exploring this subset. Starting from different nodes will allow to identify the different communities. Figures 15 and 16 show that the method is able to extract

moves which have common features, much more so that just looking at largest components of the vectors or removing the ranking vectors (as in Figs. 11–14). Small subsets of moves are disambiguated from the larger groups of the preceding figures, showing sequences which seem to go together with situations of Ko with different black dispositions (first and third eigenvector of Fig. 15), black connecting on the side of the board (fourth eigenvector of Fig. 15), and so on. Similarly, the first line of Figure 16 can be associated to attempts by black to take over an opponent's chain on the rim of the board. These examples show that the method is effective to regroup moves according to reasonably defined affinities.

We mention an alternative method which gives good results in some instances. It consists in analyzing the angles of an eigenvector components when plotted in a complex plane. This method is not systematic as there exist several real valued eigenvectors but for the complex ones one can observe interesting patterns. Either the plots show a meaningless cloud of points or they can reveal a tendency of a subset of components to be aligned. As shown in an example in Figure 17 there can be one or several directions within the same eigenvector, indicating that maybe the phases of the components can characterize moves sharing common properties. Qualitatively speaking the spatial configuration of these subgroups of moves look similar but there are also similarities between moves having different angles, and a formal understanding of the meaning of phases is still lacking. We note that for undirected networks the sign of components of eigenvectors of the adjacency matrix has been used to detect communities [40].

It is worth insisting again on the fact that in general the next to leading eigenvectors in the Google matrix represent a different information from the list of most common moves. In fact, these eigenvectors can even sometimes be highly sensitive to rare links, indeed during our analysis one impossible move was highlighted in one of the top eigenvectors. This move had only two links among the several millions, leading us to find a fake gamefile in the dataset. This shows that the network approach can detect specificities that a mere statistical analysis of the datasets will miss.

It is in principle not excluded that one should look into combinations of eigenvectors but even though we considered single vectors, the results show that it is possible to extract community of moves which share some common properties with these methods. The combination of methods outlined in this section, namely isolating top moves in eigenvectors associated to large eigenvalues, and disambiguating them through search for common ancestries, seems to yield meaningful groups of moves. We stress again that they do not merely correspond to most played moves or sequences of moves, nor to the best ranked in the PageRank or CheiRank, but give a different information related to the network structure around these moves. It is possible to play with the parameters of the method (threshold ϵ , number of eigenvectors, starting point of the common ancestry) in order to find different sets of communities, which should be analyzed in relation with the strat-

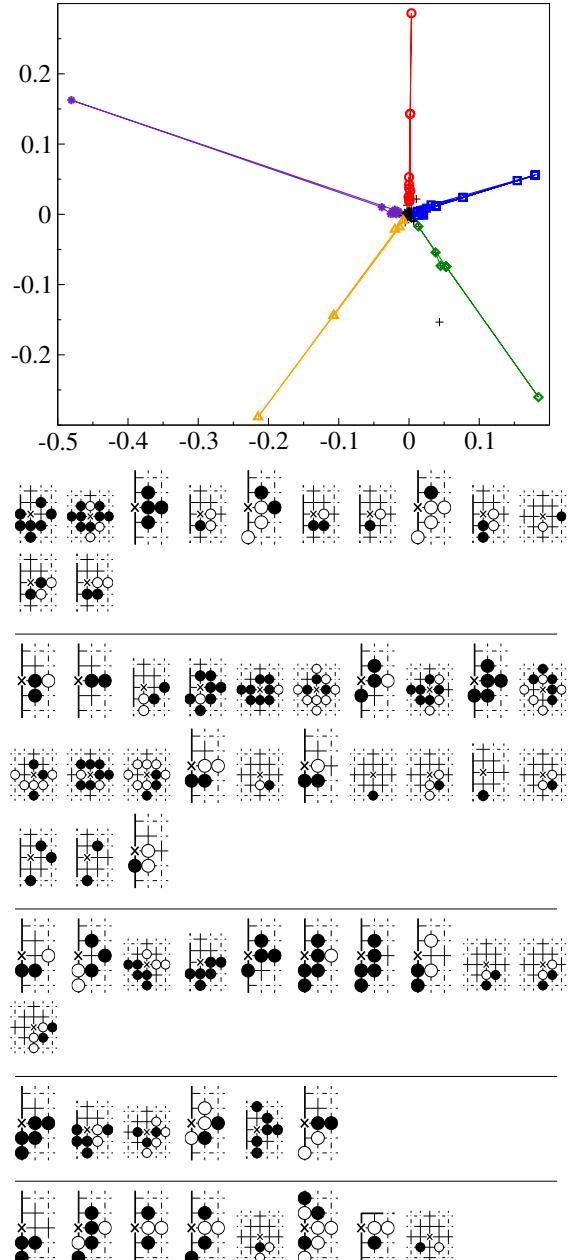


Fig. 17. Example of community extraction through phase analysis (see text) applied on the eigenvector ψ of G^* corresponding to λ_{13} . Top: eigenvector components in the complex plane; groups of plaquettes, from top to bottom, correspond to respective symbols red circles, blue squares, green diamonds, orange triangles and purple stars.

egy of the game, and then could help organize the Monte Carlo go search by running it into specific communities.

5 Generalized networks

One can refine the analysis further by disaggregating the datasets in several ways, constructing different networks from the same database. The number of nodes is still the

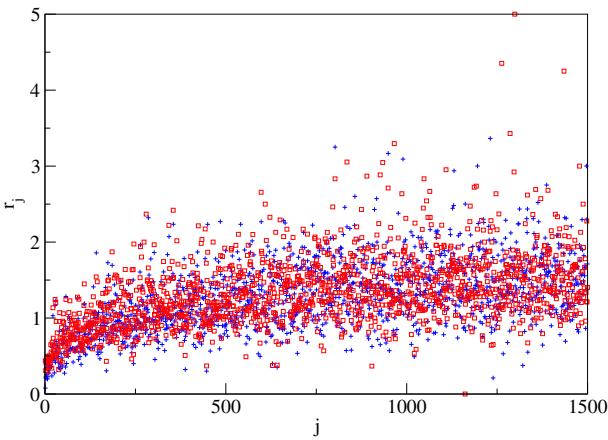


Fig. 18. Fluctuation difference $r_j = \sum_{i \leftarrow j} |k_i - k'_i| / \sum_i k_i$ of outgoing links versus move indices for top 1500 moves of diamond patterns in PageRank order (network III)(see text). An example of difference is shown between two networks built from games between 6d players (blue crosses) and two networks built respectively from games between 1d players and games between 9d players (red squares). The number of games in each case is 2731, corresponding to the number of 1d/1d games in the database [21].

same, but links are now selected according to some specific criterion and may give rise to different properties. In this section we will illustrate this by a few examples.

An important aspect of the games, especially in view of applications to computer go, is to select moves which are more susceptible of winning the game. It is possible to separate the players between winners and losers, but the presence of handicaps makes this process ambiguous. Indeed, it is possible to place up to nine stones before the beginning of the game at strategic locations, giving an advantage to a weaker player which may allow him to play against a better opponent with a fair chance of winning. Another possibility we thus investigated was to separate the players by their levels according to their dan ranking. Indeed, players are ranked from first dan (1d, lowest level) to ninth dan (9d, highest level). In the database [21] the number of dans of the players is known, and it is therefore possible to separate games played at different levels. To explore these differences, we constructed the diamond network from games played by 1d versus 1d, the one from 9d versus 9d, and the one from 6d versus 6d. Figure 18 shows the quantity $r_j = \sum_{i \leftarrow j} |k_i - k'_i| / \sum_i k_i$ defined for a pair of networks, where k_i (resp. k'_i) is the number of links from a fixed node j to node i for one network (resp. for the second network). For each node, r_j thus quantifies the difference in outgoing links between two networks. Figure 18 shows the distribution of this quantity highlighting the difference between the network 1d/1d and the network 9d/9d. One sees that they are indeed different, with a mean $\langle r_j \rangle \approx 1.33$. Nevertheless, in the same figure we add for comparison the difference between two networks of 6d/6d, showing that one can also find differences between networks built from players of the same level. In view of this, to see if the difference between 1d/1d and

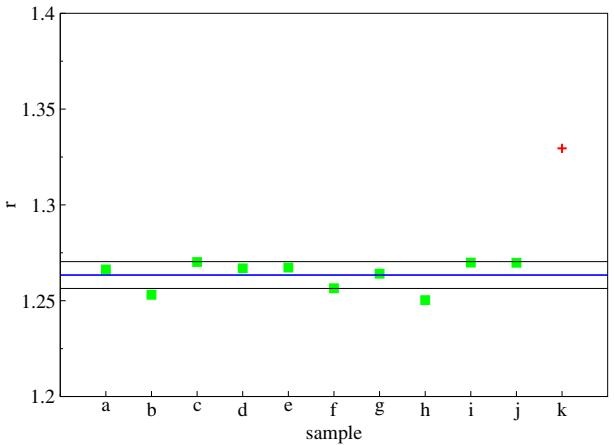


Fig. 19. Difference r (see text) between the networks built from games of 1d players and of 9d players (red cross) together with several examples of r for pairs of networks constructed from different samples of games of 6d players (green squares). The three horizontal lines mark the mean and the variance of the 6d values The number of games in each sample is 2731, corresponding to the number of 1d/1d games in the database.

9d/9d is statistically significant, Figure 19 shows the average $r = \langle r_j \rangle$ for different choices of samples of 6d versus 6d games and the value for the networks constructed from the games of 1d players and 9d players, with the average taken on top 1500 moves of the PageRank. It shows that the difference between 1d players and 9d players has some statistical significance. The quantity r is a simple way of quantifying the structural differences in the networks at the level of outgoing flows which is in our case an indication that 9d players might have an overall structurally different style of play than 1d players, even though the difference is relatively small.

An interesting possibility which might also be useful for applications is to create separate networks for different phases of the game. For instance, one can take into account when using the database of real games only the first 50 moves, the middle 50, or the final 50. Again, this does not modify the nodes of the networks, but changes the links, creating three different networks corresponding to respectively beginning, middle, and ending phases of the game. The number of links is now 6 155 936 for the beginning phase, 6 460 771 for the middle phase, and 5 947 467 for the ending phase (instead of 26 116 006 for the whole game) (the numbers without degeneracies for diamond plaquettes are respectively 613 953, 2 070 305 and 3 182 771). The spectra of the three networks for the diamond plaquettes are shown in Figure 20 (again, only the largest eigenvalues are calculated). It is clear that the spectra are quite different, indicating that the structure of the network is not equivalent for the different phases of the game. It is visible that the eigenvalue cloud is larger for the ending phase indicating that near the final stage of the game the random surfer gets trapped more easily in specific patterns, which should correspond to typical endgames. Similarly, the gap is smaller for the beginning phase, indicating that one

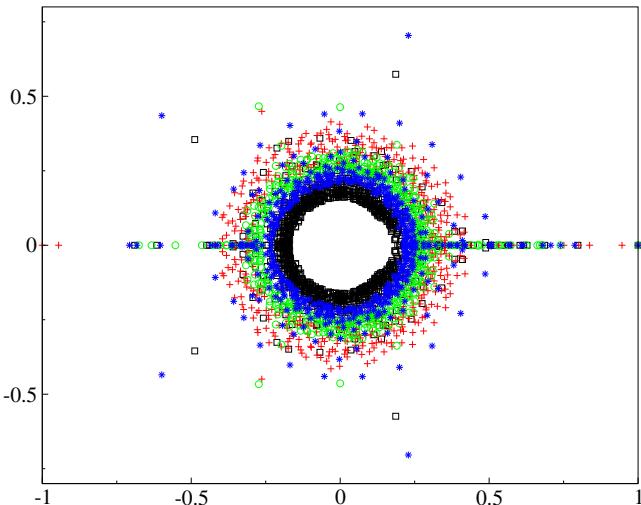


Fig. 20. Spectrum of G for diamond networks of different game phases: first 50 moves (red crosses), middle 50 moves (green circles) and last 50 moves (blue stars). The black squares correspond to the spectrum of the network when the whole game is taken into account, shown for reference.

strongly knit community exists with an eigenvalue close to the PageRank value.

The eigenvectors shown in Figure 21 highlight different sets of moves as might be expected since strategy should differ in those phases. Obviously, eigenvectors for opening moves are much more biased towards relatively empty plaquettes, indicating the start of local fights. In the middle and end of the games, communities are biased towards moves corresponding to more and more filled plaquettes, indicating ongoing fights or fight endings. We stress the fact that those sets of moves are not just the most played moves in the respective phases. Running the community detection process of Section 4 on such eigenvectors should select communities specific to these different phases of the game.

6 Conclusion

We have shown that it is possible to construct networks which describe the game of go, in a spirit similar to the ones already used for languages. We have extended the results of [4], comparing three networks of different sizes according to the size of the plaquettes which serve as nodes of the network. The three networks share structural similarities, such as a statistical correlation (but not an exact symmetry) between incoming and outgoing links. However, the largest network, besides necessitating more refined numerical tools in order to obtain the largest eigenvalues and associated eigenvectors, is also much less connected and disambiguates much better the different moves. We have also shown that specific subnetworks can be constructed, selecting links in the databases according to levels of the players or phases of the game.

Our results show that the networks constructed in this way have specific properties which reflect the peculiarities

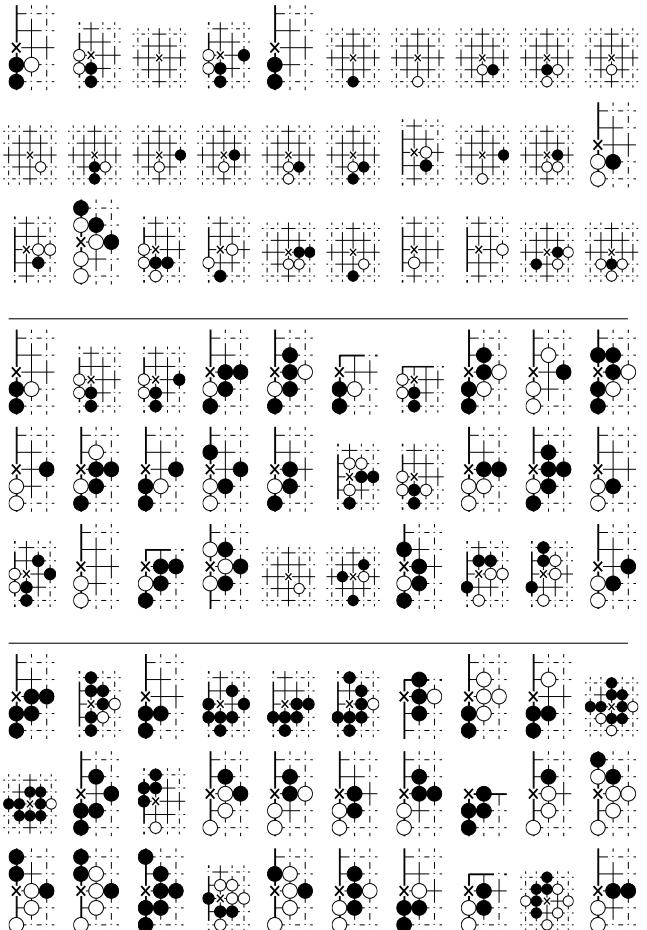


Fig. 21. Examples of set of top 30 moves where eigenvectors of G localize themselves, those examples are computed for diamond network in different game phases: starting phase and λ_4 (top), middle phase and λ_4 (middle) and ending phase λ_4 (bottom).

of the game. In particular, the PageRank and CheiRank vectors give new orderings of the moves, which do not merely correspond to most played moves or sequences of moves, but give a different information. As explained in Section 3, moves highlighted by the ranking vectors can correspond to moves which are connected to chains of important moves, even though they are not that frequent (it was this difference which made Google the famous company it is today). We have also shown that it is possible with these methods to extract communities of moves which share some common properties. A possible use of these results would be to help organize the Monte Carlo go search by running it into specific communities. Indeed, despite its limitations [41], Monte-Carlo go remains the most promising approach to computer go. The main goal of these algorithms is an efficient value function estimation [12]. We have proposed in this paper various community detection processes, and the knowledge of these communities could be used for instance to initialize the value of moves according to the local pattern, at a value given by the value of its ancestors. It could also be used to propagate the value of a move to similar moves. It would be

interesting to compare the values assigned to nodes of our networks by the different computer programs available, in order to see whether adjacency matrix properties could be used to converge more quickly to the correct value function. We think an especially interesting path in this direction corresponds to the approach outlined in Section 5: by constructing specific networks according to game phases or levels of players, one can specify communities useful in specific contexts of the game or corresponding to winning strategies. It is also possible to use “personalization” techniques (implemented by modifying the vector e in the definition of G in Sect. 3 [29]) which are currently explored in a World Wide Web context and allow to compute a ranking vector biased towards a certain group of nodes, e.g. one of the communities discussed in Section 4. All these techniques deserve further study in this context.

It will be fascinating to see if other games such as chess could be modelized this way, and how different the results will be. Besides its applicability to the simulations of go on computers, we also believe that such studies enable to get insight on the way the human brain participates in such game activities, and more generally on the human decision-making processes [9]. In this direction, an interesting extension of this work could be to compare the networks built from games played by human beings and computers, and determine how different they are.

We thank Dima Shepelyansky, Klaus Frahm, Pierre Aubourg, Yoann Séon and François Damon for discussions and insights. We thank CalMiP for access to its supercomputers. V.K. thanks the CNRS and the Région Midi-Pyrénées for funding. This research is supported in part by the EC FET Open project “New tools and algorithms for directed network analysis” (NA-DINE No. 288956).

References

1. R. Albert, A.-L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002)
2. S.N. Dorogovtsev, J.F.F. Mendes, *Evolution of Networks* (Oxford University Press, Oxford, 2003)
3. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Phys. Rep. **424**, 175 (2006)
4. B. Georgeot, O. Giraud, Europhys. Lett. **97**, 68002 (2012)
5. J. Tromp, G. Farnebäck, in *Proceedings of the 5th International Conference on Computer and Games*, edited by H.J. van den Herik, P. Ciancarini, H.H.L.M. Donkers, Lect. Notes in Comp. Sciences (Springer-Verlag, Heidelberg, 2007), Vol. 4630, p. 84
6. H.J. van den Herik, J.W.H.M. Uiterwijk, J. van Rijswijck, Artif. Intell. **134**, 277 (2002)
7. N.N. Schraudolph, P. Dayan, T.J. Sejnowski, Adv. Neural Inf. Process. **6**, 817 (1994)
8. G. Chaslot, J.T. Saito, B. Bouzy, J.W.H.M. Uiterwijk, H.J. van den Herik, in *Proceedings of the 18th BeNeLux Conf. on Artificial Intelligence*, edited by P.Y. Schobbens, W. Vanhoof, G. Schwanen (2006), Vol. 83
9. C.B. Browne, E. Powley, D. Whitehouse, S.M. Lucas, P.I. Cowling, P. Rohlfschagen, S. Tavener, D. Perez, S. Samothrakis, S. Colton, IEEE Trans. Comput. Intell. AI Games **4**, 1 (2012)
10. R. Coulom, in *Proceedings of the 5th International Conference on Computer and Games*, edited by H.J. van den Herik, P. Ciancarini, H.H.L.M. Donkers, Lect. Notes in Comp. Sciences (Springer-Verlag, Heidelberg, 2007), Vol. 4630, p. 72
11. Y. Wang, S. Gelly, in *IEEE Symposium on Computational Intelligence and Games, CIG 2007* (2007), Vol. 175
12. S. Gelly, L. Kocsis, M. Schoenauer, M. Sebag, D. Silver, C. Szepesvri, O. Teytaud, Commun. ACM **55**, 106 (2012)
13. S. Gelly, D. Silver, Artif. Intell. **175**, 1856 (2011)
14. S. Gelly, D. Silver, in *Proceedings of the 24th International Conference on Machine Learning, New York, USA, 2007*, p. 273
15. B. Bouzy, G. Chaslot, in *Proceedings IEEE Symp. Comput. Intell. Games, Colchester, UK, 2005*, p. 176
16. R. Coulom, in *Proceedings Comput. Games Workshop, Amsterdam, The Netherlands, 2007*, p. 113
17. R. Ferrer-i-Cancho, R.V. Sole, Proc. R. Soc. London Ser. B **268**, 2261 (2001)
18. S.N. Dorogovtsev, J.F.F. Mendes, Proc. R. Soc. London Ser. B **268**, 2603 (2001)
19. A.P. Masucci, G.J. Rodgers, Phys. Rev. E **74**, 026102 (2006)
20. S.-C. Huang, R. Coulom, S.-S. Lin, Lect. Notes Comp. Sci. **6515**, 81 (2011)
21. <http://www.u-go.net/>
22. G.K. Zipf, *The Psycho-Biology of Language: an Introduction to Dynamic Philology* (Houghton Mifflin, Boston, 1935)
23. X. Gabaix, Quart. J. Econ. **114**, 739 (1999)
24. K. Okuyama, M. Takayasu, H. Takayasu, Physica A **269**, 125 (1999)
25. B. Blasius, R. Tönjes, Phys. Rev. Lett. **103**, 218701 (2009)
26. D. Donato, L. Laura, S. Leonardi, S. Millozzi, Eur. Phys. J. B **38**, 239 (2004)
27. G. Pandurangan, P. Raghavan, E. Upfal, Internet Math. **3**, 1 (2005)
28. S. Brin, L. Page, Comput. Networks ISDN Syst. **33**, 107 (1998)
29. A.M. Langville, C.D. Meyer, *Google’s PageRank and Beyond: the Science of Search Engine Rankings* (Princeton University Press, Princeton, 2006)
30. A.D. Chepelianskii, [arXiv:1003.5455](https://arxiv.org/abs/1003.5455) (2010)
31. A.O. Zhirov, O.V. Zhirov, D.L. Shepelyansky, Eur. Phys. J. B **77**, 523 (2010)
32. J. Kleinberg, JACM **46**, 604 (1999)
33. R. Lambiotte, M. Rosvall, Phys. Rev. E **85**, 056107 (2012)
34. L. Ermann, D.L. Shepelyansky, Acta Physica Polonica A **120**, 158 (2011)
35. O. Giraud, B. Georgeot, D.L. Shepelyansky, Phys. Rev. E **80**, 026107 (2009)
36. B. Georgeot, O. Giraud, D.L. Shepelyansky, Phys. Rev. E **81**, 056109 (2010)
37. J.-C. Delvenne, S.N. Yaliraki, M. Barahona, Proc. Natl. Acad. Sci. **107**, 12755 (2010)
38. S. Fortunato, Phys. Rep. **486**, 75 (2010)
39. L. Ermann, K.M. Frahm, D.L. Shepelyansky, Eur. Phys. J. B **86**, 193 (2013)
40. F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, P. Zhang, Proc. Natl. Acad. Sci. **110**, 20935 (2013)
41. S.-C. Huang, M. Müller, in *Proceedings of the 8th International Conference on Computer and Games* (Springer, 2013)

Google matrix analysis of the multiproduct world trade network

Leonardo Ermann¹ and Dima L. Shepelyansky^{2,a}

¹ Departamento de Física Teórica, GIyA, CNEA, Av. Libertador 8250, C1429BNP Buenos Aires, Argentina

² Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France

Received 14 January 2015 / Received in final form 2 March 2015

Published online 1st April 2015 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2015

Abstract. Using the United Nations COMTRADE database [United Nations Commodity Trade Statistics Database, available at: <http://comtrade.un.org/db/>. Accessed November (2014)] we construct the Google matrix G of multiproduct world trade between the UN countries and analyze the properties of trade flows on this network for years 1962–2010. This construction, based on Markov chains, treats all countries on equal democratic grounds independently of their richness and at the same time it considers the contributions of trade products proportionally to their trade volume. We consider the trade with 61 products for up to 227 countries. The obtained results show that the trade contribution of products is asymmetric: some of them are export oriented while others are import oriented even if the ranking by their trade volume is symmetric in respect to export and import after averaging over all world countries. The construction of the Google matrix allows to investigate the sensitivity of trade balance in respect to price variations of products, e.g. petroleum and gas, taking into account the world connectivity of trade links. The trade balance based on PageRank and CheiRank probabilities highlights the leading role of China and other BRICS countries in the world trade in recent years. We also show that the eigenstates of G with large eigenvalues select specific trade communities.

1 Introduction

According to the data of UN COMTRADE [1] and the international trade statistics 2014 of the World Trade Organization (WTO) [2] the international world trade between world countries demonstrates a spectacular growth with an increasing trade volume and number of trade products. It is well clear that the world trade plays the fundamental role in the development of world economy [3]. According to the WTO Chief Statistician Hubert Escaith “In recent years we have seen growing demand for data on the world economy and on international trade in particular. This demand has grown in particular since the 2008–2009 crisis, whose depth and breadth surprised many experts” [2]. In global the data of the world trade exchange can be viewed as a large multi-functional directed World Trade Network (WTN) which provides important information about multiproduct commercial flows between countries for a given year. At present the COMTRADE database contains data for $N_c = 227$ UN countries with up to $N_p \approx 10^4$ trade products. Thus the whole matrix of these directed trade flows has a rather large size $N = N_p N_c \sim 10^6$. A usual approach is to consider the export and import volumes, expressed in US dollars (USD). An example of the world map of countries characterized by their import and export trade volume for year 2008 is shown in Figure 1. However,

such an approach gives only an approximate description of trade where hidden links and interactions between certain countries and products are not taken into account since only a country global import or export are considered. Thus the statistical analysis of these multiproduct trade data requires a utilization of more advanced mathematical and numerical methods.

In fact, in the last decade, modern societies developed enormous communication and social networks including the World Wide Web (WWW), Wikipedia, Twitter, etc. (see e.g. [4]). A necessity of information retrieval from such networks led to a development of efficient algorithms for information analysis on such networks appeared in computer science. One of the most spectacular tools is the PageRank algorithm developed by Brin and Page in 1998 [5], which became a mathematical foundation of the Google search engine (see e.g. [6]). This algorithm is based on the concept of Markov chains and a construction of the Google matrix G of Markov transitions between network nodes. The right eigenvector of this matrix G , known as PageRank vector, allows to rank all nodes according to their importance and influence on the network. The studies of various directed networks showed that it is useful to analyze also the matrix G^* constructed for the same network but with an inverted direction of links [7,8]. The PageRank vector of G^* is known as the CheiRank vector. The spectral properties of Google matrix for various networks are described in reference [9].

^a e-mail: dima@irsamc.ups-tlse.fr

Table 1. Codes and names of the 61 products from COMTRADE Standard International Trade Classification (SITC) Rev. 1.

Code	Name	Code	Name
00	Live animals	54	Medicinal and pharmaceutical products
01	Meat and meat preparations	55	Perfume materials, toilet & cleansing preparations
02	Dairy products and eggs	56	Fertilizers, manufactured
03	Fish and fish preparations	57	Explosives and pyrotechnic products
04	Cereals and cereal preparations	58	Plastic materials, etc.
05	Fruit and vegetables	59	Chemical materials and products, nes
06	Sugar, sugar preparations and honey	61	Leather, lthr. Manufs., nes & dressed fur skins
07	Coffee, tea, cocoa, spices & manufacs. Thereof	62	Rubber manufactures, nes
08	Feed. Stuff for animals excl. Unmilled cereals	63	Wood and cork manufactures excluding furniture
09	Miscellaneous food preparations	64	Paper, paperboard and manufactures thereof
11	Beverages	65	Textile yarn, fabrics, made up articles, etc.
12	Tobacco and tobacco manufactures	66	Non metallic mineral manufactures, nes
21	Hides, skins and fur skins, undressed	67	Iron and steel
22	Oil seeds, oil nuts and oil kernels	68	Non ferrous metals
23	Crude rubber including synthetic and reclaimed	69	Manufactures of metal, nes
24	Wood, lumber and cork	71	Machinery, other than electric
25	Pulp and paper	72	Electrical machinery, apparatus and appliances
26	Textile fibres, not manufactured, and waste	73	Transport equipment
27	Crude fertilizers and crude minerals, nes	81	Sanitary, plumbing, heating and lighting fixt.
28	Metalliferous ores and metal scrap	82	Furniture
29	Crude animal and vegetable materials, nes	83	Travel goods, handbags and similar articles
32	Coal, coke and briquettes	84	Clothing
33	Petroleum and petroleum products	85	Footwear
34	Gas, natural and manufactured	86	Scientif & control instrum, photogr gds, clocks
35	Electric energy	89	Miscellaneous manufactured articles, nes
41	Animal oils and fats	91	Postal packages not class. According to kind
42	Fixed vegetable oils and fats	93	Special transact. Not class. According to kind
43	Animal and vegetable oils and fats, processed	94	Animals, nes, incl. Zoo animals, dogs and cats
51	Chemical elements and compounds	95	Firearms of war and ammunition therefor
52	Crude chemicals from coal, petroleum and gas	96	Coin, other than gold coin, not legal tender
53	Dyeing, tanning and colouring materials		

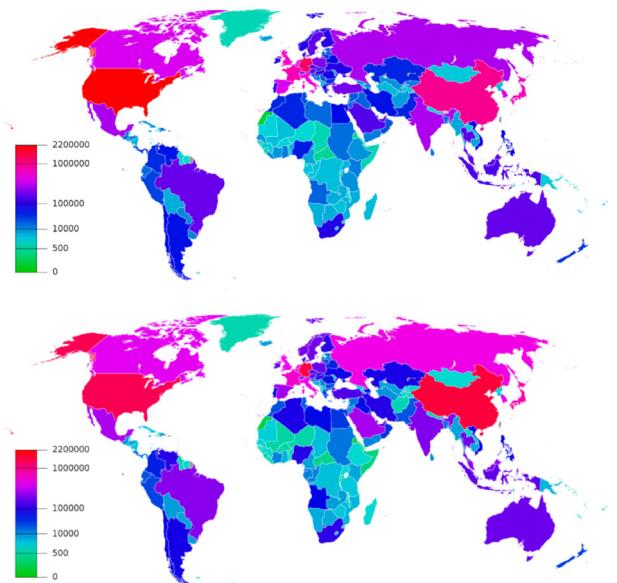


Fig. 1. World map of countries with color showing country import (top panel) and export (bottom panel) trade volume expressed in millions of USD given by numbers of the color bars. The data are shown for year 2008 with $N_c = 227$ countries for trade in all $N_p = 61$ products (from UN COMTRADE [1]). Names of countries can be found at [10].

The approach of Google matrix to the analysis of WTN was started in [11]. The striking feature of this approach is that it treats all UN countries on equal democratic grounds, independently of richness of a given country, in agreement with the principles of UN where all countries are equal. This property of G matrix is based on the property of Markov chains where the total probability is conserved to be unity since the sum of elements for each column of G is equal to unity. Even if in this approach all countries are treated on equal grounds still the PageRank and CheiRank analysis recovers about 75% of industrially developed countries of G20. However, now these countries appear at the top ranking positions not due to their richness but due to the efficiency of their trade network. Another important aspect found in [11] is that both PageRank and CheiRank vectors appear very naturally in the WTN corresponding to import and export flows.

In this work we extend the Google matrix analysis for the multiproduct WTN obtained from COMTRADE [1] with up to $N_p = 61$ trade products for up to $N_c = 227$ countries. The global G matrix of such trade flows has a size up to $N = N_p N_c = 13847$ nodes. The names and codes of products are given in Table 1 and their trade volumes, expressed in percent of the whole world trade volume, are given in Table 2 for years 1998 and 2008.

Table 2. Columns represent data: codes of 61 products of COMTRADE SITC Rev. 1, ImportRank and ExportRank $\hat{K} = \hat{K}^*$ in year 2008, product fraction in global trade volume in 2008, $\hat{K} = \hat{K}^*$ in 1998, product fraction in 1998.

Code	$\hat{K}(08)$	% Vol(08)	$\hat{K}(98)$	% Vol (98)	Code	$\hat{K}(08)$	% Vol(08)	$\hat{K}(98)$	% Vol (98)
00	53	0.10	51	0.17	54	9	2.89	16	1.88
01	27	0.69	26	0.83	55	25	0.76	28	0.79
02	34	0.44	34	0.56	56	30	0.55	43	0.36
03	28	0.63	22	0.99	57	58	0.03	57	0.04
04	21	1.07	19	1.13	58	15	1.95	13	2.07
05	19	1.16	18	1.50	59	22	1.04	20	1.13
06	49	0.23	44	0.36	61	51	0.19	42	0.37
07	33	0.47	29	0.73	62	26	0.73	24	0.85
08	38	0.39	36	0.45	63	35	0.43	32	0.61
09	40	0.34	41	0.39	64	20	1.14	17	1.79
11	31	0.54	31	0.65	65	18	1.40	11	2.46
12	47	0.24	35	0.49	66	17	1.71	14	2.01
21	56	0.05	53	0.11	67	7	3.63	10	2.74
22	37	0.39	48	0.32	68	11	2.27	15	1.95
23	44	0.26	50	0.22	69	13	2.04	12	2.12
24	39	0.35	30	0.65	71	2	11.82	1	15.03
25	43	0.29	45	0.34	72	3	10.42	3	12.26
26	50	0.22	37	0.45	73	4	10.06	2	12.38
27	41	0.33	47	0.33	81	42	0.31	46	0.34
28	16	1.92	25	0.84	82	23	0.93	21	1.03
29	48	0.24	40	0.39	83	45	0.26	49	0.26
32	24	0.82	39	0.42	84	10	2.42	6	3.44
33	1	14.88	4	5.02	85	29	0.59	27	0.79
34	14	2.04	23	0.99	86	12	2.25	8	2.95
35	46	0.26	52	0.17	89	6	3.72	5	4.54
41	57	0.03	58	0.04	91	61	0.00	61	0.00
42	32	0.49	38	0.44	93	5	3.92	9	2.92
43	54	0.08	55	0.08	94	59	0.01	59	0.01
51	8	3.01	7	3.07	95	55	0.08	54	0.11
52	52	0.11	56	0.05	96	60	0.00	60	0.00
53	36	0.39	33	0.60					

The main problem of construction of such a matrix is not its size, which is rather modest compared to those studied in [9], but the necessity to treat all countries on democratic grounds and at the same time to treat trade products on the basis of their trade volume. Indeed, the products cannot be considered on democratic grounds since their contributions to economy are linked with their trade volume. Thus, according to Table 2, in year 2008 the trade volume of *Petroleum and petroleum products* (code 33 in Tab. 1) is by a factor 300 larger than those of *Hides, skins and fur skins (undressed)* (code 21 in Tab. 1). To incorporate these features in our mathematical analysis of multiproduct WTN we developed in this work the google personalized vector method (GPVM) which allows to keep a democratic treatment of countries and at the same time to consider products proportionally to their trade volume. As a result we are able to perform analysis of the global multiproduct WTN keeping all interactions between all countries and all products. This is a new step in the WTN analysis since in our previous studies [11] it was possible to consider a trade between countries only in one product or only in all products summed together (all commodities). The new finding of such global WTN analysis is an

asymmetric ranking of products: some of them are more oriented to import and others are oriented to export while the ranking of products by the trade volume is always symmetric after summation over all countries. This result with asymmetric ranking of products confirms the indications obtained on the basis of ecological ranking [12], which also gives an asymmetry of products in respect to import and export. Our approach also allows to analyze the sensitivity of trade network to price variations of a certain product.

We think that the GPVM approach allows to perform a most advanced analysis of multiproduct world trade. The previous studies have been restricted to studies of statistical characteristics of WTN links, patterns and their topology (see e.g. [13–19]). The applications of PageRank algorithm to the WTN was discussed in [20] but effects of export had been not analyzed there, the approach based on HITS algorithm was used in [21]. In comparison to the above studies, the approach developed here for the multiproduct WTN has an advantage of analysis of ingoing and outgoing flows, related to PageRank and CheiRank, and of taking into account of multiproduct aspects of the WTN. Even if the importance to multiproduct WTN analysis is clearly understood by researchers

(see e.g. [22]) the Google matrix methods have not been efficiently used up to now. We also note that the matrix methods are extensively used for analysis of correlations of trade indexes (see e.g. [23,24]) but these matrices are Hermitian being qualitatively different from those appearing in the frame of Markov chains. Here we make the steps in multi-functional or multiproduct Google matrix analysis of the WTN extending the approach used in [11].

2 Methods

Below we give the mathematical definition for the construction of the Google matrix G , which belongs to the class of Perron-Frobenius operators and Markov chains. The matrix G is constructed for the import (ingoing) trade flows. We also use the matrix G^* built from the export (outgoing) trade flows. The matrix size N is given by the product of number of countries N_c by the number of products N_p . The main features of matrices G and G^* are: all elements are real positive numbers or zeros; the sum of elements in each matrix column is equal to unity, that gives the probability conservation required for Markov chains. We use the right eigenvectors of PageRank P and CheiRank P^* respectively for the matrices G , G^* with the largest eigenvalue $\lambda = 1$ ($GP = P$, $G^*P^* = P^*$). These vectors give the stationary distribution of probability over the nodes. The important element of G and G^* is their democratic (equal grounds) treatment of all world countries independently of their richness. This results from the construction rules of G , G^* where for each country the sum of elements in each column, corresponding to any product of given country, is equal to unity. At the same time we keep the contribution of products to be proportional to their trade volume since their effect on the trade is indeed related with their volume contribution in the world trade. Thus, the important new element of this work is the new proposed method which uses a certain personalized vector in construction of G , G^* and satisfies the above requirements.

At the same time we note that it is preferable to work in a certain fixed class of operators, e.g. Google matrix and Markov chains. Already only this requirement implies that we need to treat countries in a democratic manner since by the construction sum of elements in each column should be unity. For one product, or for a sum of all products, the construction of G , G^* is relatively straightforward as described in [11]. The most tricky part is the case of many products which contribution should be treated proportionally to their fraction in the world trade. We describe the construction method of G , G^* which takes into account both these features of the world trade. We note that, as discussed in [11], the obtained results have no significant dependence on the damping factor α , which we keep below at the fixed value $\alpha = 0.5$. The simple examples of constructions of matrices G , G^* for directed networks are illustrated in Figures 3 and 4 in [9]. Below we present all mathematical definitions and describe the main features of these matrices and eigenvectors.

2.1 Google matrix construction for the WTN

For a given year, we build N_p money matrices $M_{c,c'}^p$ of the WTN from the COMTRADE database [1] (see [11]).

$$M_{c,c'}^p = \text{product } p \text{ transfer (in USD) from country } c' \text{ to } c. \quad (1)$$

Here the country indexes are $c, c' = 1, \dots, N_c$ and a product index is $p = 1, \dots, N_p$. According to the COMTRADE database the number of UN registered countries is $N_c = 227$ (in recent years) and the number of products is $N_p = 10$ and $N_p = 61$ for 1 and 2 digits respectively from the Standard International Trade Classification (SITC) Rev. 1. For convenience of future notation we also define the volume of imports and exports for a given country and product respectively as:

$$V_c^p = \sum_{c'} M_{c,c'}^p, V_c^{*p} = \sum_{c'} M_{c',c}^p. \quad (2)$$

The import and export volumes $V_c = \sum_p V_c^p$ and $V_c^* = \sum_p V_c^{*p}$ are shown for the world map of countries in Figure 1 for year 2008.

In order to compare later with PageRank and CheiRank probabilities we define volume trade ranks in the whole trade space of dimension $N = N_p \times N_c$. Thus the ImportRank (\hat{P}) and ExportRank (\hat{P}^*) probabilities are given by the normalized import and export volumes

$$\hat{P}_i = V_c^p / V, \hat{P}_i^* = V_c^{*p} / V, \quad (3)$$

where $i = p + (c - 1)N_p$, $i = 1, \dots, N$ and the total trade volume is:

$$V = \sum_{p,c,c'} M_{c,c'}^p = \sum_{p,c} V_c^p = \sum_{p,c} V_c^{*p}.$$

The Google matrices G and G^* are defined as $N \times N$ real matrices with non-negative elements:

$$G_{ij} = \alpha S_{ij} + (1 - \alpha) v_i e_j, G^*_{ij} = \alpha S^*_{ij} + (1 - \alpha) v_i^* e_j, \quad (4)$$

where $N = N_p \times N_c$, $\alpha \in (0, 1]$ is the damping factor ($0 < \alpha < 1$), e_j is the row vector of unit elements ($e_j = 1$), and v_i is a positive column vector called a *personalization vector* with $\sum_i v_i = 1$ [6]. We note that the usual Google matrix is recovered for a personalization vector $v_i = e_i / N$. In this work, following [11], we fix $\alpha = 0.5$. As discussed in [6,9,11] a variation of α in a range (0.5, 0.9) does not significantly affect the probability distributions of PageRank and CheiRank vectors. We specify the choice of the personalization vector a bit below.

The matrices S and S^* are built from money matrices $M_{cc'}^p$ as:

$$S_{i,i'} = \begin{cases} M_{c,c'}^p \delta_{p,p'} / V_{c'}^{*p} & \text{if } V_{c'}^{*p} \neq 0 \\ 1/N & \text{if } V_{c'}^{*p} = 0 \end{cases} \quad (5)$$

$$S^*_{i,i'} = \begin{cases} M_{c',c}^p \delta_{p,p'} / V_{c'}^p & \text{if } V_{c'}^p \neq 0 \\ 1/N & \text{if } V_{c'}^p = 0 \end{cases} \quad (5)$$

where $c, c' = 1, \dots, N_c$; $p, p' = 1, \dots, N_p$; $i = p + (c-1)N_p$; $i' = p' + (c'-1)N_p$; and therefore $i, i' = 1, \dots, N$. Note that the sum of each column of S and S^* are normalized to unity and hence the matrices G, G^*, S, S^* belong to the class of Google matrices and Markov chains. The eigenvalues and eigenstates of G, G^* are obtained by a direct numerical diagonalization using the standard numerical packages.

2.2 PageRank and CheiRank vectors from GPVM

PageRank and CheiRank (P and P^*) are defined as the right eigenvectors of G and G^* matrices respectively at eigenvalue $\lambda = 1$:

$$\sum_j G_{ij} \psi_j = \lambda \psi_i, \sum_j G^*_{ij} \psi^*_j = \lambda \psi^*_i. \quad (6)$$

For the eigenstate at $\lambda = 1$ we use the notation $P_i = \psi_i, P^*_i = \psi^*_i$ with the normalization $\sum_i P_i = \sum_i P^*_i = 1$. For other eigenstates we use the normalization $\sum_i |\psi_i|^2 = \sum_i |\psi^*_i|^2 = 1$. According to the Perron-Frobenius theorem the components of P_i, P^*_i are positive and give the probabilities to find a random surfer on a given node [6]. The PageRank K and CheiRank K^* indexes are defined from the decreasing ordering of P and P^* as $P(K) \geq P(K+1)$ and $P^*(K) \geq P^*(K^*+1)$ with $K, K^* = 1, \dots, N$.

If we want to compute the reduced PageRank and CheiRank probabilities of countries for *all commodities* (or equivalently all products) we trace over the product space getting $P_c = \sum_p P_{pc} = \sum_p P(p + (c-1)N_p)$ and $P^*_c = \sum_p P^*_{pc} = \sum_p P^*(p + (c-1)N_p)$ with their corresponding K_c and K^*_c indexes. In a similar way we obtain the reduced PageRank and CheiRank probabilities for products tracing over all countries and getting $P_p = \sum_c P(p + (c-1)N_p) \sum_p P_{pc}$ and $P^*_p = \sum_c P^*(p + (c-1)N_p) \sum_p P^*_{pc}$ with their corresponding product indexes K_p and K^*_p .

In summary we have $K_p, K^*_p = 1, \dots, N_p$ and $K_c, K^*_c = 1, \dots, N_c$. A similar definition of ranks from import and export trade volume can be done in a straightforward way via probabilities $\hat{P}_p, \hat{P}^*_p, \hat{P}_c, \hat{P}^*_c, \hat{P}_{pc}, \hat{P}^*_{pc}$ and corresponding indexes $\hat{K}_p, \hat{K}^*_p, \hat{K}_c, \hat{K}^*_c, \hat{K}$.

To compute the PageRank and CheiRank probabilities from G and G^* keeping democracy in countries and proportionality of products to their trade volume we use the GPVM approach with a personalized vector in (4). At the first iteration of Google matrix we take into account the relative product volume per country using the following personalization vectors for G and G^* :

$$v_i = \frac{V_c^p}{N_c \sum_{p'} V_c^{p'}}, v_i^* = \frac{V_c^{*p}}{N_c \sum_{p'} V_c^{*p'}}, \quad (7)$$

using the definitions (2) and the relation $i = p + (c-1)N_p$. This personalized vector depends both on product and country indexes. In order to have the same value of personalization vector in countries we can define the second iteration vector proportional to the reduced PageRank and

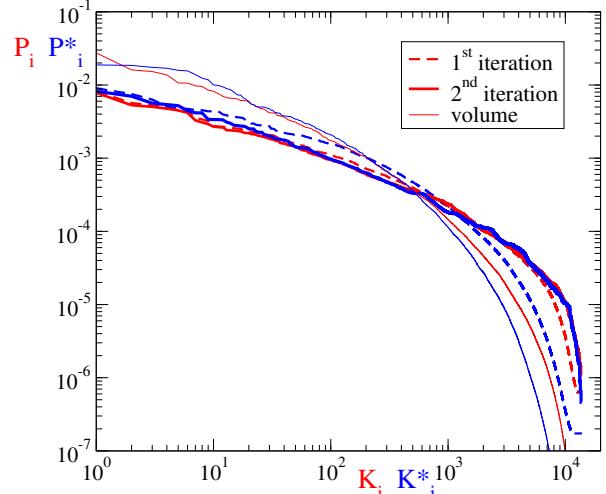


Fig. 2. Dependence of probabilities of PageRank $P(K)$, CheiRank $P^*(K^*)$, ImportRank $\hat{P}(K)$ and ExportRank $\hat{P}^*(K^*)$ as a function of their indexes in logarithmic scale for WTN in 2008 with $\alpha = 0.5$ at $N_c = 227$, $N_p = 1$, $N = 13847$. Here the results for GPVM after 1st and 2nd iterations are shown for PageRank (CheiRank) in red (blue) with dashed and solid curves, respectively. ImportRank and ExportRank (trade volume) are shown by red and blue thin curves, respectively. The fit exponents for PageRank and CheiRank are $\beta = 0.61, 0.7$ for the first iteration, $\beta = 0.59, 0.65$ for the second iteration, and $\beta = 0.94, 1.04$ for ImportRank and ExportRank (for the range $K \in [10, 2000]$).

CheiRank vectors in products obtained from the GPVM Google matrix of the first iteration:

$$v'(i) = \frac{P_p}{N_c}, v'^*(i) = \frac{P^*_p}{N_c}. \quad (8)$$

In this way we keep democracy in countries but weighted products. This second iteration personalized vectors are used for the main part of computations and operations with G and G^* . This procedure with two iterations forms our GPVM approach. The difference between results obtained from the first and second iterations is not very large (see Figs. 2 and 3) but a detailed analysis of ranking of countries and products shows that the personalized vector for the second iteration improves the results making them more stable and less fluctuating. In all figures below (except Figs. 2 and 3) we show the results after the second iteration.

The obtained results show the distribution of nodes on the PageRank-CheiRank plane (K, K^*). In addition to two ranking indexes K, K^* we use also 2DRank index K_2 which combines the contribution of these indexes as described in [8]. The ranking list $K_2(i)$ is constructed by increasing $K \rightarrow K + 1$ and increasing 2DRank index $K_2(i)$ by one if a new entry is present in the list of first $K^* < K$ entries of CheiRank, then the one unit step is done in K^* and K_2 is increased by one if the new entry is present in the list of first $K < K^*$ entries of CheiRank. More formally, 2DRank $K_2(i)$ gives the ordering of the sequence of sites, that appear inside the squares

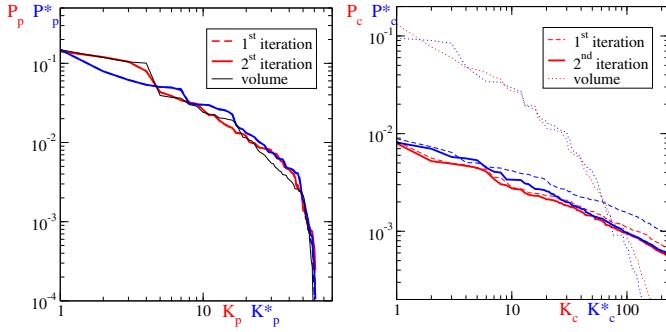


Fig. 3. Probability distributions of PageRank and CheiRank for products $P_p(K_p)$, $P_p^*(K_p^*)$ (left panel) and countries $P_c(K_c)$, $P_c^*(K_c^*)$ (right panel) in logarithmic scale for WTN from Figure 2. Here the results for the 1st and 2nd GPVM iterations are shown by red (blue) curves for PageRank (CheiRank) with dashed and solid curves, respectively. The probabilities from the trade volume ranking are shown by black curve (left) and dotted red and blue curves (right) for ImportRank and ExportRank, respectively.

[1, 1; $K = k, K^* = k; \dots$] when one runs progressively from $k = 1$ to N . Additionally, we analyze the distribution of nodes for reduced indexes (K_p, K_p^*) , (K_c, K_c^*) .

We also characterize the localization properties of eigenstates of G, G^* by the inverse participation ratio (IPR) defined as $\xi = (\sum_i |\psi_i|^2)^2 / \sum_i |\psi_i|^4$. This characteristic determines an effective number of nodes which contribute to a formation of a given eigenstate (see details in Ref. [9]).

2.3 Correlators of PageRank and CheiRank vectors

Following previous works [7,8,11] the correlator of PageRank and CheiRank vectors is defined as:

$$\kappa = N \sum_{i=1}^N P(i)P^*(i) - 1. \quad (9)$$

The typical values of κ are given in [9] for various networks.

For global PageRank and CheiRank the product-product correlator matrix is defined as:

$$\kappa_{pp'} = N_c \times \sum_{c=1}^{N_c} \left[\frac{P(p + (c-1)N_p)P^*(p' + (c-1)N_p)}{\sum_{c'} P(p + (c'-1)N_p) \sum_{c''} P^*(p' + (c''-1)N_p)} \right] - 1. \quad (10)$$

Then the correlator for a given product is obtained from (10) as:

$$\kappa_p = \kappa_{pp'} \delta_{p,p'}, \quad (11)$$

where $\delta_{p,p'}$ is the Kronecker delta.

We also use the correlators obtained from the probabilities traced over products ($P_c = \sum_p P_{pc}$) and over

countries ($P_p = \sum_c P_{pc}$) which are defined as:

$$\begin{aligned} \kappa(c) &= N_c \sum_{c=1}^{N_c} P_c P_c^* - 1, \\ \kappa(p) &= N_p \sum_{p=1}^{N_p} P_p P_p^* - 1. \end{aligned} \quad (12)$$

In the above equations (9)–(12) the correlators are computed for PageRank and CheiRank probabilities. We can also compute the same correlators using probabilities from the trade volume in ImportRank \hat{P} and ExportRank \hat{P}^* defined by (3).

We discuss the values of these correlators in Section 4.

3 Data description

All data are obtained from the COMTRADE database [1]. We used products from COMTRADE SITC Rev. 1 classification with number of products $N_p = 10$ and 61 . We choose SITC Rev. 1 since it covers the longest time interval. The main results are presented for $N_p = 61$ with up to $N_c = 227$ countries. The names of products are given in Table 1, their ImportRank index K and their fraction (in percent) of global trade volume in years 1998 and 2008 are given in Table 2. The data are collected and presented for the years 1962–2010. Our data and results are available at [25], the data for the matrices $M_{c,c'}^p$ are available at COMTRADE [1] with the rules of their distribution policy. Following [11] we use for countries ISO 3166-1 alpha-3 code available at Wikipedia.

4 Results

We apply the above methods to the described data sets of COMTRADE and present the obtained results below.

4.1 PageRank and CheiRank probabilities

The dependence of probabilities of PageRank $P(K)$ and CheiRank $P^*(K^*)$ vectors on their indexes K, K^* are shown in Figure 2 for a selected year 2008. The results can be approximately described by an algebraic dependence $P \propto 1/K^\beta$ with the exponent values given in the caption. It is interesting to note that we find approximately the same $\beta \approx 0.6$ both for PageRank and CheiRank in contrast to the WWW, universities and Wikipedia networks where usually one finds $\beta \approx 1$ for PageRank and $\beta \approx 0.6$ for CheiRank [6,9]. We attribute this to an intrinsic property of WTN where the countries try to keep economy balance of their trade. The data show that the range of probability variation is reduced for the Google ranking compared to the volume ranking. This results from a democratic ranking of countries used in the Google matrix analysis that gives a reduction of richness dispersion between countries. The results also show that the variation

of probabilities for 1st and 2nd GPVM results are not very large that demonstrates the convergence of this approach.

After tracing probabilities over countries we obtain probability distributions $P_p(K_p)$, $P_p^*(K_p^*)$ over products shown in Figure 3. The variation range of probabilities is the same as for the case of volume ranking. This shows that the GPVM approach correctly treats products keeping their contributions proportional to their volume. The difference between 1st and 2nd iterations is rather small and is practically not visible on this plot. The important result well visible here is a visible difference between PageRank and CheiRank probabilities while there is no difference between ImportRank and ExportRank probabilities since they are equal after tracing over countries.

After tracing over products we obtain probability distributions $P_c(K_c)$, $P_c^*(K_c^*)$ over countries shown in Figure 3. We see that the probability of volume ranking varies approximately by a factor 1000 while for PageRank and CheiRank such a factor is only approximately 10. Thus the democracy in countries induced by the Google matrix construction reduces significantly the variations of probabilities among countries and inequality between countries.

Both panels of Figure 3 show relatively small variations between 1st and 2nd GPVM iterations confirming the stability of this approach. In next sections we present the results only for 2nd GPVM iteration. This choice is confirmed by consideration of ranking positions of various nodes of global matrices G, G^* which show less fluctuations compared to the results of the 1st GPVM iteration.

From the global ranking of countries and products we can select a given product and then determine local ranking of countries in a given product to see how strong is their trade for this product. The results for three selected products are discussed below for year 2008. For comparison we also present comparison with the export-import ranking from the trade volume.

4.2 Ranking of countries and products

After tracing the probabilities $P(K)$, $P^*(K^*)$ over products we obtain the distribution of world countries on the PageRank-CheiRank plane (K_c, K_c^*) presented in Figure 4 for a test year 2008. In the same figure we present the rank distributions obtained from ImportRank-ExportRank probabilities of trade volume and the results obtained in [11] for trade in *all commodities*. For the GPVM data we see the global features already discussed in [11]: the countries are distributed in a vicinity of diagonal $K_c = K_c^*$ since each country aims to keep its trade balanced. The top 20 list of top K_2 countries recover 15 of 19 countries of G20 major world economies (EU is the number 20) thus obtaining 79% of the whole list. This is close to the percent obtained in [11] for trade in *all commodities*.

The global distributions of top countries with $K_c \leq 40$, $K_c^* \leq 40$ for the three ranking methods, shown in Figure 4, are similar on average. But some modifications introduced by the GPVM analysis are visible. Thus China (CHN) moves on 2nd position of CheiRank while it is in

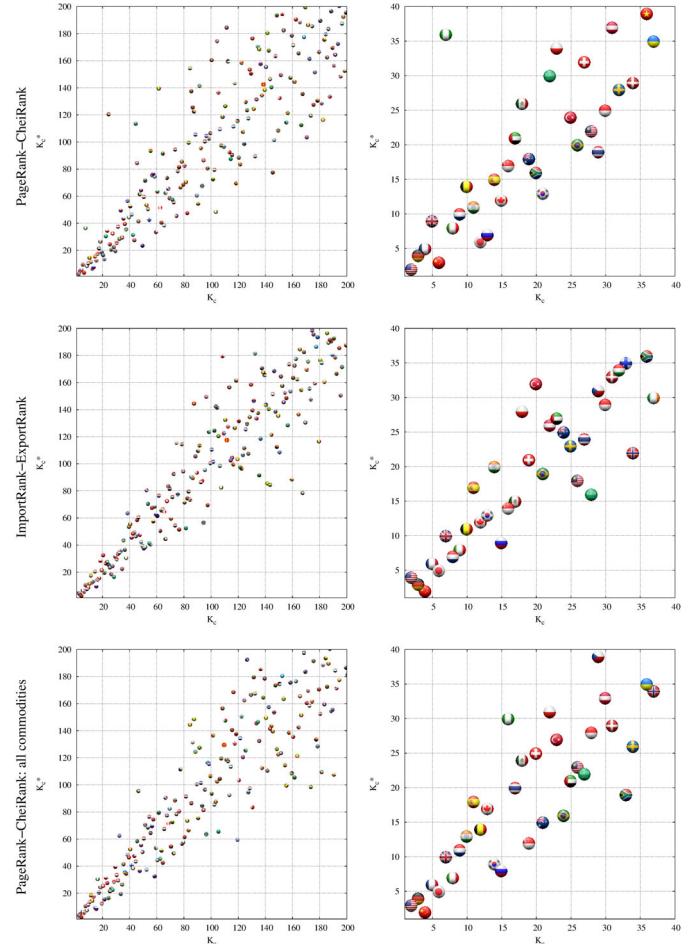


Fig. 4. Country positions on PageRank-CheiRank plane (K_c, K_c^*) obtained by the GPVM analysis (top panels), ImportRank-ExportRank of trade volume (center panels), and for PageRank-CheiRank of *all commodities* (bottom panels, data from [11]). Left panels show global scale ($K_c, K_c^* \in [1, 200]$) and right panels show zoom on top ranks ($K_c, K_c^* \in [1, 40]$). Each country is shown by circle with its own flag (for a better visibility the circle center is slightly displaced from its integer position (K_c, K_c^*) along direction angle $\pi/4$). Data are shown for year 2008.

the 1st position for trade volume ranking and CheiRank of *all commodities*. Also e.g. Saudi Arabia (SAU) and Russia (RUS) move from the CheiRank positions $K_c^* = 21$ and $K_c^* = 7$ in *all commodities* [11] to $K_c^* = 29$ and $K_c^* = 6$ in the GPVM ranking, respectively. Other example is a significant displacement of Nigeria (NGA). We explain such differences as the result of larger connectivity required for getting high ranking in the multiproduct WTN. Indeed, China is more specialized in specific products compared to USA (e.g. no petroleum production and export) that leads to its displacement in K_c^* . We note that the ecological ranking gives also worse ranking positions for China comparing to the trade volume ranking [12]. In a similar way the trade of Saudi Arabia is strongly dominated by petroleum and moreover its petroleum trade is strongly oriented on USA that makes its trade network concentrated on a few links while Russia is improving its

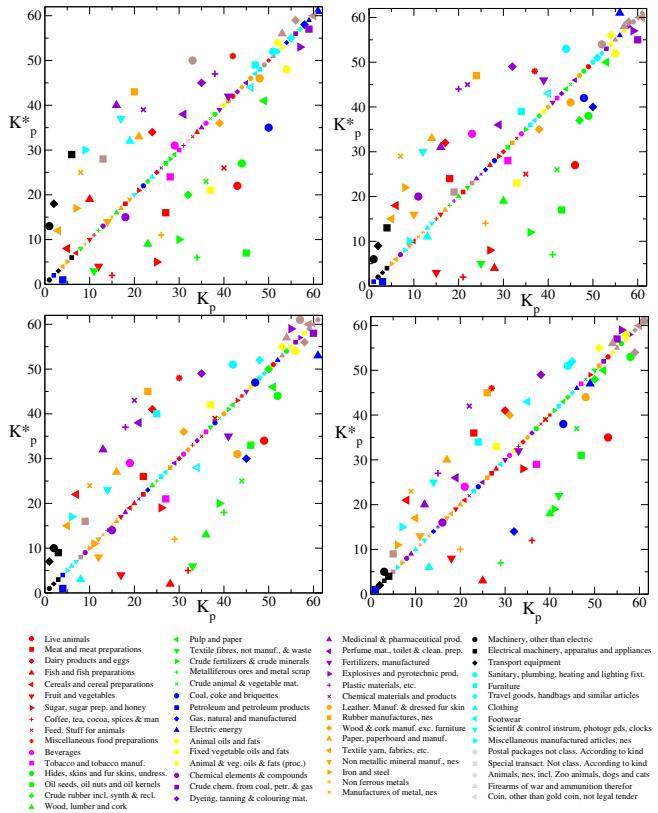


Fig. 5. Two dimensional ranking of products on the PageRank-CheiRank plane (K_p, K_p^*). Each product is represented by its specific combination of color and symbol: color illustrates the first digit of COMTRADE SITC Rev. 1 code with the corresponding name shown in the legend on the bottom; symbols correspond to product names listed in Table 1 with their code numbers; the order of names on the bottom panel of symbols of this figure is the same as in Table 1 (counting from top to bottom and left to right). The trade volume ranking via ImportRank-ExportRank is shown by small symbols at the diagonal $\hat{K}_p = \hat{K}_p^*$, after tracing over countries this ranking is symmetric in products. Top left and right panels show years 1963 and 1978, while bottom left and right panels show years 1993 and 2008, respectively.

position in K_c^* due to significant trade links with EU and Asia.

In global, the comparison of three ranks of countries shown in Figure 4 confirms that the GPVM analysis gives a reliable ranking of multiproduct WTN. Thus we now try to obtain new features of multiproduct WTN using the GPVM approach.

The main new feature obtained within the GPVM approach is shown in Figure 5 which gives the distribution of products on the PageRank-CheiRank plane (K_p, K_p^*) after tracing of global probabilities $P(K), P^*(K^*)$ over all world countries. The data clearly show that the distribution of products over this plane is asymmetric while the ranking of products from the trade volume produces the symmetric ranking of products located directly on diagonal $K_p = K_p^*$. Thus the functions of products are asymmetric: some of them are more oriented to export

(e.g. 03 Fish and fish preparations, 05 Fruit and vegetables, 26 Textile fibers, not manuf. etc., 28 Metalliferous ores and metal scrap, 84 Clothing); in last years (e.g. 2008) 34 Gas, natural and manufactured also takes well pronounced export oriented feature characterized by location in the lower right triangle ($K_p^* < K_p$) of the square plane (K_p, K_p^*). In contrast to that the products located in the upper left triangle ($K_p^* > K_p$) represent import oriented products (e.g. 02 Dairy products and eggs, 04 Cereals and cereal preparations, 64 Paper, paperboard and manuf., 65 Textile yarn, fabrics, etc., 86 Scientific & control instrum., photogr gds, clocks).

It is interesting to note that the machinery products 71, 72, 73 are located on leading import oriented positions in 1963, 1978, 1993 but they become more close to symmetric positions in 2008. We attribute this to development of China that makes the trade in these products more symmetric in import-export. It is interesting to note that in 1993 the product 33 Petroleum and petroleum products loses its first trade volume position due to low petroleum prices but still it keeps the first CheiRank position showing its trade network importance for export. Each product moves on (K_p, K_p^*) with time. However, a part of the above points, we can say that the global distribution does not manifest drastic changes. Indeed, e.g. the green symbols of first digit 2 remain export oriented for the whole period 1963–2008. We note that the established asymmetry of products orientation for the world trade is in agreement with the similar indications obtained on the basis of ecological ranking in [12]. However, the GPVM approach used here have more solid mathematical and statistical foundations with a reduced significance of fluctuations comparing to the ecological ranking.

The comparison between the GPVM and trade volume ranking methods provides interesting information. Thus in petroleum code 33 we have on top positions Russia, Saudi Arabia, United Arab Emirates while from the CheiRank order of this product we find Russia, USA, India (see Fig. 6 and Tab. 3). This marks the importance of the role of USA and India played in the WTN and in the redistribution of petroleum over nearby region countries, e.g. around India. Also Singapore is on a local petroleum position just behind India and just before Saudi Arabia, see Table 3. This happens due to strong involvement of India and Singapore in the trade redistribution flows of petroleum while Saudi Arabia has rather restricted trade connections strongly oriented on USA and nearby countries.

For electrical machinery 72 there are less modifications in the top export or CheiRank positions (see Fig. 6) but we observe significant broadening of positions on PageRank-CheiRank plane comparing to ImportRank-ExportRank. Thus, Asian countries (China, Japan, S. Korea, Singapore) are located on the PageRank-CheiRank plane well below the diagonal $K = K^*$ showing a significant trade advantages of these countries in product 72 comparing to Western countries (USA, Germany, France, UK).

Another product, shown in Figure 6, is 03 Fish and fish preparations. According to the trade volume export ranking the top three positions are attributed to China,

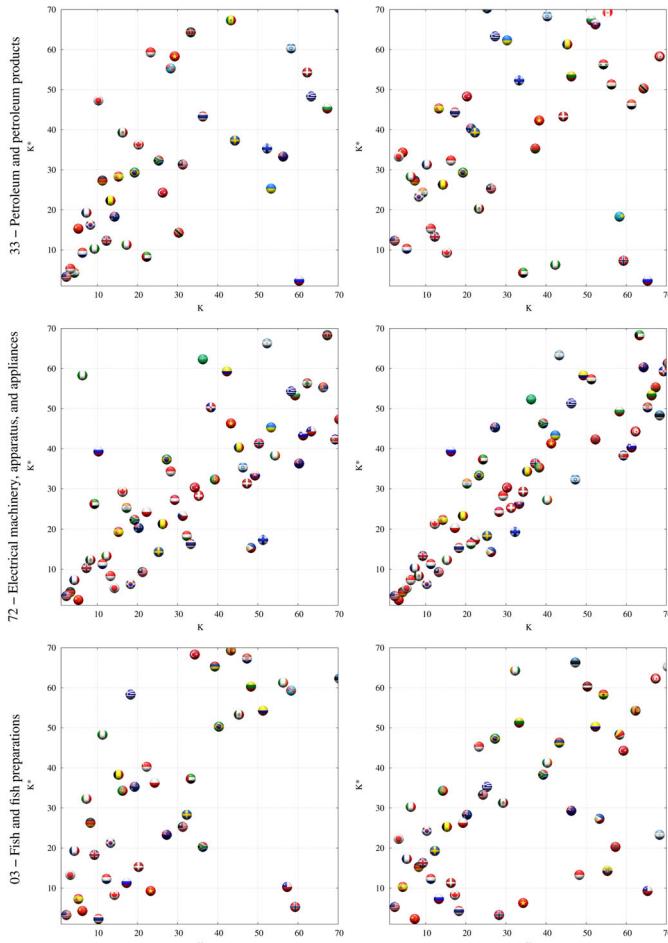


Fig. 6. Left panels show results of the GPVM data for country positions on PageRank-CheiRank plane of local rank values K, K^* ordered by (K_{cp}, K_{cp}^*) for specific products with $p = 33$ (top panel), $p = 72$ (center panel) and $p = 03$ (bottom panel). Right panels show the ImportRank-ExportRank planes respectively for comparison. Data are given for year 2008. Each country is shown by circle with its own flag as in Figure 4.

Norway, Thailand. However, from CheiRank of product 03 we find another order with Thailand, USA, China. This result stresses again the broadness and robustness of the trade connections of Thailand and USA. As another example we note a significant improvement of Spain CheiRank position showing its strong commercial relations for product 03. On the other side Russia has relatively good position in the trade volume export of 03 product but its CheiRank index becomes worse due to absence of broad commercial links for this product.

The global top 20 positions of indexes $K, K^*, K_2, \hat{K}, \hat{K}^*$ are given in Table 3 for year 2008. We note a significant improvement of positions of Singapore and India in PageRank-CheiRank positions comparing to their positions in the trade volume ranking. This reflects their strong commercial relations in the world trade. In the trade volume ranking the top positions are taken by 33 petroleum and digit 7 of machinery products. This remains mainly true for PageRank-CheiRank positions but

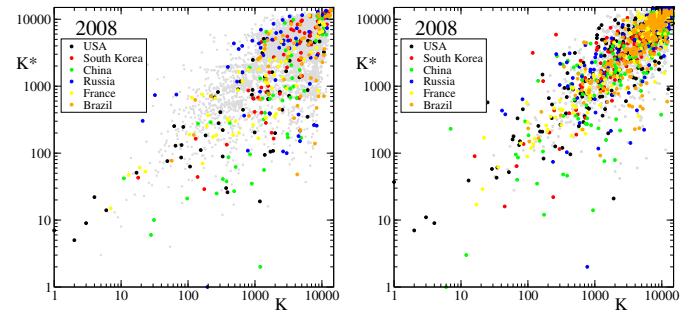


Fig. 7. Global plane of rank indexes (K, K^*) for PageRank-CheiRank (left panel) and ImportRank-ExportRank (right panel) for $N = 13847$ nodes in year 2008. Each country and product pair is represented by a gray circle. Some countries are highlighted in colors: USA with black, South Korea with red, China with green, Russia with blue, France with yellow and Brazil with orange.

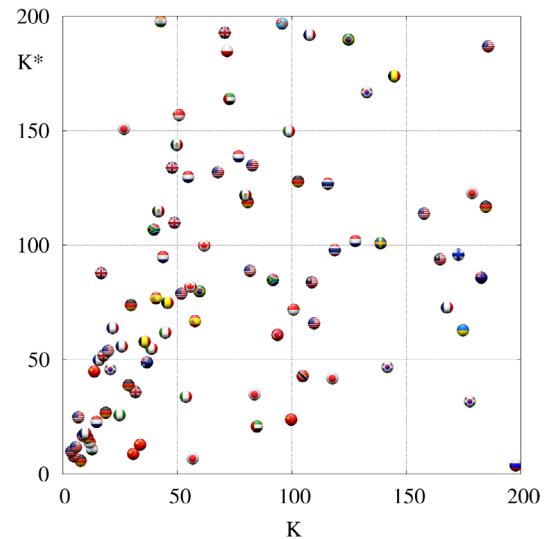


Fig. 8. Top 200 global PageRank-CheiRank indexes (K, K^*) distributions for year 2008. Each country (for different products) is represented by its flag.

we see the spectacular improvement of positions of 84 *Clothing* for China ($K^* = 2$) and 93 *Special transact.* for USA ($K = 4$) showing thus these two products have strong commercial exchange all over the world even if their trade volume is not so dominant.

We show the plane (K, K^*) for the global world ranking in logarithmic scale in 2008 in Figure 7. The positions of trade nodes of certain selected countries are shown by color. We observe that the trade volume gives a higher concentration of nodes around diagonal comparing to the GPVM ranking. We attribute this to the symmetry of trade volume in products.

In Figure 8 we show the distributions of top 200 ranks of the PageRank-CheiRank plane (zoom of left panel of Fig. 7). Among the top 30 positions of K^* there are 8 products of USA, 6 of China, 3 of Germany and other countries with less number of products. The top position at $K^* = 1$ corresponds to product 33 of Russia while Saudi

Table 3. Top 20 ranks for global PageRank K , CheiRank K^* , 2dRank K_2 , ImportRank \hat{K} and ExportRank \hat{K}^* for given country and product code for year 2008.

#	K country & code	K^* country & code	K_2 country & code	\hat{K} country & code	\hat{K}^* country & code
1	USA 33	Russia 33	Germany 73	USA 33	China 72
2	USA 73	China 84	USA 73	USA 71	Russia 33
3	USA 71	Germany 73	USA 33	USA 72	China 71
4	USA 93	Japan 73	USA 71	USA 73	Germany 73
5	Germany 73	USA 73	India 33	Japan 33	Germany 71
6	USA 72	China 72	Singapore 33	China 72	Saudi Arabia 33
7	France 73	USA 33	Germany 71	China 33	USA 71
8	Germany 71	India 33	USA 72	Germany 71	Japan 73
9	Singapore 33	USA 71	France 73	Germany 73	USA 73
10	India 33	China 71	Netherlands 33	Netherlands 33	Japan 71
11	China 33	Singapore 33	USA 93	Germany 72	USA 72
12	Netherlands 33	Saudi Arabia 33	Nigeria 33	China 71	China 89
13	France 33	Germany 71	Germany 72	USA 89	Germany 72
14	UK 71	USA 72	China 72	Italy 33	China 84
15	UK 73	France 73	China 71	Germany 33	Japan 72
16	Germany 72	Thailand 3	UK 33	South Korea 33	South Korea 72
17	USA 89	Kazakhstan 33	Germany 93	France 73	France 73
18	South Korea 33	U. Arab Emir. 33	China 33	China 28	Italy 71
19	France 71	USA 28	South Korea 33	Germany 93	U. Arab Emir. 33
20	Sudan 73	Netherlands 33	Australia 33	India 33	Germany 93

Arabia is only at $K^* = 12$ for this product. The lists of all $N = 13\,847$ network nodes with their K, K_2, K^* values are available at [25].

4.3 Time evolution of ranking

The time evolution of indexes of products K_p, K_p^* is shown in Figure 9. To obtain these data we trace PageRank and CheiRank probabilities over countries and show the time evolution of rank indexes of products K_p, K_p^* for top 15 rank products of year 2010. The product 33 *Petroleum and petroleum products* remains at the top CheiRank position $K_p^* = 1$ for the whole period while in PageRank it shows significant variations from $K_p = 1$ to 4 being at $K_p = 4$ at 1986–1999 when the petroleum had a low price. Products with first digit 7 have high ranks of K_p but especially strong variation is observed for K_p^* of 72 *Electrical machinery* moving from position 26 in 1962 to 4 in 2010. Among other indexes with strong variations we note 58 *Plastic materials*, 84 *Clothing*, 93 *Special transact.*, 34 *Gas, natural and manufactured*.

The time evolution of products 33 and 72 on the global index plane (K, K^*) is shown in Figure 10 for 6 countries from Figure 7. Thus for product 72 we see a striking improvement of K^* for China and Korea that is at the origin of the global importance improvement of K^* in Figure 9. For the product 33 in Figure 10 Russia improves significantly its rank positions taking the top rank $K^* = 1$ (see also Tab. 3).

The variation of global ranks K, K^* with time is shown for 4 products and 10 countries in Figure 11. For products 72, 73 on a scale of 50 years we see a spectacular improvement of K^* for China, Japan, Korea. For the product 33 we see strong improvement of K^* for Russia in

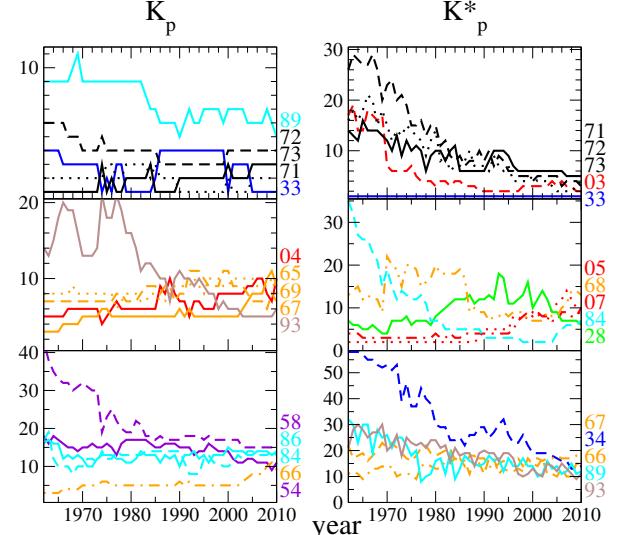


Fig. 9. Time evolution of PageRank K_p and CheiRank K_p^* indexes for years 1962 to 2010 for certain products marked on the right panel side by their codes from Table 1. Top panels show top 5 ranks of 2010, middle and bottom panels show ranks 6 to 10 and 11 to 15 for 2010, respectively. Colors of curves correspond to the colors of Figure 5 marking the first code digit.

last 15 years. It is interesting to note that at the period 1986–1992 of cheap petroleum 33 USA takes the top position $K^* = 1$ with a significant increase of its corresponding K value. We think that this is a result of political decision to make an economical pressure on USSR since such an increase of export of cheap price petroleum is not justified from the economical view point.

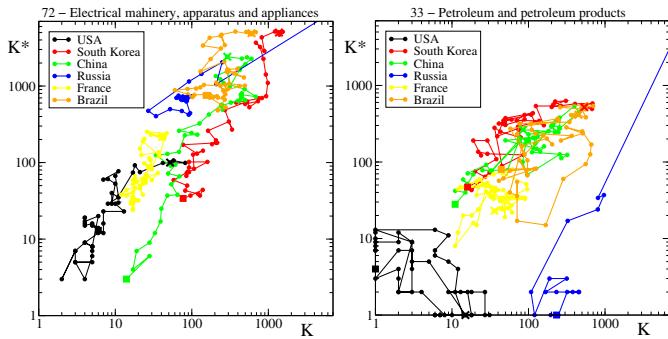


Fig. 10. Time evolution of ranking of two products 72 and 33 for 6 countries of Figure 7 shown on the global PageRank-CheiRank plane (K, K^*). Left and right panels show the cases of 72 *Electrical machinery, apparatus and appliances* and 33 *petroleum and petroleum products*, respectively. The evolution in time starts in 1962 (marked by cross) and ends in 2010 (marked by square).

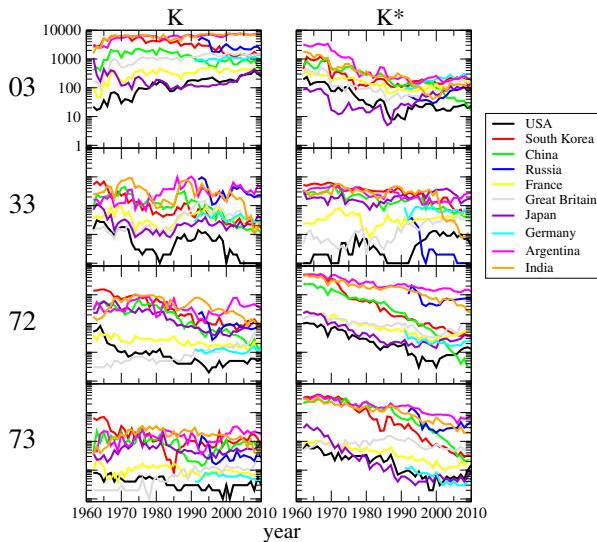


Fig. 11. Time evolution of global ranking of PageRank and CheiRank indexes K, K^* for selected 10 countries and 4 products. Left and right panels show K and K^* as a function of years for products: 03 *Fish and fish preparations*; 33 *Petroleum and petroleum products*; 72 *Electrical machinery, apparatus and appliances*; and 73 *Transport equipment* (from top to bottom). In all panels the ranks are shown in logarithmic scale for 10 given countries: USA, South Korea, China, Russia, France, Brazil, Great Britain, Japan, Germany and Argentina marked by curve colors.

For the product 33 we also note a notable improvement of K^* of India which is visible in CheiRank but not in ExportRank (see Tab. 3). We attribute this not to a large amount of trade volume but to a significant structural improvements of trade network of India in this product. We note that the strength and efficiency of trade network is also at the origin of significant improvement of PageRank and CheiRank positions of Singapore comparing to the trade volume ranking. Thus the development of trade connections of certain countries significantly improves their

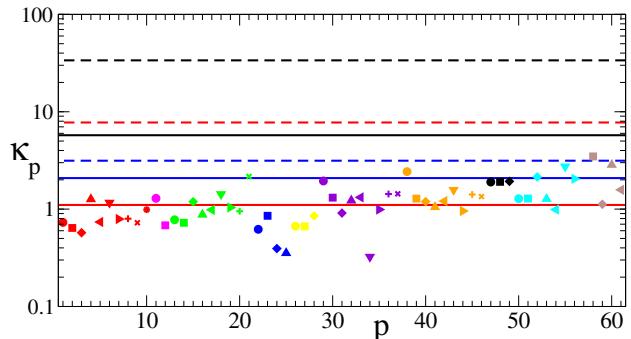


Fig. 12. PageRank-CheiRank correlators κ_p (11) from the GPVM are shown as a function of the product index p with the corresponding symbol from Figure 5. PageRank-CheiRank and ImportRank-ExportRank correlators are shown by solid and dashed lines respectively, where the global correlator κ (9) is shown in black, the correlator for countries $\kappa(c)$ (12) is shown by red lines, the correlator for products $\kappa(p)$ (12) is shown by blue lines. Here product number p is counted in order of appearance in Table 1. The data are given for year 2008 with $N_p = 61, N_c = 227, N = 13\,847$.

Google rank positions. For the product 03 we note the improvement of K^* positions of China and Argentina while Russia shows no improvements in this product trade for this time period.

4.4 Correlation properties of PageRank and CheiRank

The properties of κ correlator of PageRank and CheiRank vectors for various networks are reported in [7,9]. There are directed networks with small or even slightly negative values of κ , e.g. Linux Kernel or Physical Review citation networks, or with $\kappa \sim 4$ for Wikipedia networks and even larger values $\kappa \approx 116$ for the Twitter network.

The values of correlators defined by equations (9)–(12) are shown in Figures 12 and 13 for a typical year 2008. For the global PageRank-CheiRank correlator we find $\kappa \approx 5.7$ (9) while for Import-Export probabilities the corresponding value is significantly larger with $\kappa \approx 33.7$. Thus the trade volume ranking with its symmetry in products gives an artificial increase of κ by a significant factor. A similar enhancement factor of Import-Export remains for correlators in products $\kappa(p)$ and countries $\kappa(c)$ from equation (12) while for PageRank-CheiRank we obtain moderate correlator values around unity (see Fig. 12). The PageRank-CheiRank correlator κ_p (11) for specific products have relatively low values with $\kappa_p < 1$ for practically all products with $p \leq 45$ (we remind that here p counts the products in the order of their appearance in Tab. 1, it is different from COMTRADE code number).

The correlation matrix of products $\kappa_{pp'}$ (10) is shown in Figure 13. This matrix is asymmetric and demonstrates the existence of relatively high correlations between products 73 *Transport equipment*, 65 *Textile yarn, fabrics, made up articles, etc.* and 83 *Travel goods, handbags and similar articles* that all are related with transportation of products.

Table 4. Top 10 values of 4 different eigenvectors from Figure 16. The corresponding eigenvalues form left to right are $\lambda = 0.9548$, $\lambda = 0.9345$, $\lambda = 0.452 + i0.775$ and $\lambda = 0.424 + i0.467$. There is only one product in each of these top 10 list nodes which are: 57 *Explosives and pyrotechnic products*; 06 *Sugar, sugar preparations and honey*; 56 *Fertilizers, manufactured*; 52 *Crude chemicals from coal, petroleum and gas*.

K_i	$ \psi_i $	Country	$ \psi_i $	Country	$ \psi_i $	Country	$ \psi_i $	Country
		prod: 57		prod:06		prod:56		prod:52
1	0.052	USA	0.216	Mali	0.332	Brazil	0.288	Japan
2	0.044	Tajikistan	0.201	Guinea	0.304	Bolivia	0.279	Rep. of Korea
3	0.042	Kyrgyzstan	0.059	USA	0.274	Paraguay	0.245	China
4	0.022	France	0.023	Germany	0.031	Argentina	0.020	Australia
5	0.021	Mexico	0.021	Mexico	0.017	Uruguay	0.013	USA
6	0.018	Italy	0.021	Canada	0.009	Chile	0.012	U. Arab Em.
7	0.018	Canada	0.018	UK	0.004	Portugal	0.010	Canada
8	0.015	Germany	0.015	Israel	0.004	Angola	0.010	Singapore
9	0.013	U. Arab Em.	0.015	C. d'Ivoire	0.004	Spain	0.009	Germany
10	0.012	Qatar	0.014	Japan	0.003	France	0.008	New Zealand

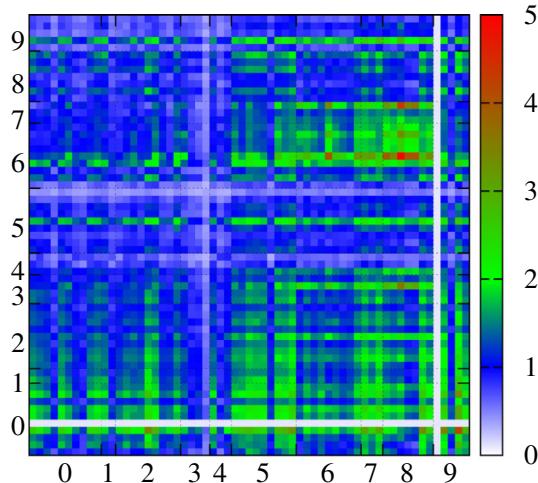


Fig. 13. Product PageRank-CheiRank correlation matrix $\kappa_{p,p'}$ (10) for year 2008 with correlator values shown by color. The code indexes p and p' of all $N_p = 61$ products are shown on x and y axes by their corresponding first digit (see Tab. 1).

4.5 Spectrum and eigenstates of WTN Google matrix

Above we analyzed the properties of eigenstates of G and G^* at the largest eigenvalue $\lambda = 1$. However, in total there are N eigenvalues and eigenstates. The results obtained for the Wikipedia network [26] demonstrated that eigenstates with large modulus of λ correspond to certain specific communities of the network. Thus it is interesting to study the spectral properties of G for the multiproduct WTN. The spectra of G and G^* are shown in Figure 14 for year 2008. It is interesting to note that for G the spectrum shows some similarities with those of Wikipedia (see Fig. 1 in [26]). At $\alpha = 1$ there are 12 and 7 degenerate eigenvalues $\lambda = 1$ for G and G^* , respectively. Thus the spectral gap appears only for $\alpha < 1$. The dependence of IPR ξ of eigenstates of G on $Re\lambda$ is shown in Figure 15. The results show that $\xi \ll N$ so that the eigenstates are well localized on a certain group of nodes.

The eigenstates ψ_i can be ordered by their decreasing amplitude $|\psi_i|$ giving the eigenstate index K_i with

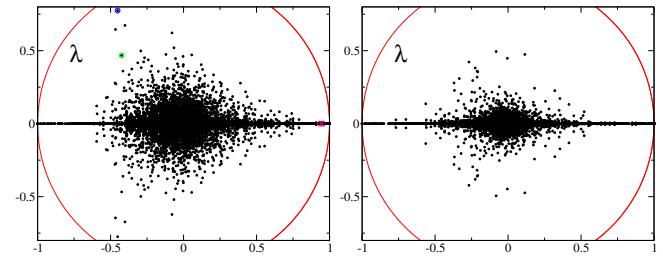


Fig. 14. Spectrum of Google matrices G (left panel) and G^* (right panel) represented in the complex plane of λ . The data are for year 2008 with $\alpha = 1$, and $N = 13847$, $N_c = 227$, $N_p = 61$. Four eigenvalues marked by colored circles are used for illustration of eigenstates in Figures 15 and 16.

the largest amplitude at $K_i = 1$. The examples of four eigenstates are shown in Figure 16. We see that the amplitude is mainly localized on a few top nodes in agreement of small values of $\xi \sim 4$ shown in Figure 15. The top ten amplitudes of these four eigenstates are shown in Table 4 with corresponding names of countries and products. We see that for a given eigenstate these top ten nodes correspond to one product clearly indicating strong links of trade between certain countries. Thus for 06 *Sugar* we see strong link between geographically close Mali and Guinea with further links to USA, Germany, etc. In a similar way for 56 *Fertilizers* there is a group of Latin American countries Brazil, Bolivia, Paraguay linked to Argentina, Uruguay, etc. We see a similar situation for products 57 and 52. These results confirm the observation established in [26] for Wikipedia that the eigenstates with large modulus of λ select interesting specific network communities. We think that it would be interesting to investigate the properties of eigenstates in further studies.

4.6 Sensitivity to price variations

Above we established the global mathematical structure of multiproduct WTN and presented results on its ranking and spectral properties. Such ranking properties bring new interesting and important information about the WTN.

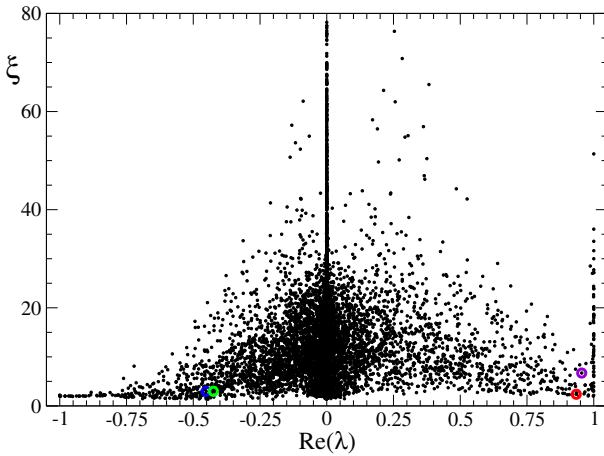


Fig. 15. Inverse participation ratio (IPR) ξ of all eigenstates of G as a function of the real part of the corresponding eigenvalue λ from the spectrum of Figure 14. The eigenvalues marked by color circles are those from Figure 14.

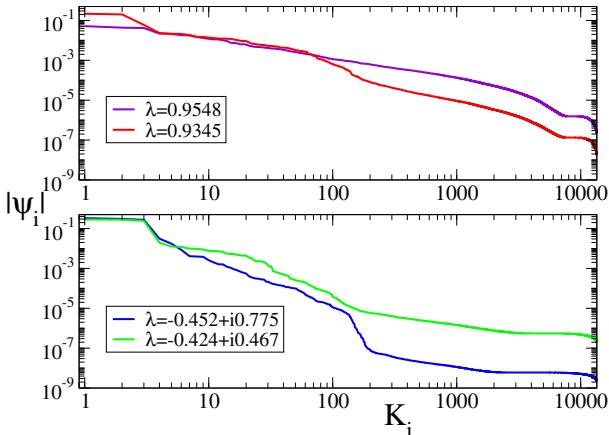


Fig. 16. Eigenstate amplitudes $|\psi_i|$ ordered by its own decreasing amplitude order with index K_i for 4 different eigenvalues of Figure 14 (states are normalized as $\sum_i |\psi_i| = 1$). Top panel shows two examples of real eigenvalues with $\lambda = 0.9548$ and $\lambda = 0.9345$ while bottom panel shows two eigenvalues with large imaginary part with $\lambda = 0.452 + i0.775$ and $\lambda = 0.424 + i0.467$. Node names (country, product) for top ten largest amplitudes of these eigenvectors are shown in Table 4.

However, from the view point of economy it is more important to analyze the effects of crisis contamination and price variations. Such an analysis represents a complex task to which we hope to return in our further investigations. However, the knowledge of the global WTN structure is an essential building block of this task and we think that the presented results demonstrate that this block is available now.

Using the knowledge of WTN structure, we illustrate here that it allows to obtain nontrivial results on sensitivity to price variations for certain products. We consider as an example year 2008 and assume that the price of product 33 *Petroleum and petroleum products* is increased by a relative fraction δ going from its unit value 1 to $1 + \delta$ (or $\delta = \delta_{33}$). Then we compute the derivatives of probabilities

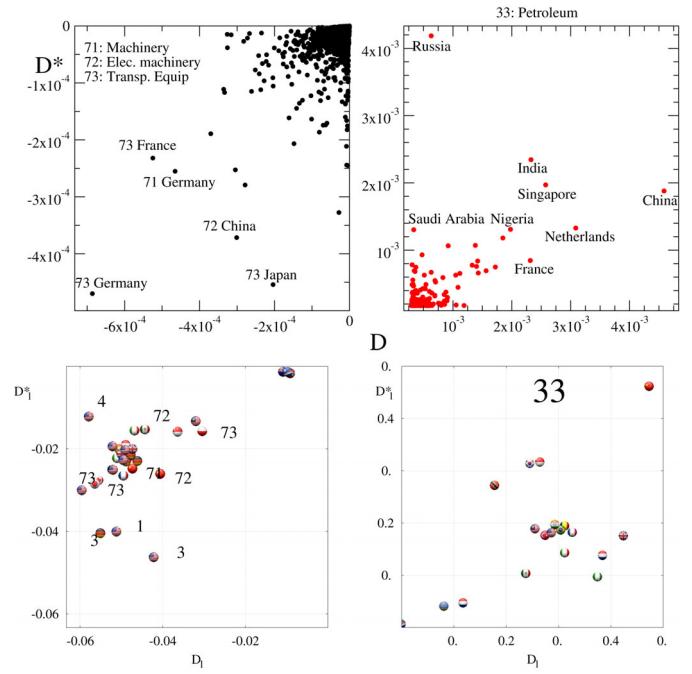


Fig. 17. Derivatives $D = dP/d\delta_{33}$ and $D^* = dP^*/d\delta_{33}$ for a price variation δ_{33} of 33 *Petroleum and petroleum products* for year 2008. Top left and right panels show the cases of negative and positive D and D^* respectively, with some products and countries labeled by their 2 digit code. Bottom panels show the positive and negative cases of the logarithmic derivatives $D_l = D/P$ and $D_l^* = D^*/P^*$ for countries and products with $K_2 \leq 50$, where the flags and 2 digit codes for countries and products are shown (in right panels only product 33 is present). Codes are described in Table 1.

of PageRank $D = dP/d\delta = \Delta P/\delta$ and CheiRank $D^* = dP^*/d\delta = \Delta P^*/\delta$. The computation is done for values of $\delta = 0.01, 0.03, 0.05$ ensuring that the result is not sensitive to a specific δ value. We also compute the logarithmic derivatives $D_l = d \ln P/d\delta$, $D_l^* = d \ln P^*/d\delta$ which give us a relative changes of P, P^* .

The results for the price variation δ_{33} of 33 *Petroleum and petroleum products* are shown in Figure 17. The derivatives for all WTN nodes are shown on the planes (D, D^*) and (D_l, D_l^*) . For (D, D^*) the nodes are distributed in two sectors with $D > 0, D^* > 0$ and $D < 0, D^* < 0$. The largest values with $D > 0, D^* > 0$ correspond to nodes of countries of product 33 which are rich in petroleum (e.g. Russia, Saudi Arabia, Nigeria) or those which have strong trade transfer of petroleum to other countries (Singapore, India, China, etc). It is rather natural that with the growth of petroleum prices the rank probabilities P, P^* of these countries grow. A more unexpected effect is observed in the sector $D < 0, D^* < 0$. Here we see that an increase of petroleum price leads to a decrease of probabilities of nodes of countries Germany, France, China, Japan trading in machinery products 71, 72, 73.

For comparison we also compute the derivatives D, D^*, D_l, D_l^* from the probabilities (3) defined by the trade volume of Import-Export instead of PageRank-CheiRank.

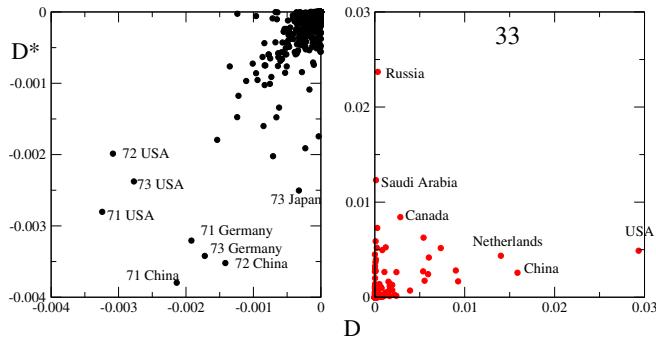


Fig. 18. Same as in top panels of Figure 17 but using probabilities from the trade volume (3).

The results are shown in Figure 18 for petroleum price variation to be compared with Figure 17. The distribution of D, D^* is rather different from those values obtained with PageRank-CheiRank probabilities. This is related to the fact that PageRank and CheiRank take into account the global network structure while the trade volume gives only local relations in trade links between countries. The difference between these two methods becomes even more striking for logarithmic derivatives D_l, D_l^* . Indeed, for the trade volume ranking the variation of probabilities P^*, P due to price variation of a given product can be computed analytically taking into account the trade volume change with δ_p . The computations give $D_{cp} = (1-f_p)P_{cp}$, $D^*_{cp} = (1-f_p)P^*_{cp}$ for a derivative of probability of product p and country c over the price of product δ_p and $D_{cp'} = -f_p P_{cp'}$, $D_{cp'} = -f_p P_{cp'}$, $D^*_{cp'} = -f_p P^*_{cp'}$ (if $p' \neq p$), where f_p is a fraction of product p in the world trade. From these expressions we see that the logarithmic derivatives are independent of country and product. Indeed, for the case of Figure 18 we obtain analytically and by direct numerical computations that $D_l = D_l^* = -0.2022$ (for all countries if $p' \neq p = 33$) and $D_l = D_l^* = 0.7916$ (for all countries if $p' = p = 33$). Due to simplicity of this case we do not show it in Figure 18.

The results for price variation of *34 Gas, natural and manufactured* are presented in Figure 19 showing derivatives of PageRank and CheiRank probabilities over δ_{34} . We see that for absolute derivatives D, D^* the mostly affected are now nodes of gas producing countries for the sector $D, D^* > 0$, while for the sector $D, D^* < 0$ the mostly affected are countries linked to petroleum production or trade, plus USA with products 71,72,73. For the sector of logarithmic derivatives $D_l, D_l^* < 0$ among top K_2 and K, K^* nodes we find nodes of countries of product 33 and also 93.

Thus the analysis of derivatives provides an interesting new information of sensitivity of world trade to price variations.

4.7 World map of CheiRank-PageRank trade balance

On the basis of the obtained WTN Google matrix we can now analyze the trade balance in various products between the world countries. Usually economists consider the ex-

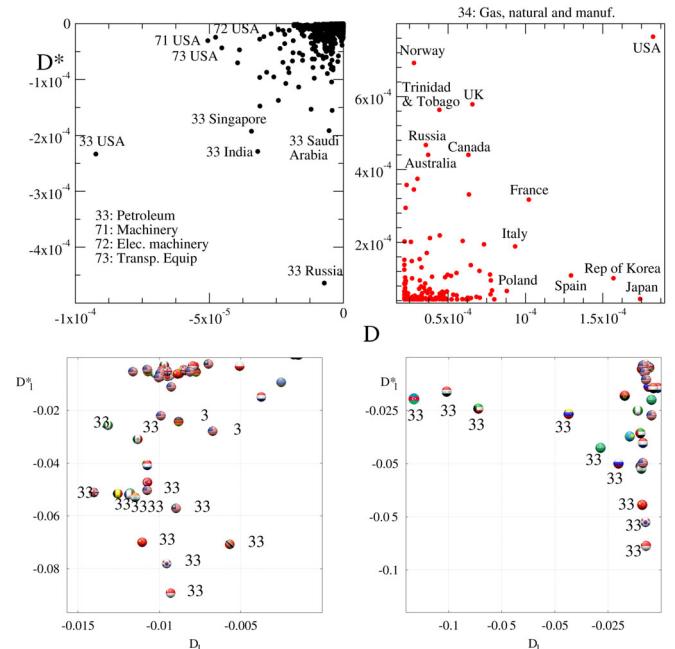


Fig. 19. Derivative of P and P^* (D and D^* respectively) for a price variation of 34 *Gas, natural and manufactured* for 2008. Top left and right panels show the cases of negative and positive sectors of D and D^* respectively, with some products and countries labeled by their 2 digit code and names (in top right panel all points correspond to product 34). Bottom panels show the cases of the logarithmic derivatives D_l and D_l^* for countries and products with $K_2 \leq 50$ (bottom left panel) and $K, K^* \leq 25$ (bottom right panel); flags and 2 digit codes for countries and products are shown. In bottom right panel ($K, K^* \leq 25$) we do not show the case of Sudan (*73 Transport equipment*) which has values of $(D_l, D_l^*) = (2 \times 10^{-4}, 1.75 \times 10^{-2})$. Codes are described in Table 1.

port and import of a given country as it is shown in Figure 1. Then the trade balance of a given country c can be defined making summation over all products:

$$\begin{aligned} B_c &= \sum_p (P_{cp}^* - P_{cp}) / \sum_p (P_{cp}^* + P_{cp}) \\ &= (P_c^* - P_c) / (P_c^* + P_c). \end{aligned} \quad (13)$$

In economy, P_c, P_c^* are defined via the probabilities of trade volume $\hat{P}_{cp}, \hat{P}_{cp}^*$ from (3). In our approach, we define P_{cp}, P_{cp}^* as PageRank and CheiRank probabilities. In contrast to the trade volume our approach takes into account the multiple network links between nodes.

The comparison of the world trade balance obtained by these two methods is shown in Figure 20. We see that the leadership of China becomes very well visible in CheiRank-PageRank balance map while it is much less pronounced in the trade volume balance. The Google matrix analysis also highlights the dis-balance of trade network of Nigeria (strongly oriented on petroleum export and machinery import) and Sudan. It is interesting to note that the positive CheiRank-PageRank balance is mainly located in the countries of BRICS (Brazil, Russia, India, China, South Africa). In contrast to that,

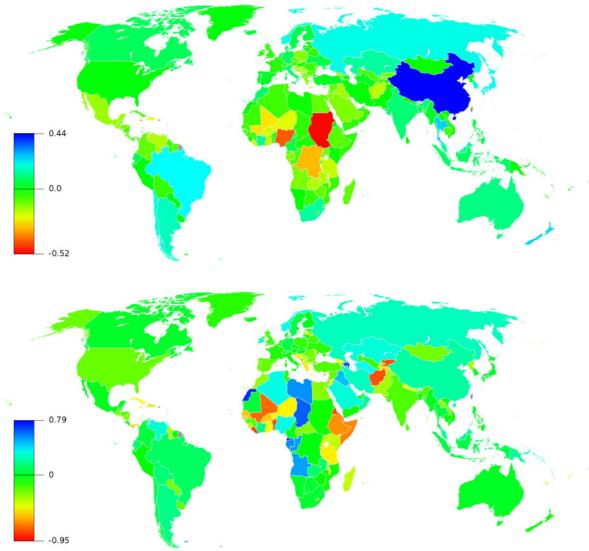


Fig. 20. World map of probabilities balance $B_c = (P_c^* - P_c) / (P_c^* + P_c)$ determined for each of $N_c = 227$ countries in year 2008. Top panel: probabilities P_c^*, P_c are given by CheiRank and PageRank vectors; bottom panel: probabilities are computed from the trade volume of Export-Import (3). Names of countries can be found at [10].

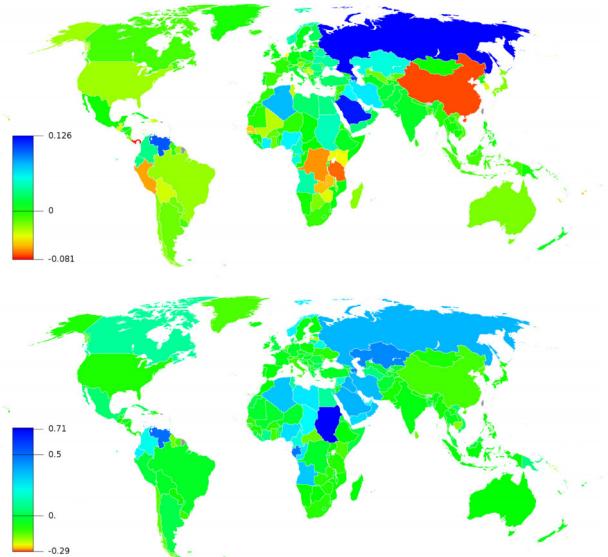


Fig. 21. Derivative of probabilities balance $dB_c/d\delta_{33}$ over petroleum price δ_{33} for year 2008. Top panel: balance of countries B_c is determined from CheiRank and PageRank vectors as in the top panel of Figure 20; bottom panel: B_c values are computed from the trade volume as in the bottom panel of Figure 20. Names of countries can be found at [10].

the usual trade volume balance highlights Western Sahara and Afghanistan at large positive and negative trade balance in 2008.

We can also determine the sensitivity of trade balance to price variation of a certain product p computing the balance derivative $dB_c/d\delta_p$. The world map sensitivity in respect to price of petroleum $p = 33$ is shown in Figure 21 for the above two methods of definition of probabilities P_c, P_c^* in (13). For the CheiRank-PageRank balance we see that the derivative $dB_c/d\delta_{33}$ is positive for countries producing petroleum (Russia, Saudi Arabia, Venezuela) while the highest negative derivative appears for China which economy is happened to be very sensitive to petroleum price. The results from the trade volume computation of $dB_c/d\delta_p$, shown in Figure 21, give rather different distribution of derivatives over countries with maximum for Sudan and minimum for the Republic of Nauru (this country has very small area and is not visible in the bottom panel of Fig. 21), while for China the balance looks to be not very sensitive to δ_{33} (in contrast to the CheiRank-PageRank method). This happens due to absence of links between nodes in the trade volume computations while the CheiRank-PageRank approach takes links into account and recover hidden trade relations between products and countries.

This absence of links in the trade volume approach becomes also evident if we consider the derivative of the partial trade balance for a given product p defined as:

$$\begin{aligned} B_{cp} &= (P_{cp}^* - P_{cp}) / \sum_p (P_{cp}^* + P_{cp}) \\ &= (P_{cp}^* - P_{cp}) / (P_c^* + P_c), \end{aligned} \quad (14)$$

so that the global country balance is $B_c = \sum_p B_{cp}$. Then the sensitivity of partial balance of a given product p in respect to a price variation of a product p' is given by the derivative $dB_{cp}/d\delta_{p'}$. The sensitivity for balance of product $p = 72$ (*72 Electrical machinery ...*) in respect to petroleum $p' = 33$ price variation δ_{33} is shown for the CheiRank-PageRank balance in Figure 22 (top panel) indicating sensitivity of trade balance of product $p = 72$ at the petroleum $p' = 33$ price variation. We see that China has a negative derivative for this partial balance. In contrast, the computations based on the trade volume (Fig. 22 bottom panel) give a rather different distribution of derivatives $dB_{cp}/d\delta_{p'}$ over countries. In the trade volume approach the derivative $dB_{cp}/d\delta_{p'}$ appears due to the renormalization of total trade volume and nonlinearity coming from the ratio of probabilities. We argue that the CheiRank-PageRank approach treats the trade relations between products and countries on a significantly more advanced level taking into account all the complexity of links in the multiproduct world trade.

Using the CheiRank-PageRank approach we determine the sensitivity of partial balance of all 61 products in respect to petroleum price variation δ_{33} for China, Russia and USA, as shown in Figure 23 (top panel). We see that the diagonal derivative $dB_{c33}/d\delta_{33}$ is positive for Russia but is negative for China and USA. Even if USA produce petroleum its sensitivity is negative due to a significant import of petroleum to USA. For non-diagonal derivatives over δ_{33} we find positive sensitivity of Russia and USA for products $p = 71, 72, 73$ while for China it is negative. Other product partial balances sensitive to petroleum are e.g. 84 *Clothing* for China for which expensive petroleum

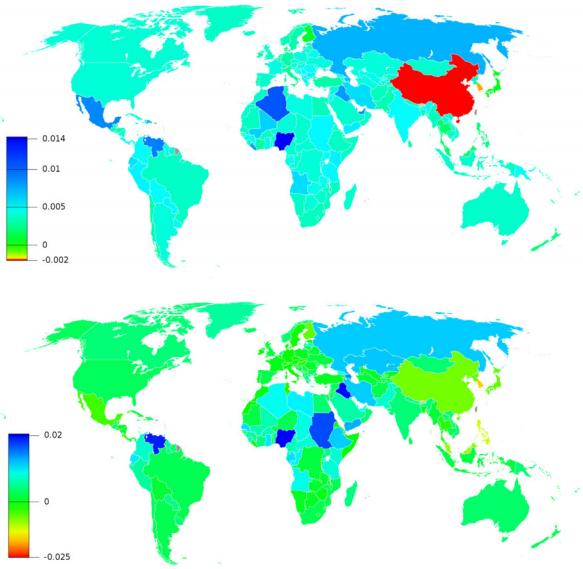


Fig. 22. Derivative of partial probability balance of product p defined as $dB_{cp}/d\delta_{33}$ over petroleum price δ_{33} for year 2008; here $B_{cp} = (P_{cp}^* - P_{cp})/(P_{cp}^* + P_c)$ and $p = 72$ (72 Electrical machinery ... from Tab. 1); the product balance of countries B_{cp} is determined from CheiRank and PageRank vectors (top-panel) and from the trade volume of Export-Import (3) (bottom-panel). Names of countries can be found at [10].

gives an increase of transportation costs; negative derivative of balance in metal products $p = 67, 68$ for Russia due to fuel price increase; positive derivative for 93 Special transact ... of USA.

The sensitivity of country balance B_c to price variation $\delta_{p'}$ for all products is shown in Figure 23 for China (middle panel) and USA (bottom panel). We find that the balance of China is very sensitive to $p' = 33, 84$ and indeed, these products play an important role in its economy with negative and positive derivatives, respectively. For USA the trade balance is also very sensitive to these two products $p' = 33, 84$ but the derivative is negative in both cases. We also present the derivative of balance without diagonal term $(d(B_c - B_{cp'})/d\delta_{p'})$ for China and USA. This quantity shows that for USA all other products give a positive derivative for $p' = 33$ but the contribution of petroleum import gives the global negative derivative of the total USA balance. In a similar way for China for $p' = 84$ all products, except the diagonal one $p' = 84$, give a negative sensitivity for balance but the diagonal contribution of $p' = 84$ gives the final positive derivative of China total balance in respect to δ_{84} .

The CheiRank-PageRank approach allows to determine cross-product sensitivity of partial trade balance computing the derivative $dB_{cp}/d\delta_{p'}$ shown in Figure 24 for China and USA. The derivatives are very different for two countries showing a structural difference of their economies. Thus for China the cross-derivative (at $p \neq p'$) are mainly negative (except a few lines around $p = 33$) but the diagonal terms $dB_{cp}/d\delta_p$ are mainly positive.

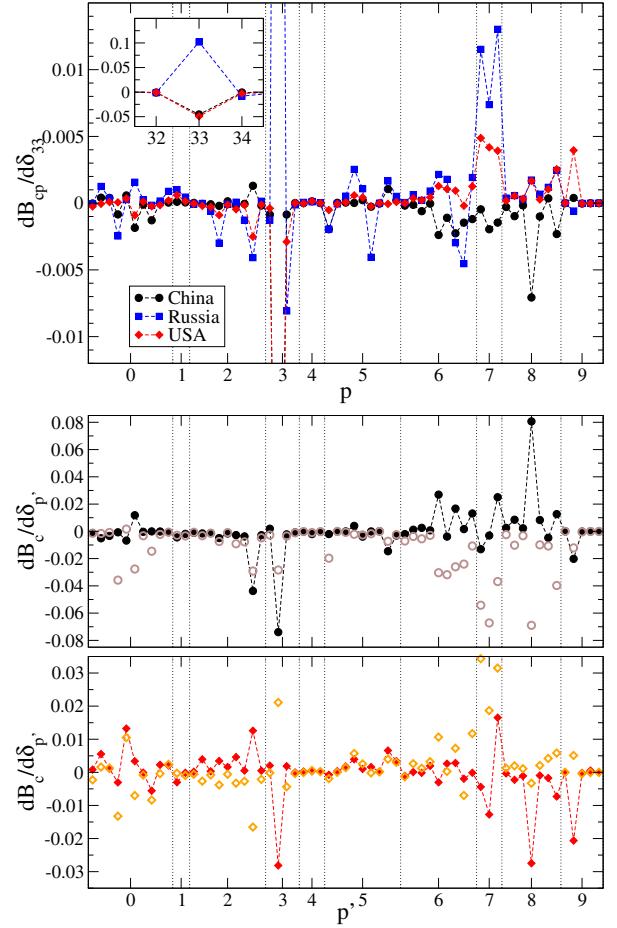


Fig. 23. Top panel: derivative $dB_{cp}/d\delta_{33}$ of partial probability balance B_{cp} of product p over petroleum price δ_{33} for year 2008 and countries: China (black circles), Russia (blue squares) and USA (red diamonds); inset panel shows the products of digit 3 including the diagonal term $p = 33$ being out of scale in the main panel; here $B_{cp} = (P_{cp}^* - P_{cp})/(P_{cp}^* + P_c)$ (14). Center (China) and bottom (USA) panels show derivative $dB_c/d\delta_{p'}$ of country total probability balance B_c over price $\delta_{p'}$ for year 2008; derivatives of balance without diagonal term ($dB_c/d\delta_{p'} - dB_{cp'}/d\delta_{p'}$) are represented by open circles and open diamonds for China and USA, respectively. The product balance of countries B_{cp} and B_c are determined from CheiRank and PageRank vectors. The vertical dotted lines mark the first digit of product index p or p' from Table 1.

In contrast, for USA the situation is almost the opposite. We attribute this to the leading role of China in export and the leading role of USA in import. However, a detailed analysis of these cross-products derivatives and correlations require further more detailed analysis. We think that the presented cross-product sensitivity plays an important role in the multiproduct trade network that are highlighted by the Google matrix analysis developed here. This analysis allows to determine efficiently the sensitivity of multiproduct trade in respect to price variations of various products.

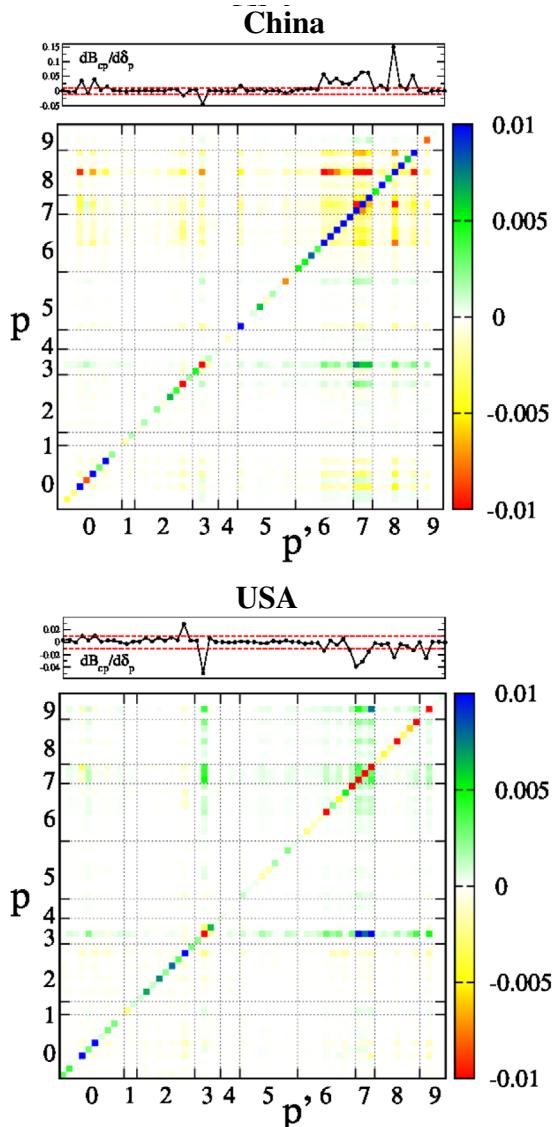


Fig. 24. China (top) and USA (bottom) examples of derivative $dB_{cp}/d\delta_{p'}$ of partial probability balance B_{cp} of product p over price $\delta_{p'}$ of product p' for year 2008. Diagonal terms, given by $dB_{cp}/d\delta_p$ vs. $p = p'$, are shown on the top panels of each example. Products p' and p are shown in x -axis and y -axis respectively (indexed as in Tab. 1), while $dB_{cp}/d\delta_{p'}$ is represented by colors with a threshold value given by -0.01 and 0.01 for negative and positive values respectively, also shown in red dashed lines on top panels with diagonal terms. Dotted lines mark the first digit of Table 1. Here B_{cp} are defined by CheiRank and PageRank probabilities.

5 Discussion

In this work we have developed the Google matrix analysis of the multiproduct world trade network. Our approach allows to treat all world countries on equal democratic grounds independently of their richness keeping the contributions of trade products proportional to their fractions in the world trade. As a result of this approach we have obtained a reliable ranking of world countries and

products for years 1962–2010. The Google analysis captures the years with crises and also shows that after averaging over all world countries some products are export oriented while others are import oriented. This feature is absent in the usual Import-Export analysis based on trade volume which gives a symmetric orientation of products after such an averaging.

The WTN matrix analysis determines the trade balance for each country not only in trade volume but also in CheiRank-PageRank probabilities which take into account multiple trade links between countries which are absent in the usual Export-Import considerations. The CheiRank-PageRank balance highlights in a clear manner the leading WTN role of new rising economies of China and other BRICS countries. This analysis also allows to determine the sensitivity of trade network to price variations of various products that opens new possibilities for analysis of cross-product price influence via network links absent in the standard Export-Import analysis.

We think that this work makes only first steps in the development of WTN matrix analysis of multiproduct world trade. Indeed, the global properties of the Google matrix of multiproduct WTN should be studied in more detail since the statistical properties of matrix elements of G , shown in Figure 25 for year 2008, are still not well understood (e.g. visible patterns present in the coarse-grained representation of G in Fig. 25).

Even if the UN COMTRADE database contains a lot of information there are still open questions if all essential economic aspects are completely captured in this database. Indeed, the COMTRADE data for trade exchange are diagonal in products since there are no interactions (trade) between products. However, this feature may be a weak point of collected data since in a real economy there is a transformation of some products into some other products (e.g. metal and plastic are transferred to cars and machinery). It is possible that additional data should be collected to take into account the existing interactions between products. There are also some other aspects of services and various other activities which are not present in the COMTRADE database and which can affect the world economy. At the same time our results show that the existing COMTRADE data allow to obtain reliable results using the Google matrix analysis: thus the ranking of countries and products are reasonable being in correspondence with results of other methods. Also sensitivity to price variations is correct from the economy view point (e.g. Fig. 22 showing a high sensitivity of China economy to petroleum price). We think that additional inter-product links will not modify significantly the results presented here but we expect that they will allow to characterize in a better way how one product is transferred to others in the result of the multiproduct world trade.

One of the important missing element of COMTRADE are financial flows between countries. Indeed, the product 93 *Special trans ...* (see Tabs. 1 and 2) partially takes into account the financial flows but it is clear that the interbank flows are not completely reported in the database. In fact the Wold Bank Web (WBW) really exists

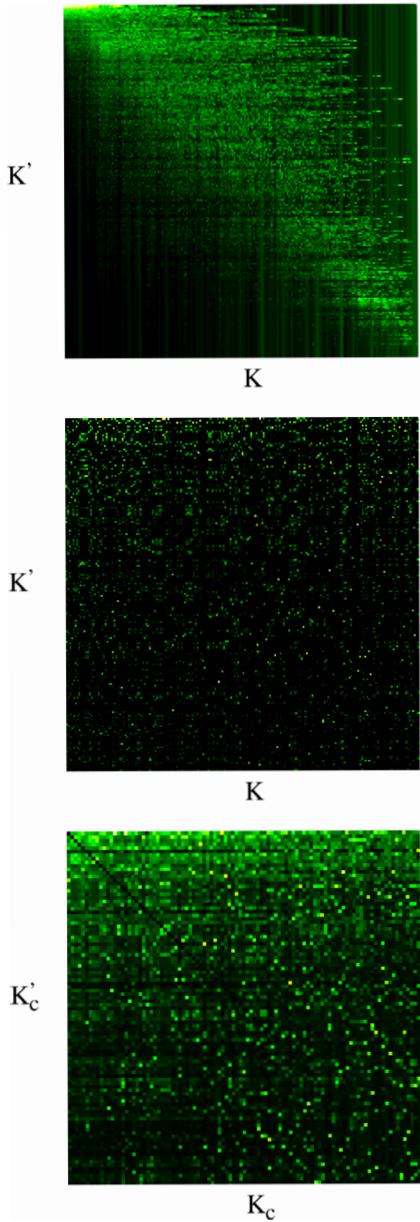


Fig. 25. Google matrix $G_{KK'}$ representation for 2008 with $\alpha = 0.5$ ordered by PageRank index K value (where $K = K' = 1$ is on top left corner). Top panel shows the whole Google matrix ($N = N_c \times N_p = 227 \times 61 = 13847$) with coarse-graining of $N \times N$ elements down to 200×200 shown cells. Center panel represents the top corner of the full Google matrix with $K, K' \leq 200$. Bottom panel shows the coarse-grained Google matrix for countries for the top 100 countries ($K_c, K'_c \leq 100$). Color changes from black at minimal matrix element to white at maximal element, $\alpha = 0.5$.

(e.g. a private person can transfer money from his bank account to another person account using SWIFT code) but the flows on the WBW remain completely hidden and not available for scientific analysis. The size of interbank networks are relatively small (e.g. the whole Federal Reserve of USA has only $N \approx 6600$ bank nodes [27] and there are only about $N \approx 2000$ bank nodes in Germany [28]).

Thus the WBW size of the whole world is about a few tens of thousands of nodes and the Google matrix analysis should be well adapted for WBW. We consider that there are many similarities between the multiproduct WTN and the WBW, where financial transfers are performed with various financial products so that the above WTN analysis should be well suited for the WBW. The network approach to the WBW flows is now at the initial development stage (see e.g. [27–29]) but hopefully the security aspects will be handled in an efficient manner opening possibilities for the Google matrix analysis of the WBW. The joint analysis of trade and financial flows between world countries would allow to reach a scientific understanding of peculiarities of such network flows and to control in an efficient way financial and petroleum crises.

The developed Google matrix analysis of multiproduct world trade allows to establish hidden dependencies between various products and countries and opens new prospects for further studies of this interesting complex system of world importance.

We thank the representatives of UN COMTRADE [1] for providing us with the friendly access to this database. This research is supported in part by the EC FET Open project “New tools and algorithms for directed network analysis” (NADINE No. 288956). We thank Barbara Meller (Deutsche Bundesbank, Zentrale) for constructive critical remarks.

References

1. United Nations Commodity Trade Statistics Database, <http://comtrade.un.org/db/>. Accessed November 2014
2. World Trade Organization, International Trade Statistics 2014, <http://www.wto.org/its2014/>. Accessed November 2014
3. P.R. Krugman, M. Obstfeld, M. Melitz, *International Economics: Theory & Policy* (Prentice Hall, New Jersey, 2011)
4. S. Dorogovtsev, *Lectures on Complex Networks* (Oxford University Press, Oxford, 2010)
5. S. Brin, L. Page, Computer Networks and ISDN Systems **30**, 107 (1998)
6. A.M. Langville, C.D. Meyer, *Google's PageRank and Beyond: the Science of Search Engine Rankings* (Princeton University Press, Princeton, 2006)
7. A.D. Chepelianskii, [arXiv:1003.5455\[cs.SE\]](https://arxiv.org/abs/1003.5455) (2010)
8. A.O. Zhirov, O.V. Zhirov, D.L. Shepelyansky, Eur. Phys. J. B **77**, 523 (2010)
9. L. Ermann, K.M. Frahm, D.L. Shepelyansky, [arXiv:1409.0428\[physics.soc-ph\]](https://arxiv.org/abs/1409.0428) (2014)
10. Web page Maps of the world, <http://www.mapsofworld.com/>. Accessed December 2014
11. L. Ermann, D.L. Shepelyansky, Acta Phys. Pol. A **120**, A158 (2011)
12. L. Ermann, D.L. Shepelyansky, Phys. Lett. A **377**, 250 (2013)
13. D. Garlaschelli, M.I. Loffredo, Physica A **355**, 138 (2005)
14. M.A. Serrano, M. Boguna, A. Vespignani, J. Econ. Interac. Coord. **2**, 111 (2007)
15. G. Fagiolo, J. Reyes, S. Schiavo, Phys. Rev. E **79**, 036115 (2009)

16. J. He, M.W. Deem, Phys. Rev. Lett. **105**, 198701 (2010)
17. G. Fagiolo, J. Reyes, S. Schiavo, J. Evol. Econ. **20**, 479 (2010)
18. M. Barigozzi, G. Fagiolo, D. Garlaschelli, Phys. Rev. E **81**, 046104 (2010)
19. T. Squartini, G. Fagiolo, D. Garlaschelli, Phys. Rev. E **84**, 046118 (2011)
20. L. De Benedictis, L. Tajoli, World Econ. **34**, 1417 (2011)
21. T. Deguchi, K. Takahashi, H. Takayasu, M. Takayasu, PLoS One **9**, e1001338 (2014)
22. C.A. Hidalgo, B. Klinger, A.-L. Barabási, R. Hausmann, Science **317**, 5837 (2007)
23. J.-P. Bouchaud, M. Potters, *Theory of Financial Risk and Derivative Pricing* (Cambridge University Press, Cambridge, 2003)
24. M.C. Munnix, R. Schaefer, T. Guhr, PLoS One **9**, e98030 (2014)
25. Web page Google matrix of multiproduct world trade, <http://www.quantware.ups-tlse.fr/QWLlib/wtnmatrix>. Accessed December 2014
26. L. Ermann, K.M. Frahm, D.L. Shepelyansky, Eur. Phys. J. B **86**, 193 (2013)
27. K. Soramäki, M.L. Bech, J. Arnold, R.J. Glass, W.E. Beyler, Physica A **379**, 317 (2007)
28. B. Craig, G. von Peter, Interbank tiering and money center bank, Discussion paper No. 12, Deutsche Bundesbank (2010)
29. R.J. Garratt, L. Mahadeva, K. Svirydzenka, Mapping systemic risk in the international banking network, Working paper No. 413, Bank of England (2011)

Google matrix of the world network of economic activities

V.Kandiah^{1,2,3}, H.Escaith^{2,3} and D.L.Shevelyansky¹

¹ Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, F-31062 Toulouse, France

² World Trade Organization, rue de Lausanne 154, CH-1211 Genève 21, Switzerland

³ Opinions are personal and do not represent WTO's position

Dated: April 14, 2015

Abstract. Using the new data from the OECD-WTO world network of economic activities we construct the Google matrix G of this directed network and perform its detailed analysis. The network contains 58 countries and 37 activity sectors for years 1995 and 2008. The construction of G , based on Markov chain transitions, treats all countries on equal democratic grounds while the contribution of activity sectors is proportional to their exchange monetary volume. The Google matrix analysis allows to obtain reliable ranking of countries and activity sectors and to determine the sensitivity of CheiRank-PageRank commercial balance of countries in respect to price variations and labor cost in various countries. We demonstrate that the developed approach takes into account multiplicity of network links with economy interactions between countries and activity sectors thus being more efficient compared to the usual export-import analysis. The spectrum and eigenstates of G are also analyzed being related to specific activity communities of countries.

PACS. 89.75.Fb Structures and organization in complex systems – 89.65.Gh Econophysics – 89.75.Hc Networks and genealogical trees – 89.20.Hh World Wide Web, Internet

1 Introduction

The recent reports of the Organisation for Economic Co-operation and Development (OECD) [1] and of the World Trade Organization (WTO) [2] demonstrate all the complexity of global manufacturing activities, exchange and trade in the modern world. This complexity is rapidly growing with time and now it becomes clear that traditional statistics are increasingly unable to provide all the necessary information. Applying modern mathematical tools and methods to new data sets can allow to understand the hidden trends of the world economic activities. Thus the matrix tools for analysis of Input-Out transactions are broadly used in economy starting from the fundamental works of Leontief [3,4] with their more recent developments described in [5]. In the last decade the development of modern society generated enormous communication and social networks including the World Wide Web (WWW), Wikipedia, Twitter and other directed networks (see e.g. [6]). It has been found that the concept of Markov chains provides a very useful and powerful mathematical approach for analysis of such networks. Thus the PageRank algorithm, developed by Brin and Page in 1998 [7] for the WWW information retrieval, became at the mathematical foundation of the Google search engine (see e.g. [8]). This algorithm constructs the Google matrix G of Markov chain transitions between network nodes and allows to rank billions of web pages of the WWW. The

spectral and other properties of the Google matrix are analyzed in [9]. The historical overviews of the development of Google matrix methods and their links with the works of Leontief are given in [10,11].

The obtained results demonstrate the efficiency of the Google matrix analysis not only for the WWW but also for various types of directed networks [9]. One of such examples is the World Trade Network (WTN) with multiproduct exchange between the world countries. The data of trade flows are available at the United Nations (UN) COMTRADE database [12] for more than 50 years. The results presented in [13,14] for the WTN show that the Google matrix analysis is well adapted to the ranking of world countries and trade products and to determination of the sensitivity of trade to price variations of various products. The new element of such an approach is a democratic treatment of world countries independently of their richness being different from the usual Import and Export ranking. At the same time the contributions of various products are considered being proportional to their trade volume contribution in the exchange flows.

Here we use the Google matrix analysis developed for the multiproduct WTN [14] showing that it can be directly used for the World Network of Economic Activities (WNEA) constructed from the OECD-WTO trade in value-added database. In a certain sense activities (or sectors) are correlated to products in the WTN. However, for the WTN there is exchange between countries but there

is no exchange between industries and commodities. Thus in [14] it was argued that certain economical features are not captured by the COMTRADE database since in real economy the traders are industries, not countries; in particular certain products are transferred to each other (e.g. metal and plastic are used for production of cars). In contrast to that, the OECD-WTO WNEA incorporates the transitions between activity sectors thus representing the economic reality of world activities in a more correct manner.

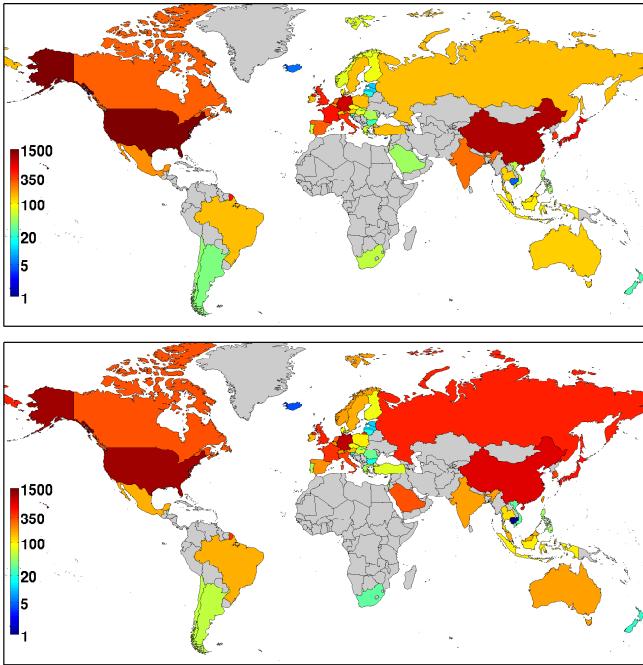


Fig. 1. World map of countries with color showing country import (top panel) and export (bottom panel) with economic activity (trade) volume expressed in billions of USD and given by numbers at color bars; the gray color marks countries attributed to the ROW group (rest of the world) with exchange values 733 (Import) and 1018 (Export) in billions of USD. The data are shown for year 2008 with $N_c = 57 + 1$ countries (with ROW) for the economic activities in all $N_s = 37$ sectors. Country names can be found in Table 1 and in the world map of countries [22].

We note that there has been a number of other investigations of the WTN reported in [15–21]. However, in this work we have the new important elements, introduced in [13, 14]: the analysis of PageRank and CheiRank probabilities corresponding to direct and inverted network flows and related to Import and Export; democratic treatment of countries combined with the contributions of sectors (or products) being proportional to their commercial exchange fractions. We point that the OECD-WTO TiVA database of economic activities between world countries and activity sectors has been created very recently (2013) and thus this work represents the first Google matrix analysis of these data. We stress that the usual Import-Export

ranking of commercial flows, shown in Fig. 1, is not able to take into account all the complexity of chains of links between various countries and various activity sectors. In contrast to that the approach developed here takes all of them into account due to the powerful method based on the Google matrix.

2 Methods and data description

Here we describe the data available for the OECD-WTO TiVA network and the mathematical methods used for the analysis of this network. The list of $N_c = 58$ countries (57 plus 1 for the Rest Of the World ROW) is given in Table 1 with their flags. Following [13] we use for countries ISO 3166-1 alpha-3 code available at Wikipedia. The list of sectors with their names is given in Table 2. The fractions of sectors in the exchange volume are given in Table 3 for years 1995, 2008.

2.1 Google matrix construction for the OECD-WTO WNEA

We use the OECD-WTO TiVA database released in May 2013 which covers years 1995, 2000, 2005, 2008, 2009 with the main emphasis for years 1995 and 2008 (2009 data are affected by the global crisis and may not be representative). The network considers $N_c = 58$ world countries given in Table 1. In fact, there are 57 countries and the rest of the world, which includes the remaining countries of the world forming one group called ROW. There are also $N_s = 37$ sectors of economic activities given in Table 2. The sectors are classified according to the International Standard Industrial Classification of All Economic Activities (ISIC) Rev.3 [23]. Here we present results for all 37 sectors of Table 2, noting that the sectors $s = 1, 2, \dots, 20$ represent production activities while $s = 21, \dots, 37$ represent service activities. The transections between service sectors are hard to extract and the future improvements of this part of TiVA database are desirable.

For a given year, the TiVA data extend OECD Input/Out tables of economic activity expressed in terms of USD for a given year. From these data we construct the matrix $M_{cc',ss'}$ of money transfer between nodes expressed in USD:

$$M_{cc',ss'} = \text{transfer from country } c', \text{ sector } s' \text{ to } c, s \quad (1)$$

Here the country indexes are $c, c' = 1, \dots, N_c$ and activity sector indexes are $s, s' = 1, \dots, N_s$ with $N_c = 58$ and $N_s = 37$. The whole matrix size is $N = N_c \times N_s = 2146$. Here each node represents a pair of country and activity sector, a link gives a transfer from a sector of one country to another sector of another country. We construct the matrix $M_{cc',ss'}$ from the TiVA Input/Output tables using the transposed representation so that the volume of products or sectors flows in a column from line to line. In the construction of $M_{cc',ss'}$ we exclude exchanges inside a given country in order to highlight the trade exchange

flows between countries (elements inside country are zeros).

The ISIC Rev.3 classification of sectors have a significant correlation with the UN Standard International Trade Classification (SITC) Rev. 1 of products used in [14]. There is a clear relationship on the production side between ISIC sectors and products of the world exports (but not at import level: if all agricultural exports are produced by the agricultural sector, agricultural products will be imported by manufacturing industries such as food processing of textile and clothing). There is also another important difference: the transfer matrix from COMTRADE is diagonal in products [14] (thus there is no transfer from product to product), while for the TIVA data there are transitions from one sector to another sector and thus the matrix of nominal values, in current prices, (1) is not diagonal in s, s' .

For convenience of future notations we also define the value of imports V_{cs} and exports V_{cs}^* for a given country c and sector s as

$$V_{cs} = \sum_{c',s'} M_{cc',ss'}, \quad V_{cs}^* = \sum_{c',s'} M_{c'c,s's}. \quad (2)$$

The import $V_c = \sum_s V_{cs}$ and export $V_c^* = \sum_s V_{cs}^*$ values for countries c are shown on the world map of countries in Fig. 1 for year 2008. We note that often one uses the notion of volume of export or import (see, e.g. [14]) but from the economic view point it more correct to speak about value of export or import.

In order to compare later with the PageRank and CheiRank probabilities we define exchange value ranks in the whole matrix space of dimension $N = N_c \times N_s$. Thus the ImportRank (\hat{P}) and ExportRank (\hat{P}^*) probabilities are given by the normalized import and export values

$$\hat{P}_i = V_{cs}/V, \quad \hat{P}_i^* = V_{cs}^*/V, \quad (3)$$

where $i = s + (c - 1)N_s$, $i = 1, \dots, N$ and the total exchange value is $V = \sum_{c,c',s,s'} M_{cc',ss'} = \sum_{c,s} V_{cs} = \sum_{cs} V_{cs}^*$.

The Google matrices G and G^* are defined as $N \times N$ real matrices with non-negative elements:

$$G_{ij} = \alpha S_{ij} + (1 - \alpha) v_i e_j, \quad G^*_{ij} = \alpha S^*_{ij} + (1 - \alpha) v_i^* e_j, \quad (4)$$

where $N = N_c \times N_s$, $\alpha \in (0, 1]$ is the damping factor ($0 < \alpha < 1$), e_j is the row vector of unit elements ($e_j = 1$), and v_i is a positive column vector called a *personalization vector* with $\sum_i v_i = 1$ [8, 14]. We note that the usual Google matrix corresponds to a personalization vector $v_i = e_i/N$ with $e_i = 1$. In this work, following [13, 14], we fix $\alpha = 0.5$ noting that a variation of α in a range (0.5, 0.9) does not significantly affect the probability distributions of PageRank and CheiRank vectors [8, 9, 13]. The choice of the personalization vector is specified below. Following [14] we call this approach the Google Personalized Vector Method (GPVM).

The matrices S and S^* are built from money matrices $M_{cc',ss'}$ as

$$S_{i,i'} = \begin{cases} M_{cc',ss'}/V_{c's'} & \text{if } V_{c's'} \neq 0 \\ 1/N & \text{if } V_{c's'} = 0 \end{cases} \\ S^*_{i,i'} = \begin{cases} M_{c'c,s's}/V_{c's'}^* & \text{if } V_{c's'}^* \neq 0 \\ 1/N & \text{if } V_{c's'}^* = 0 \end{cases} \quad (5)$$

where $c, c' = 1, \dots, N_c$; $s, s' = 1, \dots, N_s$; $i = s + (c - 1)N_s$; $i' = s' + (c' - 1)N_s$; and therefore $i, i' = 1, \dots, N$. Here $V_{c's'} = \sum_{cs} M_{cc',ss'}$. The sum of elements of each column of S and S^* is normalized to unity and hence the matrices G, G^*, S, S^* belong to the class of Google matrices and Markov chains. Thus S, G look at the import perspective and S^*, G^* at the export side of transactions.

PageRank and CheiRank (P and P^*) are the right eigenvectors of G and G^* matrices respectively at eigenvalue $\lambda = 1$. The equation for right eigenvectors have the form

$$\sum_j G_{ij} \psi_j = \lambda \psi_i, \quad \sum_j G^*_{ij} \psi^*_j = \lambda \psi^*_i. \quad (6)$$

For the eigenstate at $\lambda = 1$ we use the notation $P_i = \psi_i$, $P^*_i = \psi^*_i$ with the normalization $\sum_i P_i = \sum_i P^*_i = 1$. For other eigenstates we use the normalization $\sum_i |\psi_i|^2 = \sum_i |\psi^*_i|^2 = 1$. The eigenvalues and eigenstates of G, G^* are obtained by a direct numerical diagonalization using the standard numerical packages.

2.2 PageRank and CheiRank vectors from GPVM

The components of P_i, P^*_i are positive. In the WWW context they have a meaning of probabilities to find a random surfer on a given WWW node in the limit of large number of surfer jumps over network links [8]. In the WNEA context nodes can be viewed as markets with a random trader transitions between them. We will use in the following notation of network nodes. We define the PageRank K and CheiRank K^* indexes ordering probabilities P and P^* in a decreasing order as $P(K) \geq P(K+1)$ and $P^*(K) \geq P^*(K^*+1)$ with $K, K^* = 1, \dots, N$.

We note that the pair of PageRank and CheiRank vectors is very natural for economy and trade networks corresponding to Import and Export flows. For the directed networks the statistical properties of the pair of such ranking vectors have been introduced and studied in [24, 25, 13].

We compute the reduced PageRank and CheiRank probabilities of countries tracing probabilities over all sectors and getting $P_c = \sum_s P_{cs} = \sum_s P(s + (c - 1)N_s)$ and $P_c^* = \sum_s P_{cs}^* = \sum_s P^*(s + (c - 1)N_s)$ with the corresponding K_c and K_c^* indexes. In a similar way we obtain the reduced PageRank and CheiRank probabilities for sectors tracing over all countries and getting $P_s = \sum_c P(s + (c - 1)N_s) = \sum_c P_{cs}$ and $P_s^* = \sum_c P^*(s + (c - 1)N_s) = \sum_c P_{cs}^*$ with their corresponding sector indexes K_s and K_s^* . A similar procedure has been used for the multiproduct WTN data [14].

In summary we have $K_s, K_s^* = 1, \dots, N_s$ and $K_c, K_c^* = 1, \dots, N_c$. A similar definition of ranks from import and export exchange value can be done in a straightforward way via probabilities $\hat{P}_s, \hat{P}_s^*, \hat{P}_c, \hat{P}_c^*, \hat{P}_{cs}, \hat{P}_{cs}^*$ and corresponding indexes $\hat{K}_s, \hat{K}_s^*, \hat{K}_c, \hat{K}_c^*, \hat{K}, \hat{K}^*$.

To compute the PageRank and CheiRank probabilities from G and G^* , keeping a “democratic”, or equal, treatment of countries (independently of their richness) and at the same time keeping the proportionality of activity sectors to their exchange value, we use the Google Personalized Vector Method (GPVM) developed in [14] with a personalized vector v_i in (4). At the first iteration of Google matrix we take into account the relative product value per country using the following personalization vectors for G and G^* :

$$v_i = \frac{V_{cs}}{N_c \sum_{s'} V_{cs'}}, \quad v_i^* = \frac{V_{cs}^*}{N_c \sum_{s'} V_{cs'}^*}, \quad (7)$$

using the definitions (2) and the relation $i = s + (c-1)N_s$. This personalized vector depends both on sector and country indexes. As for the multiproduct WTN in [14] we define the second iteration vector being proportional to the reduced PageRank and CheiRank vectors in sectors, obtained from the GPVM Google matrix of the first iteration:

$$v'(i) = \frac{P_s}{N_c}, \quad v'^*(i) = \frac{P_s^*}{N_c}. \quad (8)$$

In this way we keep democracy in countries but keep contribution of sectors proportional to their exchange value. This second iteration personalized vectors are used in the following computations and operations with G and G^* giving us the PageRank and CheiRank vectors. This procedure with two iterations forms our GPVM approach. The difference between results obtained from the first and second iterations is not very large (see Figs. 2, 3), but the personalized vector for the second iteration gives a reduction of fluctuations. In all Figures after Fig. 3 we show the GPVM results after the second iteration.

As for the WTN it is convenient to analyze the distribution of nodes on the PageRank-CheiRank plane (K, K^*) . In addition to two ranking indexes K, K^* we use also 2DRank index K_2 which describes the combined contribution of two ranks as described in [25]. The ranking list $K_2(i)$ is constructed by increasing $K \rightarrow K+1$ and increasing 2DRank index $K_2(i)$ by one if a new entry is present in the list of first $K^* < K$ entries of CheiRank, then the one unit step is done in K^* and K_2 is increased by one if the new entry is present in the list of first $K < K^*$ entries of CheiRank. More formally, 2DRank $K_2(i)$ gives the ordering of the sequence of nodes, that appear inside the squares $[1, 1; K = k, K^* = k; \dots]$ when one runs progressively from $k = 1$ to N . Additionally, we analyze the distribution of nodes for reduced indexes $(K_c, K_c^*), (K_s, K_s^*)$.

The localization properties of eigenstates of G, G^* are characterized by the inverse participation ratio (IPR) defined as $\xi = (\sum_i |\psi_i|^2)^2 / \sum_i |\psi_i|^4$. This quantity determines an effective number of nodes contributing to a formation of a given eigenstate (see details in [9]).

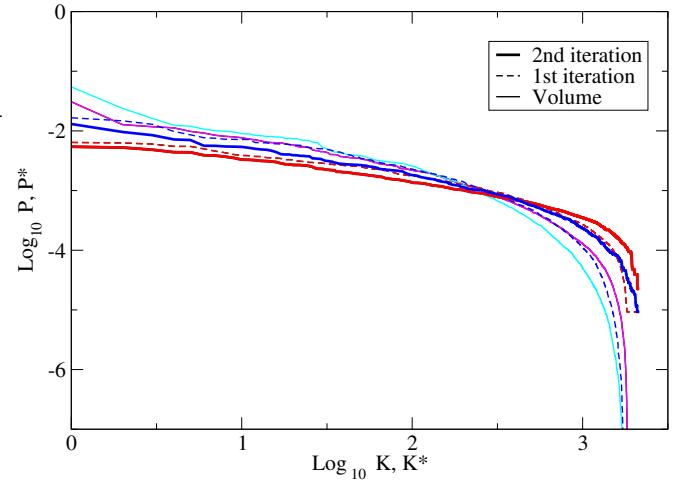


Fig. 2. Dependence of probabilities of PageRank $P(K)$, CheiRank $P^*(K^*)$, ImportRank $\hat{P}(\hat{K})$ and ExportRank $\hat{P}^*(\hat{K}^*)$ on their indexes in logarithmic scale for WNEA (or OECD-WTO TiVA network) in 2008 with $\alpha = 0.5$, $N_c = 58$, $N_s = 37$, $N = N_c \times N_s = 2146$. Here the results for the GPVM after the first and second iterations are shown for PageRank (CheiRank) in red (blue) with dashed and solid curves respectively. Probabilities for ImportRank and ExportRank from exchange value are shown by magenta and cyan thin curves respectively.

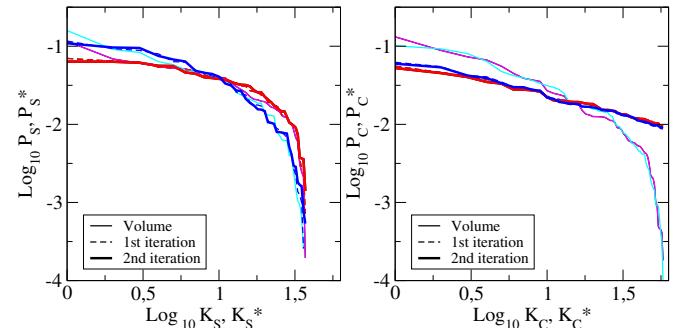


Fig. 3. Probability distributions of PageRank and CheiRank for sectors $P_s(K_s), P_s^*(K_s^*)$ (left panel) and countries $P_c(K_c), P_c^*(K_c^*)$ (right panel) in logarithmic scale for WNEA (or OECD-WTO TiVA network) from Fig.2. Here the results for the first and second GPVM iterations are shown by red (blue) curves for PageRank (CheiRank) with dashed and solid curves respectively (with a strong overlap of curves). The probabilities from the exchange value ranking are shown by thin magenta and cyan lines for ImportRank and ExportRank respectively.

2.3 Correlators of PageRank and CheiRank vectors

As in previous works [24, 25, 13] we consider the correlator of PageRank and CheiRank vectors:

$$\kappa = N \sum_{i=1}^N P(i)P^*(i) - 1. \quad (9)$$

The typical values of κ are given in [9] for various networks.

For the global PageRank and CheiRank probabilities the sector-sector correlator matrix is defined as:

$$\kappa_{ss'} = N_c \sum_{c=1}^{N_c} \left[\frac{P(s + (c-1)N_s)P^*(s' + (c-1)N_s)}{\sum_{c'} P(s + (c'-1)N_s) \sum_{c''} P^*(s' + (c''-1)N_s)} \right] - 1 \quad (10)$$

Then the correlator for a given sector is obtained from (10) as:

$$\kappa_s = \kappa_{ss'} \delta_{s,s'}, \quad (11)$$

where $\delta_{s,s'}$ is the Kronecker delta.

We also use the correlators obtained from the probabilities traced over sectors ($P_c = \sum_s P_{sc}$) and over countries ($P_s = \sum_c P_{sc}$) which are defined as

$$\kappa(c) = N_c \sum_{c=1}^{N_c} P_c P_c^* - 1, \quad \kappa(s) = N_s \sum_{s=1}^{N_s} P_s P_s^* - 1. \quad (12)$$

In the above equations (9)-(12) the correlators are computed for PageRank and CheiRank probabilities. We can also compute the same correlators using probabilities from the exchange value in ImportRank \hat{P} and ExportRank \hat{P}^* defined by (3).

The obtained results are presented in the next Section and at the web site [26].

3 Results

We apply the GPVM approach to the data sets of OECD-WTO TiVA of WNEA and present the obtained results below.

3.1 PageRank and CheiRank probabilities

The dependence of probabilities of PageRank $P(K)$ and CheiRank $P^*(K^*)$ vectors on their indexes K, K^* are shown in Fig. 2 for a selected year 2008. The results can be approximately described by an algebraic dependence $P \propto 1/K^\beta$, $P^* \propto 1/K^{*\beta}$ with the fit exponent value $\beta = 0.385 \pm 0.014$ for PageRank and $\beta = 0.486 \pm 0.02$ for CheiRank for $K, K^* \leq 10^3$. In contrast to WWW and Wikipedia networks (see e.g. [9]) there is no significant difference of β between two ranks that can be attributed to an intrinsic property of economy networks to keep economy balance of commercial exchange. The probability variation is reduced for the Google ranking compared to the value ranking. This results from a “democratic”, or equal grounds ranking of countries used in the Google matrix analysis. The obtained data also show that the variation of probabilities for 1st and 2nd GPVM iterations are not very large that demonstrates the convergence of this approach.

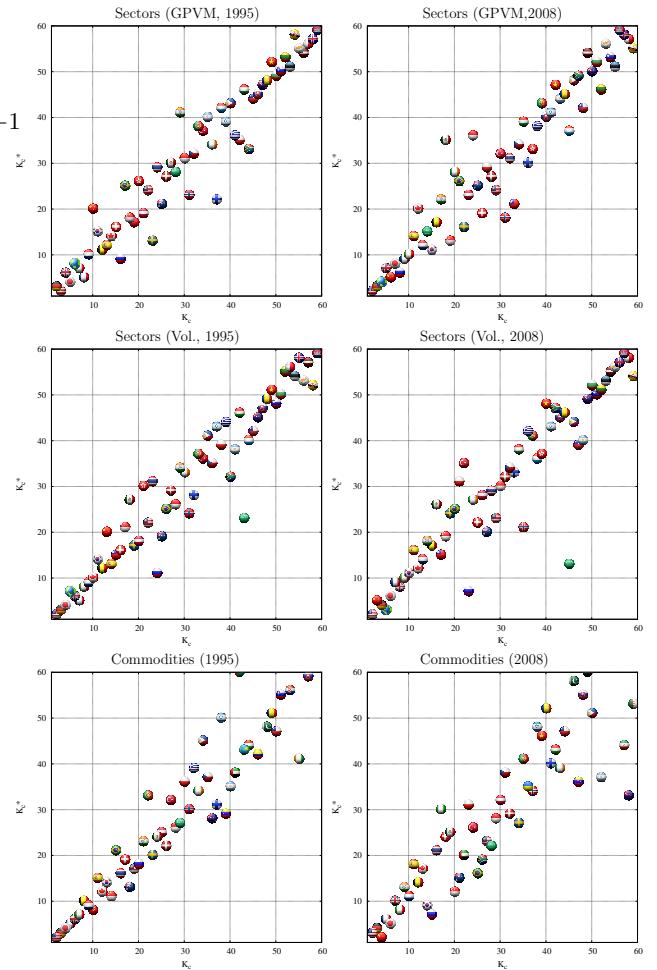


Fig. 4. Country positions on PageRank-CheiRank plane (K_c, K_c^*) obtained for the WNEA by the GPVM analysis (top panels), ImportRank-ExportRank of exchange value (middle panels), and PageRank-CheiRank plane of WTN ranking of trade in *all commodities* from [13] (bottom panels) shown for $K_c, K_c^* \leq 60$. Left (right) panels show year 1995 (2008).

3.2 Ranking of countries and sectors

After tracing the probabilities $P(K), P^*(K^*)$ over sectors we obtain the distribution of world countries on the PageRank-CheiRank plane (K_c, K_c^*) presented in Fig. 4 for WNEA in years 1995, 2008. In the same figure we present the rank distributions obtained from ImportRank-ExportRank probabilities of exchange value and the results obtained in [13] for the WTN with *all commodities*. For the GPVM data we see the global features already discussed in [13]: the countries are distributed in a vicinity of diagonal $K_c = K_c^*$ since for each country the size of imports is correlated with the size of exports, even if trade is never exactly balanced and some countries can sustain significant trade surplus or deficit. The top 20 list of top K_2 countries recover 13 of 19 countries of G20 major world economies (EU is the number 20) thus obtaining 68% of the whole list. This is close to the percent obtained in

[13] for trade in *all commodities*. The Google ranking for WNEA and WTN (top and bottom panels in Fig. 4) gives different positions for specific countries (e.g. Russia improves its position for WNEA with the opposite trend for China) but the global features of distributions of WNEA and WTN remain similar corresponding to the same economical forces.

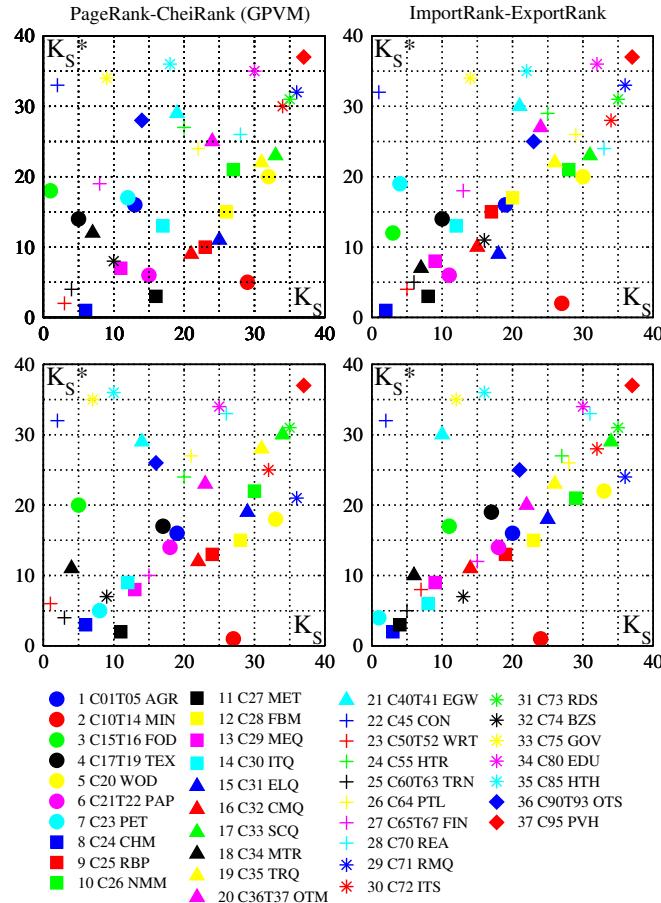


Fig. 5. Two dimensional ranking of sectors on the (K_s, K_s^*) plane using the GPVM approach for PageRank and CheiRank (left panels) and ImportRank-ExportRank (right panels). Each sector is represented by its specific combination of color and symbol. The list of all 37 sectors are given in Table 2. Top panels show the case for the year 1995 and bottom panels for the year 2008.

After tracing over countries we obtain the PageRank-CheiRank plane of activity sectors shown in Fig. 5. We see that some sectors are export oriented (e.g. $s = 2$ *C10T14 Mining* at $K_s^* = 1$ in 2008) others are import oriented (e.g. $s = 23$ *C50T52 World Retail and Trade of motors etc.* at $K_s = 1$ in 2008). The ImportRanking gives a rather different import leader $s = 7$ *C23 Manufacture of coke, refined petroleum products etc.* with $K_s = 1$ in 2008. Thus the Google ranking highlights highly connected network nodes while Import-Export gives preference to high value neglecting existing network relations between vari-

ous countries and activity sectors. We can also order sectors by 2DRank index K_2 getting for PageRank-CheiRank top sectors $s = 25, 23, 8$ at $K_2 = 1, 2, 3$ while Import-Export gives $s = 8, 11, 14$ for top K_2 values in 2008 (more data are given at [26]). We note that $s = 25$ corresponds to Transport which has many network connections thus taking the top K_2 position. We note that asymmetry of ranking of products has been discussed in [14] for COM-TRADE data, however, the comparison with these data is not so simple since the correspondence between products and activity sectors is not straightforward. Of course, for the WNEA the asymmetry of sector ranking exists even for Export-Import ranking, in a drastic difference from the WTN, since there are interactions between activity sectors.

The global ranks of top 20 countries and their activities are given in Table 4 for 2008. The top 3 places of PageRank $K = 1, 2, 3$ are taken by Germany (*Manufacture of motors etc.* $s = 18$), USA (*Public administration and defence* $s = 33$), ROW (also $s = 33$). Thus imports of arms and weapons play a very important role. In contrast for ImportRank $K = 1, 2, 3$ we find rather different results with USA (*petroleum* $s = 7$), Japan (also $s = 7$), and only then USA ($s = 33$). For CheiRank $K^* = 1, 2, 3$ we find ROW, Russia, Saudi Arabia ($s = 2$ *C10T14 Mining*) while for ExportRank we have ROW, Saudi Arabia, Russia ($s = 2$ *C10T14 Mining*) respectively. Thus Russia goes ahead of Saudi Arabia due to a broad network of activity and trade connections (a similar effect has been found in [13, 14] for trade in petroleum). The top 3 positions of 2DRank $K_2 = 1, 2, 3$ are taken by Germany ($s = 8$ *Manufacture of chemicals etc.*), USA ($s = 27$ *Finance etc.*), Germany ($s = 13$ *Manufacture of machinery etc.*).

We can fix a certain activity sector s and then consider local ranking of countries in (K_c, K_c^*) plane. Three examples are shown in Fig. 6 for $s = 21$ (*Electricity, gas, water*), 28 (*Real estate activity*), 1 (*Agriculture*). The comparison of Google ranking (left column) with value Import-Export ranking (right column) shows importance of network connections highlighted by the GPVM, thus Russia moves from $K_c^* = 4$ on right panel to $K_c^* = 2$ on left panel for $s = 21$ due to its broad links with Europe and Asia. For $s = 1$ case in bottom panels of Fig. 6 we find that the Import-Export ranking distribution is more close to diagonal comparing to the PageRank-CheiRank case that we attribute to effect of indirect links present in the later case.

The distribution of nodes on the global (K, K^*) plane is shown in Fig. 7 for Google ranking (left panel) and Import-Export ranking (right panel) in 2008. The majority of countries are shown by gray squares while 6 selected countries are marked by colors. The comparison of two panels show that in the Google ranking the positions of USA are improved (more black symbols at top K_2 positions) while for China the positions (green symbols) are weakened. We attribute this to a broader network connections of USA in important activity sectors world wide (e.g. military activities and defense).

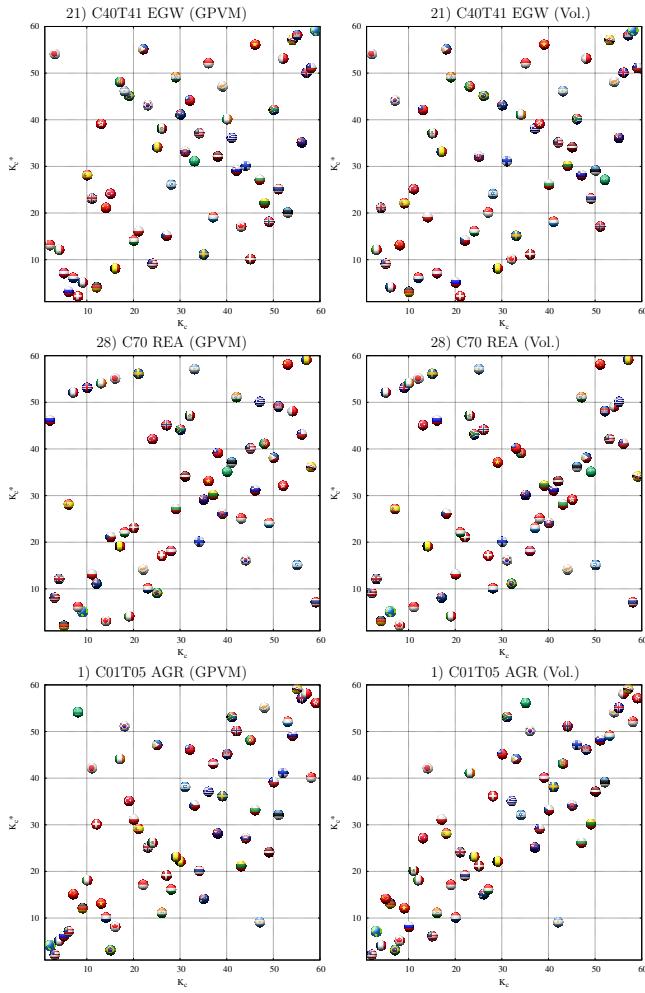


Fig. 6. Left column panels show results of the GPVM data for country positions on PageRank-CheiRank plane of local rank values K_c, K_c^* ordered by (K_{cs}, K_{*cs}) for specific sectors with $s = 21$ (top), $s = 28$ (center) and $s = 1$ (bottom). Right column panels show the ImportRank-ExportRank planes respectively for comparison. Data are given for year 2008. Each country is shown by its own flag as in Fig 4.

3.3 Correlation properties of PageRank and CheiRank

The directed networks can be characterized by the correlator κ of PageRank and CheiRank vectors. For various networks the properties of κ are reported in [24,9]. There are directed networks with small or even slightly negative values of κ , e.g. Linux Kernel or Physical Review citation networks, or with $\kappa \sim 4$ for Wikipedia networks and even larger values $\kappa \approx 116$ for the Twitter network.

The correlators of WNEA for various sectors are shown in Fig. 7. Almost all correlators κ_s are positive being distributed in a range $(0, 1)$. A small negative value appears only for $s = 37$ (*Private households etc.*) corresponding to anti-correlation between buyers and sellers. The largest correlator κ_s is for $s = 29$ (*Renting of machinery etc.*) shows that sales of machinery correlates with their purchases probably because components are needed to pro-

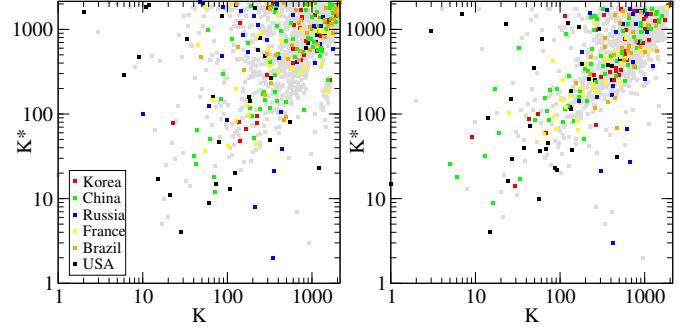


Fig. 7. Global plane of rank indexes (K, K^*) for PageRank-CheiRank (left panel) and ImportRank-ExportRank (right panel) for $N = 2146$ nodes in year 2008. Each country and sector pair is represented by a gray square. Some countries are highlighted in colors : USA in black, South Korea in red, China (and Taiwan) in green, Russia in blue, France in yellow and Brazil in orange.

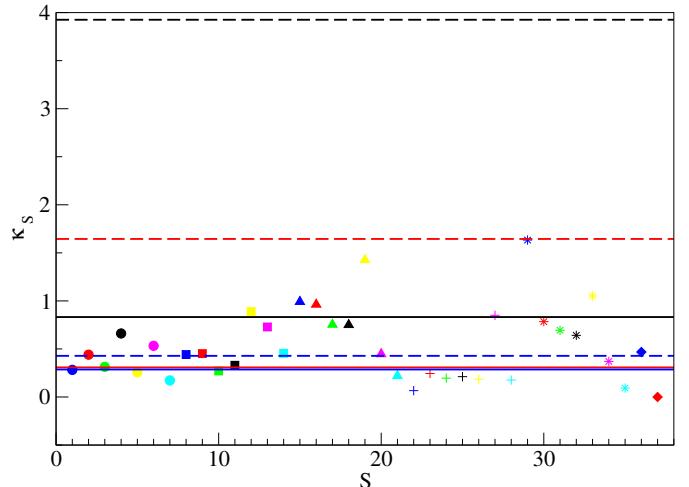


Fig. 8. PageRank-CheiRank correlators κ_s from the GPVM (see (10), (11)) are shown as a function of the sector index s with the corresponding symbol from Fig.5. PageRank-CheiRank and ImportRank-ExportRank correlators are shown by solid and dashed lines respectively, where the global correlator κ (9) is shown in black, the correlator for countries $\kappa(c)$ (12) is shown by red lines, the correlators for sectors $\kappa(s)$ (12) is shown by blue lines. Here sector index s is counted in order of appearance in Table 2. The data are given for year 2008 with $N_s = 37$, $N_c = 58$, $N = 2146$.

duce machines produced by firms in the same industrial sectors.

The matrix of correlators between sectors s, s' is shown in Fig. 8 for years 1995, 2008. It is interesting to see a significant shift of line of maximal correlators located in 1995 at $s' = 28$ (*Real estate activities*) to $s = 29$ (*Renting of machinery etc.*) in 2008. We also see that there are less correlations between sectors in 2008 compared to 1995. A further more detailed analysis of correlations would bring a better understanding of hidden inter-relations between various sectors of economic activity.

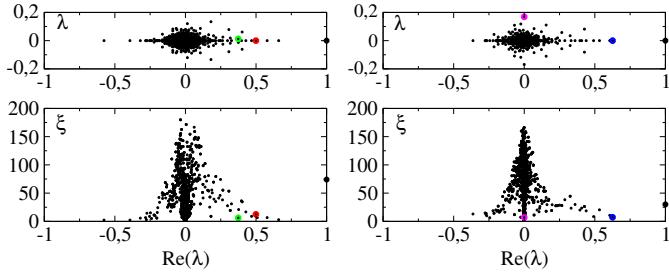


Fig. 9. Top panels: Spectrum of Google matrices G (left) and G^* (right) represented in the complex plane of λ . The data are for year 2008 with $\alpha = 1$, $N = 2146$, $N_c = 58$, $N_s = 37$. Four eigenvalues marked by colored circles are used for illustration of eigenstates in Fig. 10 and Table 5. Bottom panels: Inverse participation ratio (IPR) ξ of all eigenstates of G (left) and G^* (right) as a function of the real part of the corresponding eigenvalue λ from the spectrum above.

3.4 Spectrum and eigenstates of WNEA Google matrix

The results obtained for the Wikipedia network [29] and the multiproduct WTN [14] demonstrated that the eigenvectors of G and G^* with large eigenvalue modulus $|\lambda|$ select certain specific communities. Thus it is interesting to analyze the properties of eigenvalues for the WNEA. At $\alpha = 1$ the gap between $\lambda = 1$ and other eigenvalues characterize the rate of system relaxation to the equilibrium stationary PageRank state (for G). The presence of small gap indicates that the mixing and relaxation in the system are developed only after many iterations of G matrix (see more discussion in [9]).

The matrix size of WNEA is relatively small and the whole spectrum λ of G, G^* can be determined by direct matrix diagonalization. The spectrum is shown in top panels of Fig. 9. It is characterized by a significant gap between $\lambda = 1$ and other eigenvalues with $|\lambda| < 0.7$ at $\alpha = 1$. We attribute this to a large number of inter-connected links between matrix nodes (countries and sectors) which is usually responsible for appearance of the spectral gap (see [27], where the gap increases with the increase of number of random links per node). We also note that the maximal value of $|Im\lambda| < 0.2$ is relatively small due to presence of links going in direct and inverse directions between nodes. These features show that the relaxation processes to the steady-state PageRank vector are relatively rapid on the WNEA. Indeed, the relaxation is governed by the exponent $\exp(-\Delta\lambda t)$ where $\Delta\lambda \approx 0.25$ the gap for for WNEA in Fig. 9 and t is number of iterations of G .

The properties of eigenstates are characterized by the IPR ξ shown in bottom panels of Fig. 9. We find that the main part of states have $\xi \ll N$ so that they occupy only a small fraction of nodes corresponding to localized states (see discussion about the Anderson localization of Google matrix eigenstates in [9,28]).

The dependence of amplitudes $|\psi_i|$ of a few eigenstates, ordered by a local rank index K_i corresponding to a monotonic amplitude decrease, are shown in Fig. 10. The names of top 10 nodes of these eigenstates are given in Table 5. The red curve in Fig. 10 selects mainly the sector

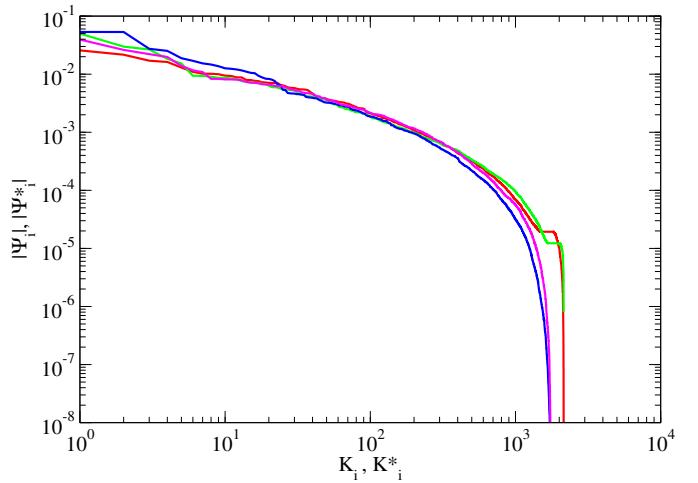


Fig. 10. Eigenstates amplitudes $|\psi_i|$ ordered by its own decreasing amplitude order with local rank index K_i for 4 different eigenvalues of Fig. 9 (states are normalized as $\sum_i |\psi_i| = 1$). The four examples are $\lambda = 0.4993$ (red), $\lambda = 0.3746 + 0.0126i$ (green), $\lambda = 0.6256$ (blue) and $\lambda = -0.0001 + 0.1687i$ (magenta). Node names (country, sector) for top ten largest amplitudes of these eigenvectors are shown in Table 5.

$s = 4$ (*Manufacture of textiles etc.*) with close links between China, Italy, USA and ROW; the green one selects $s = 18$ (*Manufacture of motor vehicles etc.*) with close links between Argentina, Brasil, Japan and Germany; the blue state corresponds to $s = 16$ (*Manufacture of radio, television and communication equipment and apparatus*) in the Asian region (China, Korea, Chinese Taipei, Singapore, Malaysia); the magenta state represents sector $s = 2$ (*Mining etc.*) with related countries like Russia, Saudi Arabia, ROW, Norway. These results coincide with the previous observations for Wikipedia-type network [29] that the eigenstates of G and G^* select specific communities of the network nodes. Similar properties of eigenstates of G of the multiproduct WTN have been found in [14].

3.5 Sensitivity to price variations

The ranking of WNEA nodes provides interesting and important information. In addition, the established matrix structure of G, G^* of WNEA also allows to study the sensitivity of the world economic activities to price variations. There are certain parallels with the multiproduct WTN analyzed in [14] but there are also new elements specific to the WNEA.

To analyze the sensitivity of price variation in a certain activity sector s we increase from 1 to $1 + \delta_s$ the money transfer in the sector s in $M_{cc's'}$ in (1), where δ_s is a dimensionless fraction variation of price in this sector. After that the matrices G, G^* are recomputed in the usual way described above and their rank probabilities P, P^* are determined. Then we compute the derivatives of probabilities of PageRank $D = dP/d\delta_s = \Delta P/\delta_s$ and CheiRank $D^* = dP^*/d\delta_s = \Delta P^*/\delta_s$. We do these computations at

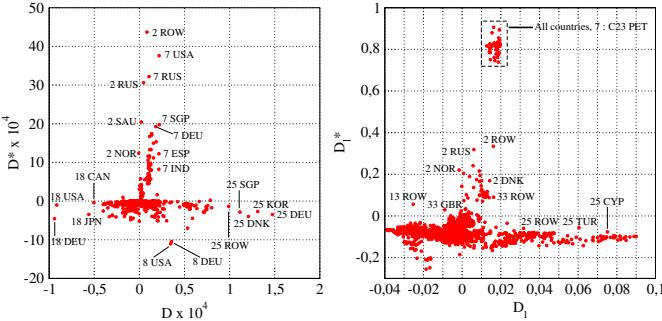


Fig. 11. Left panel: Derivatives $D = dP/d\delta_7$ and $D^* = dP^*/d\delta_7$ for a price variation δ_7 of 7 C23 PET (*Manufacture of coke, refined petroleum products and nuclear fuel*) for year 2008. Right panel: Logarithmic derivatives $D_l = D/P$ and $D_l^* = D^*/P^*$ for the same case as left panel. Codes in panels give sector number $s = 1, \dots, 37$ described in Table 2, country codes are from Table 1. The group of points, highlighted by the dashed box, represents 58 nodes of the form (*country, $s = 7$*) where $s = 7$ is C23 PET (*Manufacture of coke, refined petroleum products and nuclear fuel*).

sufficiently small δ_s values checking that the variations of P, P^* are linear in δ_s . In addition we also compute the logarithmic derivatives $D_l = d \ln P/d\delta_s$, $D_l^* = d \ln P^*/d\delta_s$ which give us relative changes of P, P^* .

The sensitivities to price of $s = 7$ (*Manufacture of coke, refined petroleum products and nuclear fuel*) are shown in Fig. 11. The data for D, D^* in the left panel show a rather complex picture with a significant derivatives not only for $s = 7$ but also for countries with sectors: $s = 18$ (*Manufacture of motor vehicles, trailers and semi-trailers*) at strongly negative D for Germany, USA, Japan; $s = 25$ (*Land transport; transport via pipelines etc*) at significant positive D for Germany, Korea, Denmark, Singapore; of course, for $s = 7$ we have positive D^* , but also for $s = 2$ related to mining and negative D^* for $s = 8$ (*Manufacture of chemicals and chemical products*) for USA and Germany. The logarithmic derivatives provide strong relative changes and are shown in the right panel of Fig. 11.

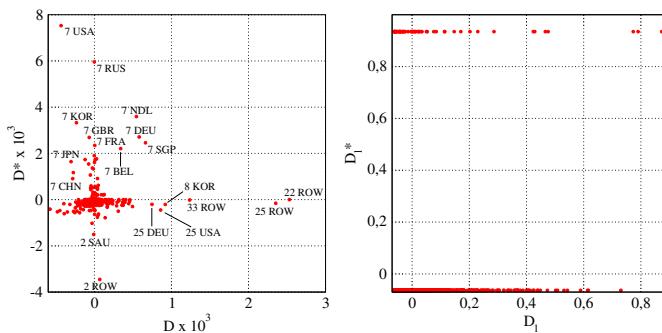


Fig. 12. Same as the left panel of Fig. 11 but using probabilities from the trade value. In the right panel, $D_l^* = 0.9348$ if $s = s'$ and $D_l^* = -0.0633$ if $s \neq s'$.

A similar analysis can be done using the probabilities \hat{P}, \hat{P}^* from the exchange value probabilities (3) instead of the above PageRank and CheiRank probabilities. The results for the value probabilities are presented in Fig. 12 for the same case as in Fig. 11. We see that the results are drastically different especially for the logarithmic derivatives D_l, D_l^* . In fact D_l, D_l^* cannot give correct picture of sensitivity to price variations since for the monetary exchange the network links between nodes are not taken into account and there is only a mechanical re-computation of the value normalization. A similar situation appears also for the multiproduct WTN [14]. Thus we see from Fig. 11 and Fig. 12 that the Google matrix approach provides new elements for the economic activity analysis going significantly beyond the usual consideration of Import-Export method.

The new element of the WNEA, compared to the multiproduct WTN, is existence of transfers between sectors of the same economy. This allows us to consider the sensitivity not only to sectoral prices but also the sensitivity to labor cost in a given country c (e.g. price shock affecting all industries in the same country). This can be taken into account by the introduction of the dimensionless labor cost change in a given country c by replacing the related monetary flows from coefficient 1 to $1 + \sigma_c$ in $M_{cc',ss}$; (1) for a selected country c .

Of course, the above derivatives over price of activity sector and labor country cost give only an approximate consideration of effects of price variations which is a very complex phenomenon. For an economic discussion of the effect of price shocks on international production networks we address a reader to the research performed in [30]. We will see below that our approach gives results being in a good agreement with economic realities thus opening complementary possibilities of economic activity analysis based on the underlying network relations between countries and activity sectors which are absent in the usual Import-Export consideration. We present the results on sensitivity to sector prices and labor cost in next subsections.

3.6 Price shocks and trade balance sensitivity

On the basis of the obtained WNEA Google matrix we can now analyze the trade balance in various activity sectors for all world countries. Usually economists consider the export and import of a given country as it is shown in Fig. 1. Then the trade balance of a given country c can be defined making summation over all sectors:

$$B_c = \sum_s (P_{cs}^* - P_{cs}) / \sum_s (P_{cs}^* + P_{cs}) = (P_c^* - P_c) / (P_c^* + P_c). \quad (13)$$

In economy, P_c, P_c^* are defined via the probabilities of trade value $\hat{P}_{cs}, \hat{P}_{cs}^*$ from (3). In our matrix approach, we define P_{cs}, P_{cs}^* as PageRank and CheiRank probabilities. In contrast to the Import-Export value our approach takes into account the multiple network links between nodes.

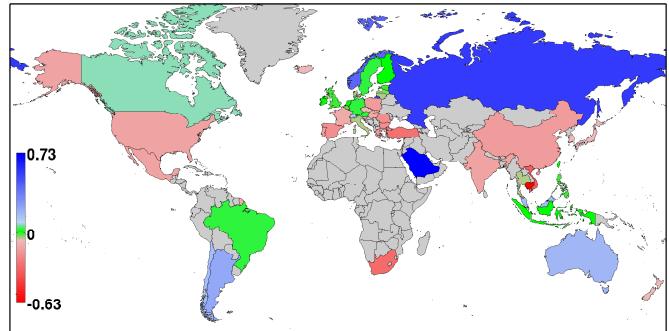
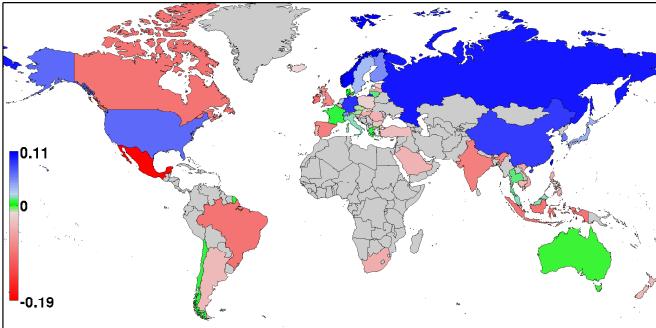


Fig. 13. World map of CheiRank-PageRank balance $B_c = (P_c^* - P_c)/(P_c^* + P_c)$ determined for all $N_c = 58$ countries in year 2008. Top panel shows the probabilities P and P^* given by PageRank and CheiRank vectors; the value of ROW group is $B_{c=58} = 0.023$. Bottom panel shows the probabilities P and P^* computed from the Export and Import value; the value of ROW group is $B_{c=58} = 0.16$. Names of the countries are given in Table 1 and in the world map of countries [22].

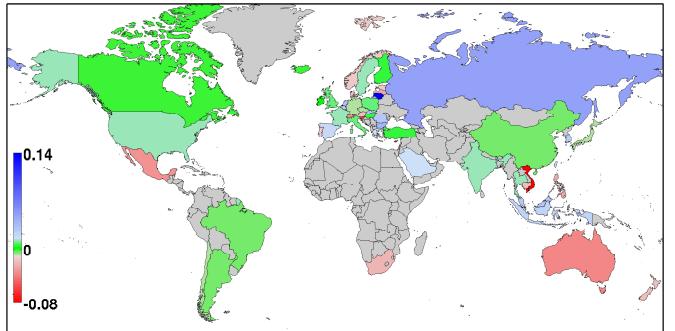
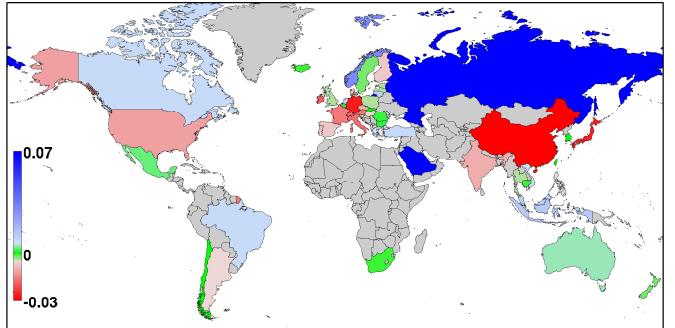


Fig. 14. Derivative of probabilities balance $dB_c/d\delta_7$ over price of sector $s = 7$ C23PET for year 2008. Top panel shows the case when B_c is determined by CheiRank and PageRank vectors as in the top panel of Fig.13; the value of ROW group is $dB_{58}/d\delta_7 = 0.04$. Bottom panel shows the case when B_c is computed from the Export-Import value as in the bottom panel of Fig.13; the value of ROW group is $dB_{58}/d\delta_7 = -0.07$. Names of the countries can be found in Table 1 and in the world map of countries [22].

The comparison of CheiRank-PageRank balance with Export-Import balance for the world countries shown in Fig. 13 for year 2008. Each country is shown by color which is proportional to the country balance B_C (13) with the color bar given on the figure. For Export-Import balance we see the dominance of petroleum producing countries Saudi Arabia, Russia, Norway with the largest values. The CheiRank-PageRank balance highlights new features placing on the top Russia, Norway, Germany, China. In fact, USA has now a slightly positive balance in top panel of Fig. 13) while it was negative before in bottom panel of same figure. We see that the broad network of economic activity relations and links makes the economies of the above countries more important in the world economy while Saudi Arabia, with the largest positive Export-Import balance, loses its leading position. Indeed, the trade of this country is mainly oriented to USA and nearby countries that reduces its importance for world economy (a similar effect has been observed with COMTRADE data [13,14]).

The sensitivity of country balance $dB_c/d\delta_7$ to price variation of sector $s = 7$ *Manufacture of coke, refined petroleum products and nuclear fuel* is shown in Fig. 14. For Export-Import in bottom panel the most sensitive countries are Lithuania (positive) and Vietnam (negative). Lithuania does not produce petroleum, but in fact in 2008

there was a large oil refinery company there which had a large exportation value (see e.g. http://en.wikipedia.org/wiki/Economy_of_Lithuania). The Export-Import approach shows that Russia is slightly positive, even less positive is Saudi Arabia, China and Germany are close to zero change, USA is only very slightly positive. The results of CheiRank-PageRank sensitivity (top panel) are significantly different showing strongly positive sensitivity for Saudi Arabia, Russia and strongly negative sensitivity for China, Germany and Japan; USA goes from slightly positive side in bottom panel to moderate negative one in top panel. The CheiRank-PageRank balance demonstrates much higher sensitivity of Russia, Saudi Arabia and China to price variations of $s = 7$ sector comparing to the case of Export-Import value analysis. The economies of Germany, China and Japan are also very sensitive to petroleum prices that is correctly captured by our analysis. We consider that the CheiRank-PageRank approach describes the economic reality from a new complementary angle and that provides new useful information about complex trade systems. We also note that the highly negative sensitivity of China to petroleum prices has been also obtained on the basis of Google matrix analysis of COMTRADE data (see Fig.21 in [14]).

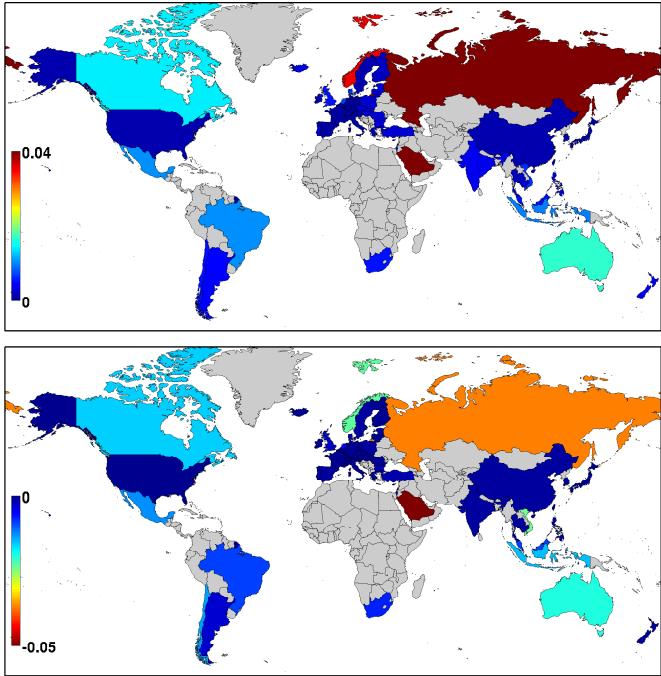


Fig. 15. Derivative of partial probability balance of sector s defined as $dB_{cs}/d\delta_{s'}$ over sector $s' = 7$ C23PET price δ_7 for year 2008. Here $B_{cs} = (P_{cs}^* - P_{cs})/(P_c^* + P_c)$ and $s = 2$ (*C10T14MIN, Mining, extraction,...*) from Table 2. The sector balance sensitivity of countries B_{cs} is determined from CheiRank and PageRank vectors (top panel) and from the exchange value of Export-Import (bottom panel); the values of ROW group are $dB_{58.2}/d\delta_7 = 0.05$ and $dB_{58.2}/d\delta_7 = -0.03$ respectively. Names of the countries can be found at Table 1 and in the world map of countries [22].

It is also possible to determine the cross-sensitivity of activity sectors to price variation. For that we determine the partial exchange balance for a given sector s defined as

$$B_{cs} = (P_{cs}^* - P_{cs}) / \sum_s (P_{cs}^* + P_{cs}) = (P_{cs}^* - P_{cs}) / (P_c^* + P_c), \quad (14)$$

so that the global country balance is $B_c = \sum_s B_{cs}$. Then the sensitivity of partial balance of a given sector s in respect to a price variation of a sector s' is given by the derivative $dB_{cs}/d\delta_{s'}$. The results for $s = 2, s' = 7$ are shown in Fig. 15. We see that two methods give results with even opposite signs. According to the Google matrix analysis the increase of petroleum prices stimulates development of mining while for the Export-Import approach the result is the opposite. In our opinion, the absence of links and next step relations between countries and sectors in the Export-Import methods does not allow to take into account all complexity of economy relations. In contrast the CheiRank-PageRank approach captures effects of all links providing more advanced indications.

The sensitivities $dB_c/d\delta_{s'}$ of CheiRank-PageRank balance of China and USA to price variation of sectors s' are presented in Fig. 16. We see two rather different profiles.

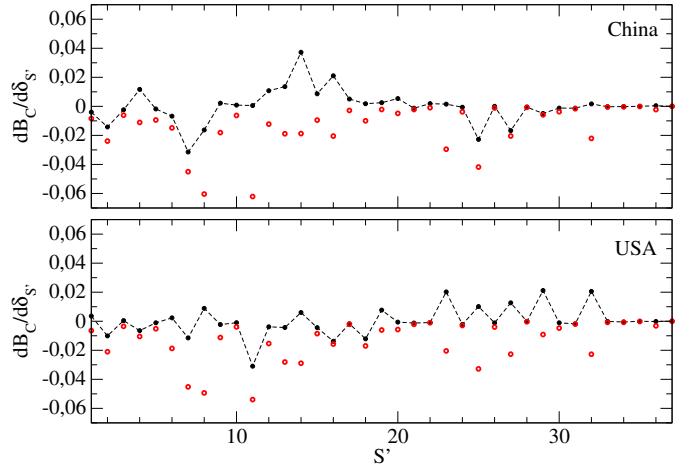


Fig. 16. Top (China) and bottom (USA) panels show derivative $dB_c/d\delta_{s'}$ of country total probability balance B_c over price $\delta_{s'}$ of sector s' for year 2008 (black points connected by dashed line); derivatives of balance without diagonal term ($dB_c/d\delta_{s'} - dB_{cs'}/d\delta_{s'}$) are represented by open red circles. The sector balance of countries B_{cs} and B_c are determined from CheiRank and PageRank vectors. The sectors corresponding to sector index s or s' are listed in Table2.

Thus, for China the derivative $dB_c/d\delta_{s'}$ is positive for sectors $s = 4, 14, 16$ (*Manufacture of textiles; office machinery; radio etc.*) and negative for $s = 7, 25, 27$ (*Petroleum; Land transport etc.; Financial intermediation etc.*). For USA the sensitivity is significantly positive for $s = 23, 29, 32$ (*Sale of motor vehicles etc.; Renting of machinery and equipment etc.; Other business activities*) and negative for $s = 11$ (*Manufacture of basic metals*). Thus the economic activities of these two countries have very different strong and weak points. We note that the sensitivity without the diagonal term ($dB_c/d\delta_{s'} - dB_{cs'}/d\delta_{s'}$) has negative values for almost all sectors for both countries.

The matrices of cross-sector sensitivity $dB_{cs}/d\delta_{s'}$ are shown for China and USA in Fig. 17. Such matrices provide a detailed information of interconnections of various activity sectors. Thus for USA we see that its $s = 8$ (*Manufacture of chemicals etc.*) has a significant negative sensitivity to $s' = 7, 23, 25$ (*Petroleum; Renting of machinery and equipment etc.; Land transport etc.*). Indeed, chemical production is linked with petroleum, machinery and transport. For China we find that its sector $s = 11$ (*Manufacture of basic metals*) has a negative sensitivity to $s' = 8, 23$ (*Manufacture of chemicals etc.; Renting of machinery and equipment etc.*); also $s = 14, 16$ have a negative derivative in respect to $s' = 11$).

Of course, the cross sensitivity to price variations in one sector and their effects on another sector, based on (14), is a very delicate thing since a price in one sector can affect prices in other sectors also in other manner since economic systems learn and adapt while here we considered only linear algebraic relations without any adaptation features. However, even being linear, the Google matrix approach provides a detailed information on hidden interactions and inter-dependencies of various economic

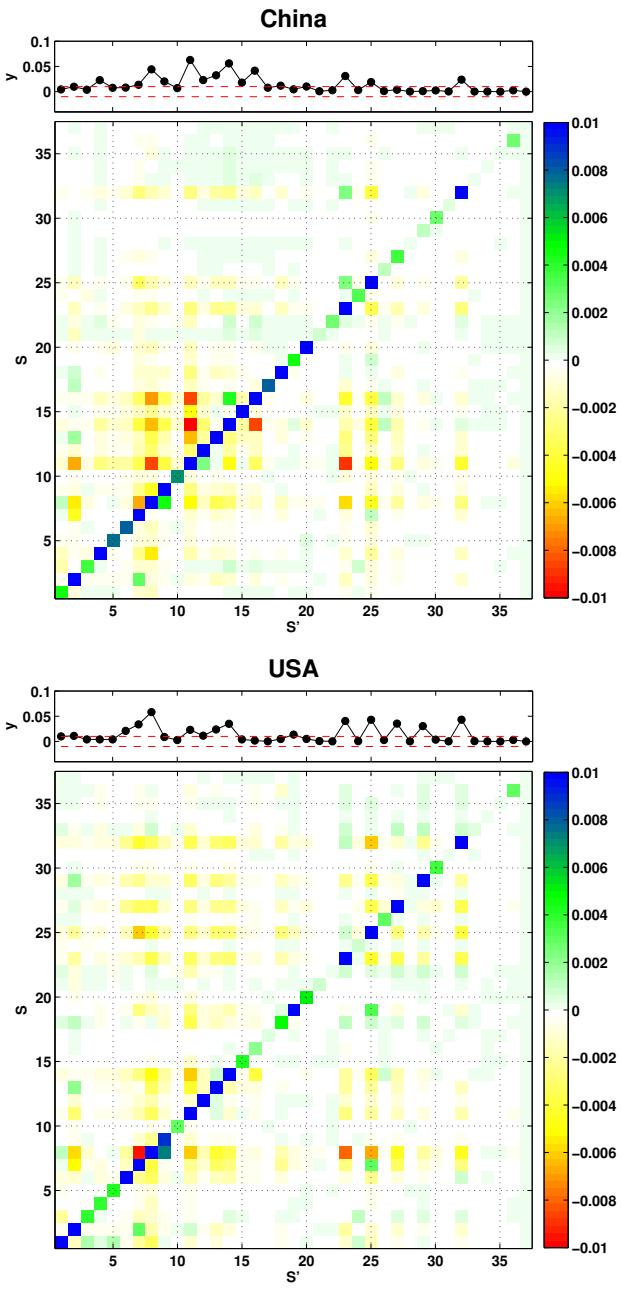


Fig. 17. China (top) and USA (bottom) examples of derivative $dB_{cs}/d\delta_{s'}$ of partial probability balance B_{cs} of sector s over price $\delta_{s'}$ of sector s' for year 2008. Diagonal terms, given by $y = dB_{cs}/d\delta_s$ for $s = s'$, are shown on the top panels of each example. Sectors s' and s are shown in x -axis and y -axis respectively (indexed as in Table 2 from 1 to 37), while $dB_{cs}/d\delta_{s'}$ is represented by colors with a threshold value given by $+\epsilon$ and $-\epsilon$ for negative and positive values respectively, also shown in red dashed lines on top panels with diagonal terms. Here $\epsilon = 0.01$ for USA and China; partial balance B_{cs} is defined by CheiRank and PageRank probabilities.

activities for various countries that can provide a useful message even for nonlinear adapting systems.

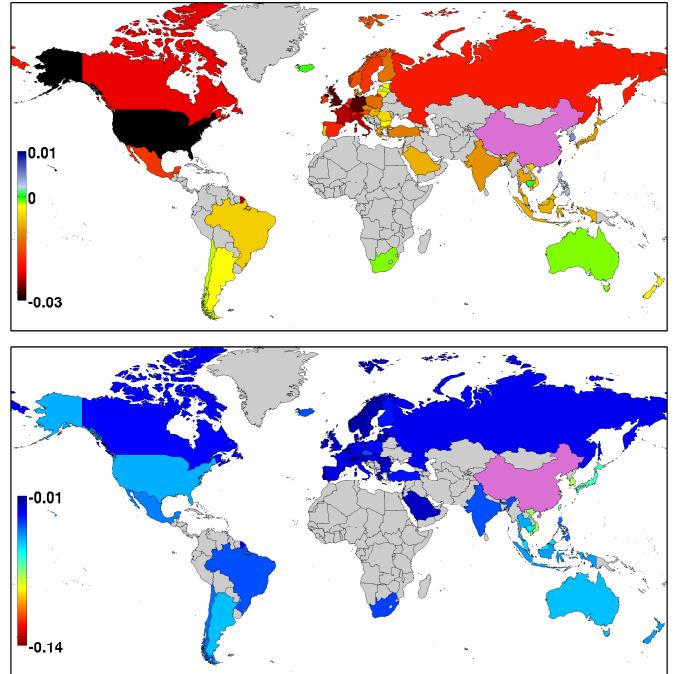


Fig. 18. Derivative of probabilities balance $dB_c/d\sigma_{c'}$ over labor cost of China $c' = 37$ for year 2008. Top panel shows the case when B_c is determined by CheiRank and PageRank vectors; here the special values are $dB_{58}/d\sigma_{37} = -0.0146$ for ROW group (gray) and $dB_{37}/d\sigma_{37} = 0.3217$ for China (magenta). Bottom panel shows the case when B_c is computed from the Export-Import value; the special values are $dB_{58}/d\sigma_{37} = -0.0352$ for ROW group (gray) and $dB_{37}/d\sigma_{37} = 0.4810$ for China (magenta). Names of the countries can be found in Table 1 and in the world map of countries [22].

3.7 World map of sensitivity to labor cost

Using the established structure of WNEA we can study the sensitivity of country balance $dB_c/d\sigma_c'$ to the labor cost in different countries. At the difference of sectoral shocks on one product, here the price shock affects all industries in a country. As before, the change in price has to be small enough for the resulting simulation to remain in a neighbourhood of the original data. Indeed, larger shocks would trigger a series of substitution effects diverting trade to other partners.

The derivative $dB_c/d\sigma_c'$ is computed numerically as described in Sec. 3.5. The world sensitivity to the labor cost of China is shown in Fig. 18. Of course, the largest derivative is found for China itself ($dB_c/d\sigma_c$ at $c = 37$ from Table 1). The effect on other countries is given by non-diagonal derivatives at $c \neq c' = 37$. From the CheiRank-PageRank balance we find that the most strong negative effect (minimal negative $dB_c/d\sigma_c'$) is obtained for USA, Germany, UK; a positive derivative is visible only for Chinese Taipei ($s = 38$) and S.Korea ($s = 19$). For the Export-Import balance the results are rather different: at first all derivatives at $c \neq c'$ are negative; among the most negative values are such countries as Hong Kong (most negative with dark red color but hardly visible due to its

small size), Chinese Taipei, S.Korea, Vietnam. Thus the Google matrix approach bring a new perspective for analysis of complex of economical relations between countries and sectors.

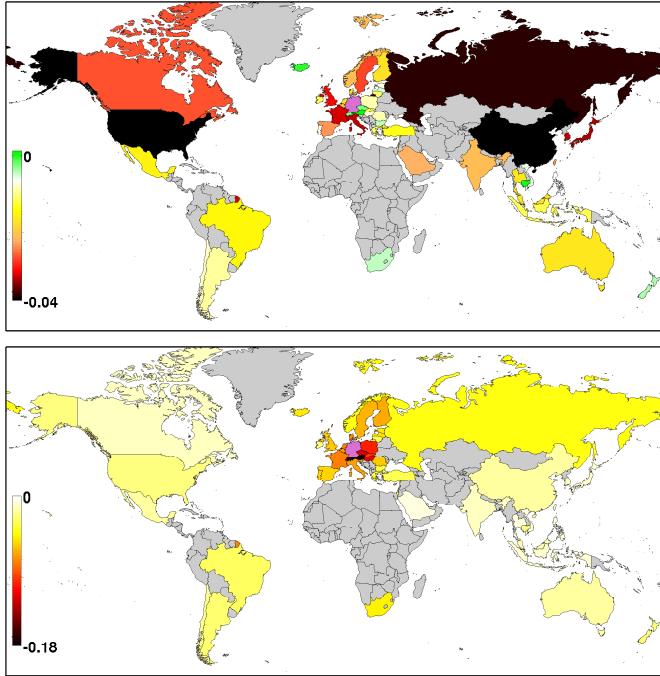


Fig. 19. Same as in Fig. 18 with the derivative $dB_c/d\sigma_c'$ over the labor cost $c' = 11$ of Germany for year 2008. Top panel shows the case when B_c is determined by CheiRank and PageRank vectors; the special values are $dB_{58}/d\sigma_{11} = -0.0367$ for ROW group (gray) and $dB_{11}/d\sigma_{11} = 0.3248$ for Germany (magenta). Bottom panel shows the case when B_c is computed from the Export-Import value; the special values are $dB_{58}/d\sigma_{11} = -0.0280$ fro ROW group (gray) and $dB_{11}/d\sigma_{11} = 0.4911$ for Germany (magenta). Names of the countries can be found in Table 1 and in the world map of countries [22].

Another results for the effects of labor cost in Germany and in USA are shown in Fig. 19 and Fig. 20. In the case of Germany the most strong negative sensitivity is for USA, Russia, China for CheiRank-PageRank balance while for Import-Export it is Switzerland and Austria. However, USA and Russia are relatively weakly affected. This again stresses the qualitative difference between these two approaches.

The increase of USA labor cost in Fig. 20 produces positive derivatives of CheiRank-PageRank balance for Canada and Mexico that looks reasonable from a view point of economy since these countries will profit from higher production costs in USA. In opposite, Export-Import gives most strong negative derivatives for Canada and Mexico.

The whole matrix of labor cost derivatives $dB_c/d\sigma_c'$ of the CheiRank-PageRank balance B_c is shown in Fig. 21 (numerical values of derivatives are given at [26]). Of course, the diagonal terms have the strongest positive derivatives,

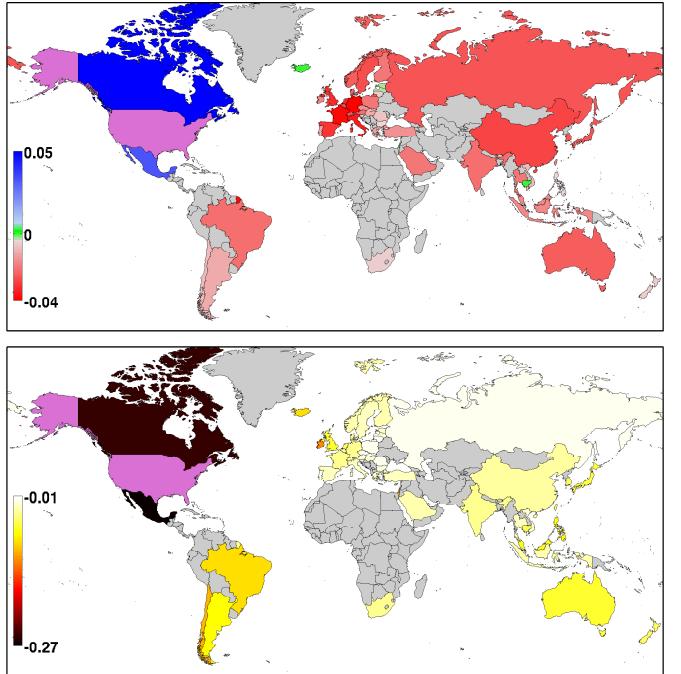


Fig. 20. Same as in Fig. 18 with the derivative $dB_c/d\sigma_c'$ over labor cost $c' = 34$ of USA for year 2008. Top panel shows the case when B_c is determined by CheiRank and PageRank vectors; the special values are $dB_{58}/d\sigma_{34} = -0.0257$ for ROW group (gray) and $dB_{34}/d\sigma_{34} = 0.3148$ for USA (magenta). Bottom panel shows the case when B_c is computed from the Export-Import value; the special values are $dB_{58}/d\sigma_{34} = -0.0632$ for ROW group (gray) and $dB_{34}/d\sigma_{34} = 0.4852$ for USA (magenta). Names of the countries can be found in Table 1 and in the world map of countries [22].

but off-diagonal terms change signs and characterize the sensitivity of one country to labor cost in other country. The vertical lines with high derivative values correspond to Germany ($c' = 11$), Japan ($c' = 18$), S.Korea ($c' = 19$), USA ($c' = 34$), China ($c' = 37$), Russia ($c' = 41$). The rest of the world (ROW) group also have a visible effect of other countries ($c' = 58$). Thus is it desirable to obtain individual OECD data for countries of the ROW group.

In Fig. 21 we considered the effects of the labor cost in various countries. We can also see the effect of price variation $\delta_{s'}$ in a given sector s' on the CheiRank-PageRank balance B_c of country c . This sensitivity is given by the rectangular matrix of derivatives $dB_c/d\delta_{s'}$ shown in Fig. 22 (numerical data are given at [26]). The strongest positive derivatives (blue squares) are for $s' = 2, c = 50$ (mining and Saudi Arabia), $s' = 23, c = 44$ (motors and Hong Kong), $s' = 27, c = 20$ (finance and Luxembourg). The strongest negative derivatives (red squares) are for $s' = 2, c = 3$ (mining and Belgium), $s' = 2, c = 42$ (mining and Singapore which economy is very sensitive to mining products), $s' = 7, c = 11$ (petroleum and Germany), $s' = 7, c = 18$ (petroleum and Japan), $s' = 7, c = 37$ (petroleum and China), $s' = 11, c = 34$ (manufacture of basic metals and USA), $s' = 11, c = 42$ (manufacture of

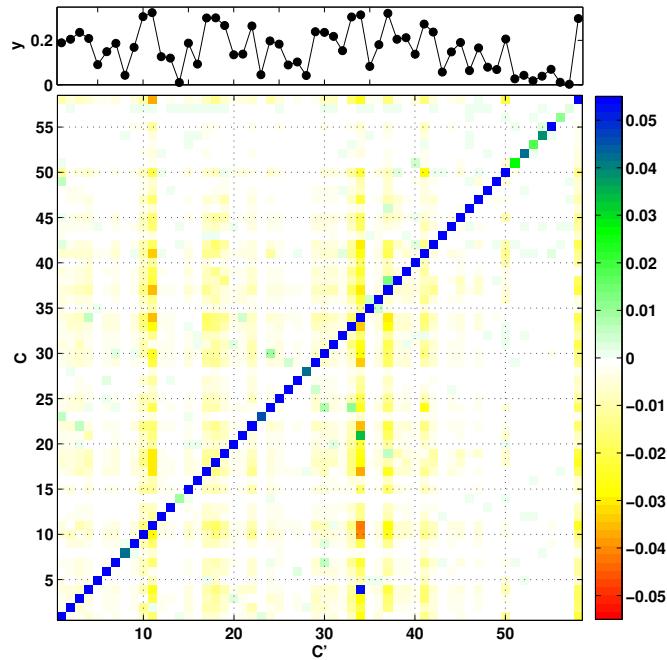


Fig. 21. Global view of the effect of labor cost variation in country c' on country c in 2008. Matrix elements $dB_c/d\sigma_{c'}$ are given in colors shown by the truncated color scale; matrix elements above the scale (diagonal terms) are shown in the top inset where $y = dB_c/d\sigma_{c'}$. In the matrix of derivatives shown by color, x -axis shows the index c' of country where a labor cost variation $\sigma_{c'}$ takes place and y -axis shows the country c affected by the change. Here B_c is computed from CheiRank and PageRank probabilities. Country identification numbers $c = 1, \dots, 58$ are given in Table 1.

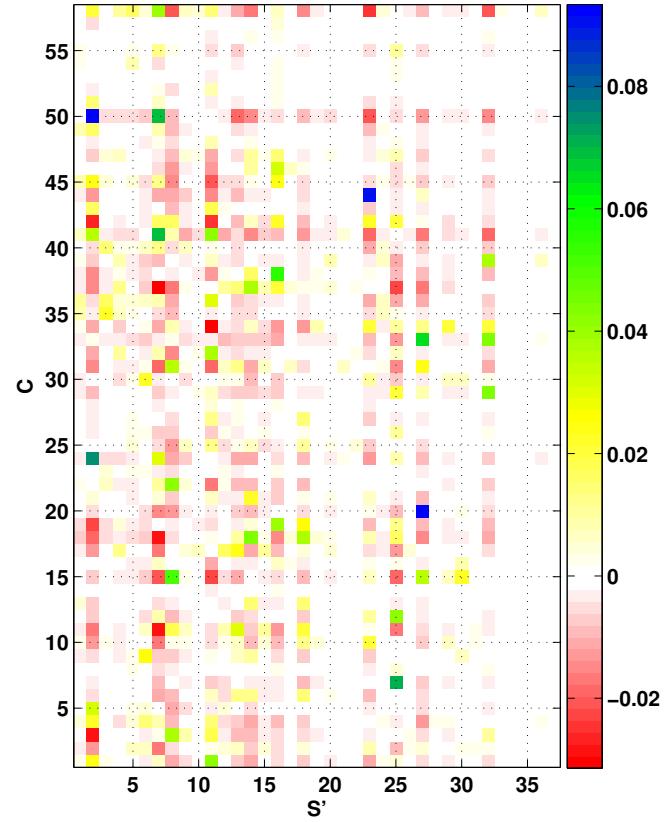


Fig. 22. Global view of the effect of sector s' price variation on balance of country c in 2008. Colors are proportional to matrix elements $dB_c/d\delta_{s'}$, x -axis shows the sector index s' (sectors are given in Table 2) and y -axis gives the country index c affected by the change (countries are given in Table 1). Here B_c is computed from CheiRank and PageRank probabilities.

basic metals and Singapore). All these results are in agreement with the economic realities of sensitivity of the above countries to given activity sectors. This shows the strength of the Google matrix approach to analysis of WNEA.

3.8 World transformation matrix of activity sectors

From the obtained Google matrices G, G^* of WNEA we can analyze the transformation of the activity sectors by the world economy. For this analysis we compute the transfer matrix

$$T = (1 - \eta)(1 - \eta G^*)^{-1}G, \quad (15)$$

where η is a numerical constant. Our study show that as in the case of damping factor α the results are robust to variations of η in the range $0.5 < \eta < 0.9$ and thus in the following we present the results for $\eta = 0.7$. We note that a similar construction for ImpactRank has been used for Wikipedia networks [27] and the C.elegans neural network [31]. In a certain sense (15) can be considered as a scattering matrix of particles entering in a system by G term and then going out by the expansion term $1 + \eta G^* + (\eta G^*)^2 \dots = 1/(1 - \eta G^*)$. In this approach η describes a relaxation rate in the system. We note that T belongs to the Google matrix class.

From the global matrix T of size N we obtain the reduced matrix $R_{ss'}(c)$ of size N_s describing the transformation for activity sectors for a country c . We have $R_{ss'}(c') = \sum_c T_{s,s',c,c'}$ where c' is a target country we are interested in. The matrices $R_{ss'}(c')$ giving the transformation of sector s' to all other sectors s for c' of China, USA, Germany are given in [26]. The reduced transformation matrix for the whole world is obtained by averaging over countries with $R_{ss'} = \sum_{c'} R_{ss'}(c')/N_{c'}$ (see Fig. 23). The results of Fig. 23 show a few characteristic features: the reduced transfer matrix has a strong diagonal element (this is because each product is strong projection on itself), there are characteristic horizontal lines corresponding to important sectors (e.g. $s = 2, 7, 11, 25$).

By considering a transformation of a given sector to all other sectors for a given country. For $s' = 2$ (mining) we present the resulting transformed vector $v(s)$ in Fig. 24 for France, Germany, Switzerland and USA. The global profiles are similar but there are significant enhancement for Germany at sector $s = 7$ (petroleum) and for Switzerland at sector $s = 20$ (manufacturing and recycling). For comparison we show the results of transformation of input/output matrix M of (1). The comparison shows a

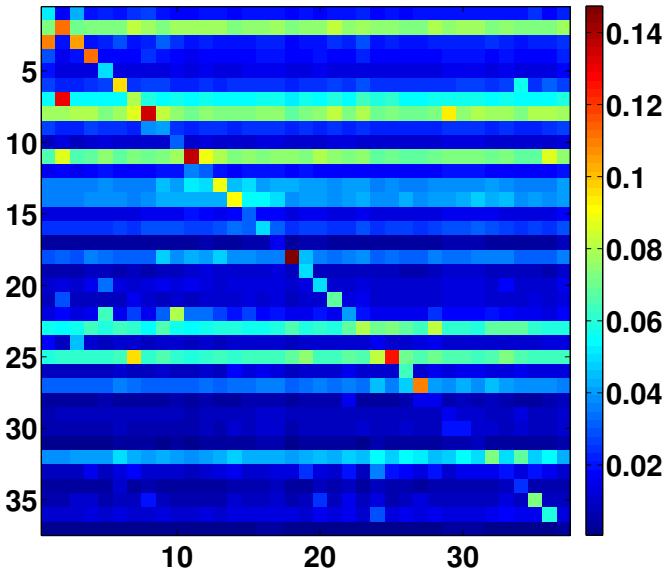


Fig. 23. Image of the average reduced transfer matrix $R_{s,s'}$ of sectors to sectors for the whole world (averaged over countries) for year 2008. Here x -axis represents the initial sector s' and y -axis represents the final sectors s into which s' is transformed. The sector numbering is given in Table 2. Colors are proportional to matrix elements and $\eta = 0.7$.

drastic difference between two approaches which we attribute to the fact that M does not take into account the multiple network transitions.

The transformation for the sector $s' = 34$ are shown in Fig. 25 for Cyprus (blue), Singapore (red), Luxembourg (green) and Malta (black). We see that for Luxembourg there is a strong transformation of $s' = 20$ to $s = 6$ (publishing). At the same time the global profile, being different from the case of Fig. 24 with $s' = 2$, has similar features for different countries. The comparison with the transformation results from value exchange matrix $M_{ss',cc'}$ are again very different as in the case of Fig. 24.

The obtained results for the activity sector transformation by the WNEA open new possibilities for analysis of interactions between the world economic activities. The Google matrix approach provides new type of results being very different from usual Input/Output matrix approach. This is related to the fact that the transformation matrix (14) takes into account summation over various cycles over the network.

4 Discussion

In this work we have developed the Google matrix analysis of the world network of economic activities from the OECD-WTO TiVA database. The PageRank and CheiRank probabilities allowed to obtain ranking of world countries independently of their richness being mainly determined by the efficiency of their economic relations. The developed approach demonstrated the asymmetry in the economic activity sectors some of which are export oriented

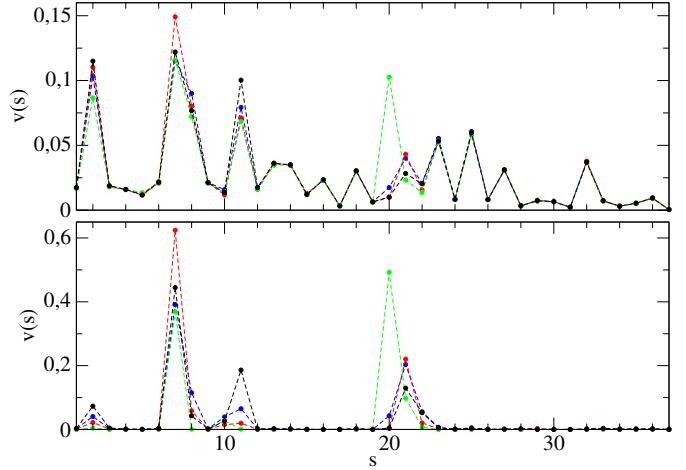


Fig. 24. *Top panel:* Examples of profile $v(S)$ for transformation vector from the reduced transfer matrix for several countries in 2008. Here the initial sector is $s = 2$ (mining) while the transformed vector $v(s)$ is formed by the matrix defined in Fig. 23; the countries are France (blue), Germany (red), Switzerland (green) and USA (black). *Bottom panel:* For comparison, we show here the same as top panel but instead of T, R matrices we use the input/output matrix M with normalized columns (dangling nodes are not replaced here, transitions inside one country are taken to be zero); a column s' of such a matrix for country c' is given by $\sum_c M_{ss',cc'}$; here the same countries are shown by same colors as in top panel..

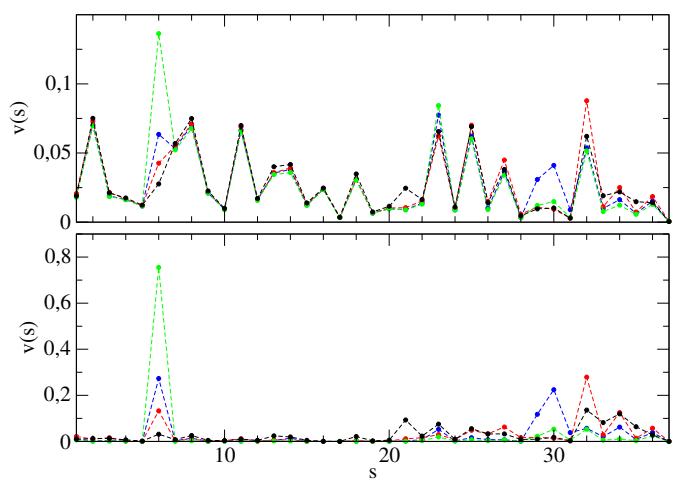


Fig. 25. Same as in Fig. 24 for the initial sector $s' = 34$ (education). The results are shown for Cyprus (blue), Singapore (red), Luxembourg (green) and Malta (black).

and others are import oriented. We also showed that the eigenstates of the WNEA Google matrix select specific quasi-isolated communities oriented to specific activity sectors. The CheiRank-PageRank balance B_c allows to determine economically rising countries with robust network of economic relations. The sensitivity of this B_c to price variations and labor cost in various countries determines the hidden relations between world economies being not visible via usual Export-Import exchange analysis. The

Google matrix analysis determines also the transformation features of world activity sectors.

The comparison with the multiproduct world trade network from UN COMTRADE shows certain similarities between the two networks of WNEA and WTN. At the same time the WNEA data provides new elements for interactions of activity sectors while there are no direct interactions of products in COMTRADE database. From this viewpoint the OECD-WTO data captures the economic reality on a deeper level. But at the same time the OECD-WTO network is less developed compared to COMTRADE (less countries, years, sectors). Thus it is highly desirable to extend the OECD-WTO database.

We think that the Google matrix analysis developed here and in [13, 14] captures better the new reality of multifunctional directed tensor interactions and that the universal features of this approach can be also extended to multifunctional financial network flows which now attract an active interest of researchers [32, 33]. Unfortunately, the data on financial flows have much less accessibility compared to the networks discussed here.

We point that recently some of the matrix methods, developed in physics community, started to find active application for economy systems (see e.g. [34, 35]). However, usually for physicists these matrices have been from the unitary or Hermitian ensembles, where the Random Matrix Theory allowed to obtain certain universal results. Here, we show that the directed networks and tensors appearing in the interacting economy systems are described by the matrices of Perron-Frobenius operators which had not been studied much in physics. Thus the new field of research is now opened for physicists, mathematicians and computer scientists with application to complex interacting economy systems.

5 Acknowledgments

We thank the representatives of OECD [1] and WTO [2] for providing us with the friendly access to the data sets investigated in this work. One of us (VK) thanks the Economic Research and Statistics Division, WTO Genève for hospitality during his internship there. We thank L.Ermann for useful discussions and advices on preparation of figures. This research is supported in part by the EC FET Open project “New tools and algorithms for directed network analysis” (NADINE No 288956).

References

1. OECD, *Secretary-general's report to ministers 2014*, Available: <http://www.oecd.org>. Accessed April 2015
2. World Trade Organization (2014) *International Trade Statistics 2014* Available: <http://www.wto.org/>. Accessed April 2015
3. W.W. Leontief, *Domestic production and foreign trade: the American capital position re-examined*, Proc. American Phil. Soc. **97**(4), 332 (1953)
4. W.W. Leontief, *Input-Output economics*, Oxford University Press, New York, NY (1986)
5. R.E. Miller and P.D. Blair, *Input-output analysis: foundations and extensions*, Cambridge University Press, Cambridge UK (2009)
6. S. Dorogovtsev, *Lectures on complex networks*, Oxford University Press, Oxford (2010)
7. S.Brin and L.Page, *The anatomy of a large-scale hyper-textual Web search engine*, Computer Networks and ISDN Systems **30**, 107 (1998)
8. A.M. Langville and C.D. Meyer, *Google's PageRank and beyond: the science of search engine rankings*, Princeton University Press, Princeton (2006)
9. L.Ermann, K.M. Frahm and D.L. Shepelyansky *Google matrix analysis of directed networks*, arXiv:1409.0428 [physics.soc-ph] (2014)
10. M. Franceschet, *PageRank: standing on the shoulders of giants*, Communications of the ACM **54**(6), 92 (2011)
11. S. Vigna, *Spectral ranking*, arXiv:0912.0238v13 [cs.IR] (2013)
12. United Nations Commodity Trade Statistics Database Available: <http://comtrade.un.org/db/>. Accessed April 2015
13. L.Ermann and D.L. Shepelyansky, *Google matrix of the world trade network*, Acta Physica Polonica A **120**, A158 (2011)
14. L.Ermann and D.L. Shepelyansky, *Google matrix analysis of the multiproduct world trade network*, Eur. Phys. J. B **88**, 84 (2015)
15. D. Garlaschelli and M.I.Loffredo, *Structure and evolution of the world trade network*, Physica A: Stat. Mech. Appl. **355**, 138 (2005)
16. J. He and M.W. Deem, *Structure and response in the world trade network*, Phys. Rev. Lett. **105**, 198701 (2010)
17. G. Fagiolo, J.Reyes and S. Schiavo, *The evolution of the world trade web: a weighted-network analysis*, J. Evol. Econ. **20**, 479 (2010)
18. M. Barigozzi, G. Fagiolo and D. Garlaschelli, *Multinetwork of international trade: a commodity-specific analysis*, Phys. Rev. E **81**, 046104 (2010)
19. L. De Benedictis and L. Tajoli, *The world trade network*, World Economy **34**(8), 1417 (2011)
20. T. Deguchi, K.Takahashi, H.Takayasu and M. Takayasu, *Hubs and authorities in the world trade network using a Weighted HITS algorithm*, PLoS ONE **9**(7), e1001338 (2014)
21. A. Kireyev and A. Leonidov, *Network effects of international shock spillovers*, Working paper, International Monetary Fund, N.Y (2015)
22. Web page *Maps of the world* Available: <http://www.mapsofworld.com/>. Accessed April 2015
23. United Nations ISIC Rev.3 Available: <http://unstats.un.org/unsd/cr/registry/regcst.asp?C1=2>. Accessed February 2015
24. A.D. Chepeliantskii, *Towards physical laws for software architecture*, arXiv:1003.5455 [cs.SE] (2010)
25. A.O.Zhirov, O.V.Zhirov and D.L. Shepelyansky, *Two-dimensional ranking of Wikipedia articles*, Eur. Phys. J. B **77**, 523 (2010)
26. Web page *Google matrix of the world network of economic activities* Available: <http://www.quantware.ups-tlse.fr/QWLlib/wneamatrix>. Accessed April 2015.

27. K.M.Frahm, Y.-H.Eom, and D.L. Shepelyansky, *Google matrix of the citation network of Physical Review* , Phys. Rev. E **89**, 052814 (2014)
28. O.V.Zhirov and D.L.Shevelyansky, *Anderson transition for Google matrix eigenstates*, arXiv:1502.00584[cond-mat.dis-nn] (2015)
29. L. Ermann, K.M. Frahm and D.L. Shepelyansky, *Spectral properties of Google matrix of Wikipedia and other networks*, Eur. Phys. J. B **86**, 193 (2013)
30. H. Escaith and F. Gonguet, *International supply chains as real transmission channels of financial shocks* Capco Inst. J. of Financial Transformation **31**, 83 (2011)
31. V.Kandiah and D.L.Shevelyansky, *Google matrix analysis of C.elegans neural network*, Phys. Lett. A **378**, 1932 (2014)
32. B. Craig and G. von Peter, *Interbank tiering and money center bank*, Discussion paper N 12, Deutsche Bundesbank (2010)
33. R.J. Garratt, L. Mahadeva and K. Sviridzenka, *Mapping systemic risk in the international banking network*, Working paper N 413, Bank of England (2011)
34. J.-P. Bouchaud and M. Potters, *Theory of financial risk and derivative pricing*, Cambridge University Press, Cambridge UK (203)
35. M.C. Munnix, R. Schaefer and T. Guhr, *A random matrix approach to credit risk*, PLoS ONE **9**(5), e98030 (2014)

	country name	country code	country flag		country name	country code	country flag
1	Australia	AUS		30	Sweden	SWE	
2	Austria	AUT		31	Switzerland	CHE	
3	Belgium	BEL		32	Turkey	TUR	
4	Canada	CAN		33	United Kingdom	GBR	
5	Chile	CHL		34	United States	USA	
6	Czech Republic	CZE		35	Argentina	ARG	
7	Denmark	DNK		36	Brazil	BRA	
8	Estonia	EST		37	China	CHN	
9	Finland	FIN		38	Chinese Taipei	TWN	
10	France	FRA		39	India	IND	
11	Germany	DEU		40	Indonesia	IDN	
12	Greece	GRC		41	Russia	RUS	
13	Hungary	HUN		42	Singapore	SGP	
14	Iceland	ISL		43	South Africa	ZAF	
15	Ireland	IRL		44	Hong Kong	HKG	
16	Israel	ISR		45	Malaysia	MYS	
17	Italy	ITA		46	Phillippines	PHL	
18	Japan	JPN		47	Thailand	THA	
19	Korea	KOR		48	Romania	ROU	
20	Luxembourg	LUX		49	Vietnam	VNM	
21	Mexico	MEX		50	Saudi Arabia	SAU	
22	Netherlands	NLD		51	Brunei Darussalam	BRN	
23	New Zealand	NZL		52	Bulgaria	BGR	
24	Norway	NOR		53	Cyprus	CYP	
25	Poland	POL		54	Latvia	LVA	
26	Portugal	PRT		55	Lithuania	LTU	
27	Slovak Republic	SVK		56	Malta	MLT	
28	Slovenia	SVN		57	Cambodia	KHM	
29	Spain	ESP		58	Rest of the World	ROW	

Table 1. List of $N_c = 58$ countries (with rest of the world ROW) with country name, code and flag.

	OECD ICIO Category	ISIC Rev. 3 correspondence
1	C01T05 AGR	01 - Agriculture, hunting and related service activities 02 - Forestry, logging and related service activities 05 - Fishing, operation of fish hatcheries and fish farms; service activities incidental to fishing
2	C10T14 MIN	10 - Mining of coal and lignite; extraction of peat 11 - Extraction of crude petroleum and natural gas; service activities incidental to oil and gas extraction excluding surveying 12 - Mining of uranium and thorium ores 13 - Mining of metal ores 14 - Other mining and quarrying
3	C15T16 FOD	15 - Manufacture of food products and beverages 16 - Manufacture of tobacco products
4	C17T19 TEX	17 - Manufacture of textiles 18 - Manufacture of wearing apparel; dressing and dyeing of fur 19 - Tanning and dressing of leather; manufacture of luggage, handbags, saddlery, harness and footwear
5	C20 WOD	20 - Manufacture of wood and of products of wood and cork, except furniture; Manufacture of articles of straw and plaiting materials
6	C21T22 PAP	21 - Manufacture of paper and paper products 22 - Publishing, printing and reproduction of recorded media
7	C23 PET	23 - Manufacture of coke, refined petroleum products and nuclear fuel
8	C24 CHM	24 - Manufacture of chemicals and chemical products
9	C25 RBP	25 - Manufacture of rubber and plastics products
10	C26 NMM	26 - Manufacture of other non-metallic mineral products
11	C27 MET	27 - Manufacture of basic metals
12	C28 FBM	28 - Manufacture of fabricated metal products, except machinery and equipment
13	C29 MEQ	29 - Manufacture of machinery and equipment n.e.c.
14	C30 ITQ	30 - Manufacture of office, accounting and computing machinery
15	C31 ELQ	31 - Manufacture of electrical machinery and apparatus n.e.c.
16	C32 CMQ	32 - Manufacture of radio, television and communication equipment and apparatus
17	C33 SCQ	33 - Manufacture of medical, precision and optical instruments, watches and clocks
18	C34 MTR	34 - Manufacture of motor vehicles, trailers and semi-trailers
19	C35 TRQ	35 - Manufacture of other transport equipment
20	C36T37 OTM	36 - Manufacture of furniture; manufacturing n.e.c. 37 - Recycling
21	C40T41 EGW	40 - Electricity, gas, steam and hot water supply 41 - Collection, purification and distribution of water
22	C45 CON	45 - Construction
23	C50T52 WRT	50 - Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel 51 - Wholesale trade and commission trade, except of motor vehicles and motorcycles 52 - Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods
24	C55 HTR	55 - Hotels and restaurants
25	C60T63 TRN	60 - Land transport; transport via pipelines 61 - Water transport 62 - Air transport 63 - Supporting and auxiliary transport activities; activities of travel agencies
26	C64 PTL	64 - Post and telecommunications
27	C65T67 FIN	65 - Financial intermediation, except insurance and pension funding 66 - Insurance and pension funding, except compulsory social security 67 - Activities auxiliary to financial intermediation
28	C70 REA	70 - Real estate activities
29	C71 RMQ	71 - Renting of machinery and equipment without operator and of personal and household goods
30	C72 ITS	72 - Computer and related activities
31	C73 RDS	73 - Research and development
32	C74 BZS	74 - Other business activities
33	C75 GOV	75 - Public administration and defense; compulsory social security
34	C80 EDU	80 - Education
35	C85 HTH	85 - Health and social work
36	C90T93 OTS	90 - Sewage and refuse disposal, sanitation and similar activities 91 - Activities of membership organizations n.e.c. 92 - Recreational, cultural and sporting activities 93 - Other service activities
37	C95 PVH	95 - Private households with employed persons

Table 2. List of sectors considered by Input/Output matrices from OECD database, their correspondence to the ISIC classification is also given.

Sector	\hat{K} (1995)	% vol (1995)	\hat{K}^* (1995)	% vol (1995)	\hat{K} (2008)	% vol (2008)	\hat{K}^* (2008)	% vol (2008)
1	19	2.2979	16	2.9763	20	1.9532	16	2.0902
2	27	1.2993	2	8.6183	24	1.5245	1	15.8784
3	3	6.0117	12	3.3271	11	3.9327	17	1.9835
4	10	3.9579	14	3.0831	17	2.0934	19	1.8634
5	30	1.108	20	1.9037	33	0.60075	22	1.3001
6	11	3.5687	6	4.2128	18	2.0608	14	2.3736
7	4	5.9126	19	2.2783	1	11.589	4	6.34
8	2	6.251	1	10.6954	3	6.0558	2	9.1103
9	17	2.4035	15	3.0546	19	1.9785	13	2.5549
10	28	1.2	21	1.8337	29	1.0389	21	1.3177
11	8	4.4393	3	8.0658	4	5.4907	3	8.3184
12	20	2.2646	17	2.7194	23	1.6212	15	2.2182
13	9	4.0642	8	4.0365	9	4.0117	9	4.0597
14	12	3.3353	13	3.158	8	4.0642	6	5.0066
15	18	2.3789	9	4.0148	25	1.456	18	1.8673
16	15	2.7053	10	3.8054	14	2.7844	11	3.6339
17	31	1.0034	23	1.1434	34	0.31041	29	0.40161
18	7	5.2722	7	4.1643	6	5.1478	10	3.9907
19	26	1.3665	22	1.7813	26	1.3028	23	1.2752
20	24	1.6331	27	0.67546	22	1.6652	20	1.3858
21	21	2.1673	30	0.34377	10	3.946	30	0.39969
22	1	6.538	32	0.22022	2	6.8692	32	0.15209
23	5	5.8472	4	7.9296	7	4.6893	8	4.6745
24	25	1.5283	29	0.37682	27	1.2377	27	0.62202
25	6	5.8385	5	6.5023	5	5.2454	5	5.8065
26	29	1.1862	26	0.6839	28	1.2179	26	0.62929
27	13	2.7584	18	2.3006	15	2.5623	12	3.3487
28	33	0.70446	24	0.93849	31	0.84772	33	0.105
29	36	0.16329	33	0.18955	36	0.21276	24	0.81082
30	34	0.53799	28	0.39581	32	0.67481	28	0.61668
31	35	0.36919	31	0.33351	35	0.24684	31	0.24177
32	16	2.618	11	3.372	13	3.0455	7	4.7163
33	14	2.7071	34	0.064931	12	3.3939	35	0.06377
34	32	0.89993	36	0.0416	30	1.036	34	0.09439
35	22	1.8912	35	0.045551	16	2.2601	36	0.025979
36	23	1.7326	25	0.7136	21	1.8131	25	0.72283
37	37	0.03899	37	0	37	0.019524	37	0

Table 3. First column gives the sectors from OECD database, for each of them the following columns give the ImportRank \hat{K} with the sector fraction in global trade value and ExportRank \hat{K}^* with sector fraction in global trade value. Data are shown for 1995 and 2008.

	K	K^*	K_2	\hat{K}	\hat{K}^*
1	DEU C34 MTR	ROW C10T14 MIN	DEU C24 CHM	USA C23 PET	ROW C10T14 MIN
2	USA C75 GOV	RUS C10T14 MIN	USA C65T67 FIN	JPN C23 PET	SAU C10T14 MIN
3	ROW C75 GOV	SAU C10T14 MIN	DEU C29 MEQ	USA C75 GOV	RUS C10T14 MIN
4	SAU C85 HTH	USA C24 CHM	DEU C34 MTR	ROW C45 CON	USA C24 CHM
5	GBR C85 HTH	DEU C24 CHM	DEU C27 MET	CHN C32 CMQ	CAN C10T14 MIN
6	USA C34 MTR	DEU C27 MET	USA C74 BZS	CHN C27 MET	DEU C24 CHM
7	ROW C45 CON	NOR C10T14 MIN	DEU C50T52 WRT	USA C45 CON	NOR C10T14 MIN
8	ROW C15T16 FOD	RUS C27 MET	USA C24 CHM	DEU C34 MTR	AUS C10T14 MIN
9	USA C15T16 FOD	USA C50T52 WRT	DNK C60T63 TRN	KOR C23 PET	CHN C30 ITQ
10	RUS C50T52 WRT	DEU C29 MEQ	GBR C74 BZS	DEU C23 PET	USA C30 ITQ
11	USA C45 CON	USA C74 BZS	JPN C34 MTR	JPN C40T41 EGW	JPN C30 ITQ
12	USA C85 HTH	CHN C27 MET	GBR C65T67 FIN	ROW C75 GOV	DEU C29 MEQ
13	DEU C15T16 FOD	USA C60T63 TRN	CHN C32 CMQ	CHN C24 CHM	DEU C34 MTR
14	ROW C60T63 TRN	GBR C65T67 FIN	CHN C24 CHM	USA C34 MTR	KOR C32 CMQ
15	USA C65T67 FIN	USA C23 PET	DEU C60T63 TRN	USA C24 CHM	USA C23 PET
16	GBR C50T52 WRT	GBR C74 BZS	FRA C50T52 WRT	CHN C30 ITQ	USA C74 BZS
17	DEU C24 CHM	USA C65T67 FIN	USA C50T52 WRT	CHN C23 PET	TWN C32 CMQ
18	DEU C29 MEQ	CHN C30 ITQ	CHN C50T52 WRT	ROW C60T63 TRN	CHN C27 MET
19	DEU C50T52 WRT	DEU C34 MTR	CHN C29 MEQ	CHN C29 MEQ	DEU C27 MET
20	DEU C27 MET	USA C30 ITQ	ROW C60T63 TRN	DEU C29 MEQ	GBR C74 BZS

Table 4. Top 20 ranks for global PageRank K , CheiRank K^* , 2DRank K_2 , ImportRank K and ExportRank K^* for the year 2008.

K_i	$ \psi_i $	node						
1	0.037606	ROW C17T19 TEX	0.050431	ARG C34 MTR	0.054681	CHN C32 CMQ	0.052248	RUS C10T14 MIN
2	0.025695	CHN C17T19 TEX	0.049991	BRA C34 MTR	0.053306	KOR C32 CMQ	0.03948	SAU C10T14 MIN
3	0.021618	ITA C17T19 TEX	0.029753	JPN C34 MTR	0.053253	TWN C32 CMQ	0.026187	ROW C10T14 MIN
4	0.017075	USA C17T19 TEX	0.026592	DEU C34 MTR	0.027361	SGP C32 CMQ	0.022125	NOR C10T14 MIN
5	0.016216	CHN C32 CMQ	0.018372	THA C34 MTR	0.025189	MYS C32 CMQ	0.019764	USA C71 RMQ
6	0.013003	CHN C30 ITQ	0.01531	IDN C34 MTR	0.018824	USA C30 ITQ	0.013899	USA C50T52 WRT
7	0.010963	FRA C17T19 TEX	0.0093875	ROW C21T22 PAP	0.016965	PHL C32 CMQ	0.011638	ROW C29 MEQ
8	0.010175	TUR C17T19 TEX	0.0093382	DEU C15T16 FOD	0.01534	JPN C30 ITQ	0.010871	RUS C27 MET
9	0.010161	USA C75 GOV	0.0090288	USA C15T16 FOD	0.014664	GBR C65T67 FIN	0.0082943	DEU C29 MEQ
10	0.0099839	USA C65T67 FIN	0.0086552	USA C75 GOV	0.013713	CHN C30 ITQ	0.0082905	RUS C23 PET

Table 5. Top 10 values of 4 different eigenvectors from Fig.9, Fig. 10. The corresponding eigenvalues from left to right are $\lambda = 0.4993$ (red), $\lambda = 0.3746 + 0.0126i$ (green), $\lambda = 0.6256$ (blue) and $\lambda = -0.0001 + 0.1687i$ (magenta).

Top 100 historical figures of Wikipedia

The **top 100 historical figures of Wikipedia** were determined by researchers from the University of Toulouse in France using mathematical and statistical methods from the Wikipedia database, and published in two scientific papers. In the statistical respects this top 100 list is of differs from the historical, cultural and other type arguments used by such historians like Michael H. Hart. The various mathematical methods and results obtained by different groups are described below. In spite of the mathematical and statistical grounds of those approaches they have overlap of about 43 percent with the top 100 list of Hart. The distribution of top PageRank historical figures over world countries is shown in Fig.1.

Approaches of different groups

The early ranking of top people of Wikipedia was done on the basis of PageRank algorithm and HITS algorithm for English Wikipedia edition (2005) by F.Belloni and R.Bonato [1]. For top people of PageRank they found Jesus, Paul the Apostle, Saint Peter and for HITS George W. Bush, Adolf Hitler, Bill Clinton.

Later studies of Quantware group analyzed English Wikipedia edition

Aug 2009 using PageRank, CheiRank and 2DRank algorithms.^[2] The top persons found are: Napoleon, George W. Bush, Elizabeth II for PageRank; Michael Jackson, Frank Lloyd Wright, David Bowie for 2DRank; Kasey S. Pipes, Roger Calmel, Yury G. Chernavsky for CheiRank. For this study the distributions of top 100 historical figures of PageRank, CheiRank and Hart's list are shown in Fig.2.

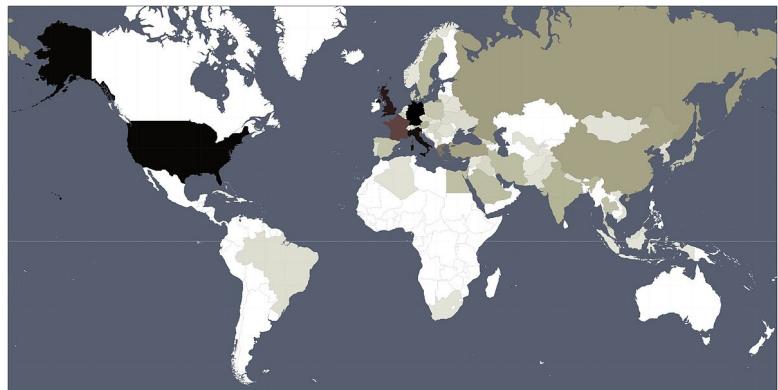


Fig1. World map of top 100 historical figures of 24 Wikipedia editions from PageRank for the period from BC 15th century to AD 20th century (darkness is proportional to a number of appearances of top persons born in a given country of AD 21st century geographical borders). After.

Time evolution of Wikipedia ranking of historical figures was investigated for English Wikipedia editions for 2003 - 2011 using the approach developed for Wikipedia Aug 2009.^[3] The distribution over fields of human activity was established there for various years.

Independently, a study with 15 largest Wikipedia language editions was done by Barcelona Media group.^[4] This group considered network of links between biographical articles of Wikipedia. However, a number of such biographical articles is relatively small compared to the total number of articles of a given edition that led to fluctuating ranking results.

The investigations of 9 Wikipedia editions have been reported by Eom and Shepelyansky producing a reliable ranking of top 30 persons for each edition. However, a selection of historical figures from the whole list of ranked edition articles was done manually that was restricting efficiency of the approach.

In parallel, the Stony-Brook group performed ranking of English Wikipedia edition combining PageRank method with other methods.^[5] This group found the top figures: Jesus, Napoleon, Muhammad. However, even if this group used the public Wikipedia database the whole list of their top 100 people is not publicly available.

The Pantheon MIT project produced the ranking list of top 100 persons using all language editions of Wikipedia counting number of editions and clicks on an article about a given person.^[6] This group found at the top: Aristotle, Plato, Jesus.

A list of the top 100 historical figures was created from Wikipedia pages in 24 different languages, using computer algorithms to analyze the importance of people based on the links to those people's pages.^{[1][7]}

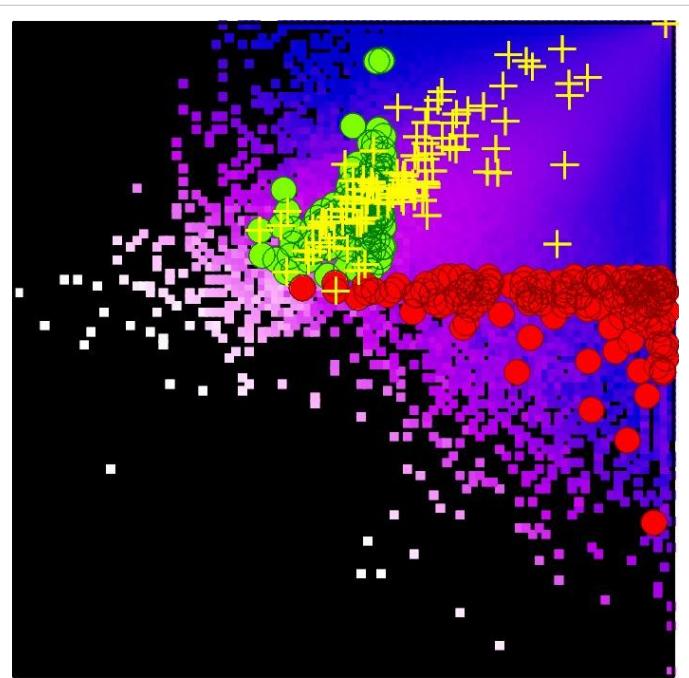


Fig2. Density distribution of Wikipedia English articles (Aug 2009) in the plane of PageRank and CheiRank indexes
 $0 < \ln K, \ln K^* < \ln N$ shown by color with blue for minimum and white for maximum (black for zero); green/red points show top 100 personalities from PageRank/CheiRank, yellow pluses show top 100 personalities from Hart's book, number of articles $N = 3282257$.
 After.

Ranking Methodology

The researchers used several different page-ranking algorithms, including Google PageRank, 2DRank, and CheiRank. They retrieved data from the text of Wikipedia pages in the 24 languages, and applied the algorithms to the data to create culturally-specific list of influential people, as well as a list across all the cultures examined in the project.

Among the data elements specifically targeted as indicators of importance were each person's birth country, date of birth, century of birth, and quantity of hyperlinks. In the case of hyperlinks for people's Wikipedia pages, both links to a person's page and links from that person's page were included in the analysis.

Other methods are described at and, they are not directly related to link analysis, network theory and Markov chains.

Results

For the global list of 24 editions of Wikipedia, the top 10 historical figures, identified by averaging over PageRank lists, were as follows:

1. Carl Linnaeus
2. Jesus
3. Aristotle
4. Napoleon
5. Hitler
6. Julius Caesar
7. Plato
8. William Shakespeare
9. Albert Einstein
10. Elizabeth II

The top global persons of 2DRank are Adolf Hitler, Michael Jackson, Madonna (entertainer). The top women of human history are Elizabeth II, Mary (mother of Jesus), Queen Victoria for PageRank list and Madonna (entertainer), Elizabeth II, Mary (mother of Jesus) for 2DRank list. Top 100 historical figures for 24 Wikipedia editions are available at.^[8] The overlap of top 100 people of Quantware, Stony-Brook, MIT Pantheon groups with the Hart list is found to be on a level of 42-44 percents. This shows that the mathematical methods of determination of top 100 historical figures of humanity via Wikipedia database give the reliable results.

Discussion of Wikipedia ranking of historical figures in public press can be found at.

References

- [1] Belloni F and Bonato R, "Network analysis for Wikipedia . Proceedings of Wikimania 2005, The First International Wikimedia Conference. Frankfurt, Germany. Retrieved September 16, 2006" (<http://www.fran.it/blog/2005/08/network-analysis-for-wikipedia.html>)
- [2] Zhirov A O, Zhirov O V and Shepelyansky D L, "Two-dimensional ranking of Wikipedia articles, Eur. J. Phys. B 77: 523 (2010); DOI: 10.1140/epjb/e2010-10500-7" (<http://www.quantware.ups-tlse.fr/QWLIB/2drankwikipedia/>)
- [3] Eom Y-H, Frahm K M, Benczur A, and Shepelyansky D L, "Time evolution of Wikipedia network ranking, Eur. Phys. J. B. 86: 492 (2013); DOI: 10.1140/epjb/e2013-40432-5" (<http://www.quantware.ups-tlse.fr/QWLIB/wikirankevolution/>)
- [4] Aragon P, Kaltenbrunner A, Laniado D, and Volkovich Y, "Biographical Social Networks on Wikipedia - A cross-cultural study of links that made history, Proceedings of WikiSym, 2012 ArXiv:1204.3799(cs.SI) (2012)" (<http://arxiv.org/abs/1204.3799>)
- [5] Skiena S, and Charles W, "Who's Bigger?: Where Historical Figures Really Rank. Cambridge University Press. ISBN 978-110704137" (<http://www.maa.org/publications/maa-reviews/whos-bigger-where-historical-figures-really-rank>)
- [6] The Pantheon MIT project "The Pantheon MIT project" (<http://pantheon.media.mit.edu>)
- [7] "Wikipedia Mining Algorithm Reveals The Most Influential People In 35 Centuries Of Human History" (<https://medium.com/the-physics-arxiv-blog/wikipedia-mining-algorithm-reveals-the-most-influential-people-in-35-centuries-of-human-history-ede5ef827b76>)
- [8] Top people of Wikipedia "Top people of Wikipedia database" (<http://www.quantware.ups-tlse.fr/QWLIB/topwikipeople/>)

External links

- The Anatomy of a Search Engine, describes the PageRank algorithm developed by Google founders Sergey Brin and Lawrence Page (<http://infolab.stanford.edu/~backrub/google.html>)
- Dima Shepelyansky, Laboratoire de Physique Théorique, Université Paul Sabatier, France (<http://www.quantware.ups-tlse.fr/dima>)

Article Sources and Contributors

Top 100 historical figures of Wikipedia *Source:* <http://en.wikipedia.org/w/index.php?oldid=616877760> *Contributors:* Animalparty, Frmorrison, HarryBoston, OccultZone, Roscelese, Shepelyansky, Spinningspark, Topbanana

Image Sources, Licenses and Contributors

Image:PageRank24x100.jpg *Source:* <http://en.wikipedia.org/w/index.php?title=File:PageRank24x100.jpg> *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Sfan00 IMG, Shepelyansky

Image:CheiRank3.jpg *Source:* <http://en.wikipedia.org/w/index.php?title=File:CheiRank3.jpg> *License:* Public Domain *Contributors:* Shepelyansky

License

Creative Commons Attribution-Share Alike 3.0
[/creativecommons.org/licenses/by-sa/3.0/](http://creativecommons.org/licenses/by-sa/3.0/)

Convergence of rank based degree-degree correlations in random directed networks

Pim van der Hoorn* Nelly Litvak†

October 31, 2014

Abstract

We introduce, and analyze, three measures for degree-degree dependencies, also called degree assortativity, in directed random graphs, based on Spearman’s rho and Kendall’s tau. We proof statistical consistency of these measures in general random graphs and show that the directed Configuration Model can serve as a null model for our degree-degree dependency measures. Based on these results we argue that the measures we introduce should be preferred over Pearson’s correlation coefficients, when studying degree-degree dependencies, since the latter has several issues in the case of large networks with scale-free degree distributions.

Keywords: Degree-degree dependencies, rank correlations, directed random graphs, directed configuration model, Spearman’s rho, Kendall’s tau

1 Introduction

This paper investigates statistical consistency of rank correlation measures for dependencies between in- and/or out-degrees on both sides of a randomly sampled edge in large directed networks, such as the World Wide Web, Wikipedia, or Twitter. These dependencies, also called the assortativity of the network, degree correlations, or degree-degree dependencies, represent an important topological property of real-world networks, and they have received a vast attention in the literature, starting with the work of Newman [12, 13].

The underlying question that motivates analysis of degree-degree dependencies is whether nodes of high in- or out-degree are more likely to be connected to nodes of high or low in- or out-degree. These dependencies have been shown to influence many topological features of networks, among others, behavior of epidemic spreading [1], social consensus in Twitter [9], stability of P2P networks under attack [15] and network observability [6]. Therefore, being able to properly measure degree-degree dependencies is essential in modern network analysis.

Given a network, represented by a directed graph, a measurement of degree-degree dependency usually consists of computing some expression that is defined by the degrees at both sides of the edges. Here the value on each edge can be seen as a realization of some unknown ‘true’ parameter that characterizes the degree-degree dependency.

Currently, the most commonly used measure for degree-degree dependencies is a so-called *assortativity coefficient*, introduced in [12, 13], that computes Pearson’s correlation coefficient for the degrees at both sides of an edge. However, this dependency measure suffers from the fact that most real-world networks have highly skewed degree distributions, also called *scale-free* distributions, formally described by power laws, or more formally, regularly varying distributions. Indeed, when the (in- or out-) degree at the end of a random edge has infinite variance, then Pearson’s coefficient is ill-defined. As a result, the dependency measure suggested in [12, 13] depends on the graph size and converges to a non-negative number in the infinite network size

*University of Twente, w.l.f.vanderhoorn@utwente.nl

†University of Twente, n.litvak@utwente.nl

limit, as was pointed out in several papers [5, 8]. The detailed mathematical analysis and examples for undirected graphs have been given in [7], and for directed graphs in our recent work [17]. Thus, Pearson’s correlation coefficient is not suitable for measuring degree-degree dependencies in most real-world directed networks.

The fact that the most commonly used degree correlation measure has obvious mathematical flaws, motivates for design and analysis of new estimators. Despite the importance of degree-degree dependencies and vast interest from the research community, this remains a largely open problem.

In [7] it was suggested to use a rank correlation measure, Spearman’s rho, and it was proved that under general regularity conditions, this measure indeed converges to its correct population value. Both configuration model and preferential attachment model [16] were proved to satisfy these conditions. In [17] we proposed three rank correlation measures, based on Spearman’s rho and Kendall’s tau, as defined for integer valued random variables, cf. [11], and we compared these measures to Pearson’s correlation coefficient on Wikipedia graphs for nine different languages.

In this paper we first prove that, under the convergence assumption of the empirical two-dimensional distributions of the degrees on both sides of a random edge, the rank correlations defined in [17] are indeed statistically consistent estimators of degree-degree dependencies. We obtain their limiting values in terms of the limiting distributions of the degrees.

Next, we apply our results to the recently developed directed Configuration Model [2]. Roughly speaking, in this model, each node is given a random number of in- and out-bound stubs, that are subsequently connected to each other at random. Since multiple edges and self-loops may appear as a result of such random wiring, [2] presents two versions of the directed Configuration Model. The *repeated* version repeats the wiring until the resulting graph is simple, while the *erased* version merges multiple edges and removes self-loops to obtain a simple graph.

We analyze our suggested rank correlation measures in the Repeated and Erased Configuration Model, as described in [2], and prove that all three measures converge to zero in both models. This result is not very surprising for the repeated model, since we connect vertices uniformly at random. However, in the erased scenario, the graph is made simple by design, and this might contribute to the network showing negative degree-degree dependencies as observed and discussed in, for instance, [10, 14]. Our result shows that such negative degree-degree dependencies vanish for sufficiently large graphs, and thus both flavors of the directed Configuration Model can be used as ‘null model’ for our three rank correlation measures.

By proving consistency of three estimators for degree-degree dependencies in directed networks, and providing an easy-to-construct null model for these estimators, this paper makes an important step towards assessing statistical significance of degree-degree dependencies in a mathematically rigorous way.

This paper is structured as follows. In Section 2 we introduce notations, used throughout this paper. Then, in Section 3, we prove a general theorem concerning statistical consistency of estimators for Spearman’s rho and Kendall’s tau on integer-valued data. This result is applied in Section 4 in the setting of random graphs to prove the convergence in the infinite size graph limit of the three degree-degree dependency measures from [17], based on Spearman’s rho and Kendall’s tau. We analyze both the Repeated and Erased Directed Configuration Model in Section 5.

2 Notations and definitions

Throughout the paper, if X and Y are random variables we denote their distribution functions by F_X and F_Y , respectively, and their joint distribution by $H_{X,Y}$. For integer valued random variables X, Y and $k, l \in \mathbb{Z}$ we will often use the following notations:

$$\mathcal{F}_X(k) = F_X(k) + F_X(k-1), \quad (1)$$

$$\mathcal{H}_{X,Y}(k, l) = H_{X,Y}(k, l) + H_{X,Y}(k-1, l) + H_{X,Y}(k, l-1) + H_{X,Y}(k-1, l-1). \quad (2)$$

If Z is a random element, we define the function $F_{X|Z} : \mathbb{R} \times \Omega \rightarrow [0, 1]$ by

$$F_{X|Z}(x, \omega) = \mathbb{E}[I\{X \leq x\}|Z](\omega),$$

where $I\{X \leq x\}$ denotes the indicator of the event $\{\omega : X(\omega) \leq x\}$. We furthermore define the random variable $F_{X|Z}(Y)$ by

$$F_{X|Z}(Y)(\omega) = F_{X|Z}(Y(\omega), \omega),$$

and we write $F_{X|Z}(x)$ to indicate the random variable $\mathbb{E}[I\{X \leq x\}|Z]$. With these notations it follows that if X' is an independent copy of X , then

$$\begin{aligned} \mathbb{E}[I\{X' \leq X\}|Z] &= \int_{\mathbb{R}} \int_{\mathbb{R}} I\{z \leq x\} d\mathbb{P}(z|Z) d\mathbb{P}(x|Z) \\ &= \int_{\mathbb{R}} \mathbb{E}[I\{X' \leq x\}|Z] d\mathbb{P}(x|Z) \\ &= \mathbb{E}[F_{X|Z}(X)|Z]. \end{aligned}$$

Using similar definitions for $H_{X,Y|Z}(x, y, \omega)$ and $H_{X,Y|Z}(X, Y)$ we get, if (X', Y') and (X'', Y'') are independent copies of (X, Y) , that

$$\mathbb{E}[I\{X' \leq X\} I\{Y'' \leq Y\}|Z] = \mathbb{E}[H_{X,Y|Z}(X, Y)|Z].$$

For integer valued random variables X and Y , the random variables $\mathcal{F}_{X|Z}(k)$ and $\mathcal{H}_{X,Y|Z}(k, l)$ are defined similarly to (1) and (2), using $F_{X|Z}(k)$ and $H_{X,Y|Z}(k, l)$, respectively.

We introduce the following notion of convergence, related to convergence in distribution.

Definition 2.1. Let $\{X_n\}_{n \in \mathbb{N}}$ and X be random variables and $\{Z_n\}_{n \in \mathbb{N}}$ be a sequence of random elements. We say that X_n converges in distribution to X conditioned on Z_n and write

$$(X_n|Z_n) \Rightarrow X \quad \text{as } n \rightarrow \infty$$

if and only if for all continuous, bounded $h : \mathbb{R} \rightarrow \mathbb{R}$

$$\mathbb{E}[h(X_n)|Z_n] \xrightarrow{\mathbb{P}} \mathbb{E}[h(X)] \quad \text{as } n \rightarrow \infty.$$

Here $\xrightarrow{\mathbb{P}}$ denotes convergence in probability. Note that if h is bounded then $\mathbb{E}[h(X_n)|Z_n]$ is bounded almost everywhere, hence $\lim_{n \rightarrow \infty} \mathbb{E}[h(X_n)] = \lim_{n \rightarrow \infty} \mathbb{E}[\mathbb{E}[h(X_n)|Z_n]] = \mathbb{E}[h(X)]$. Therefore, $(X_n|Z_n) \Rightarrow X$ implies that $X_n \Rightarrow X$, where we write \Rightarrow for convergence in distribution. Similar to convergence in distribution, it holds that Definition 2.1 is equivalent to

$$F_{X_n|Z_n}(k) \xrightarrow{\mathbb{P}} F_X(k) \quad \text{as } n \rightarrow \infty, \quad \text{for all } k \in \mathbb{Z}.$$

In this paper we use a continuization principle, applied for instance in [11], where we transform given discrete random variables in continuous ones. From here on we will work with integer valued random variables instead of arbitrary discrete random variables.

Definition 2.2. Let X be an integer valued random variable and U a uniformly distributed random variable on $[0, 1)$ independent of X . Then we define the continuization of X as

$$\tilde{X} = X + U.$$

We will refer to U as the *continuous part* of \tilde{X} . We remark that although we have chosen U to be uniform we could instead take any continuous random variable on $[0, 1)$ with strictly increasing cdf, cf. [4].

3 Rank correlations for integer valued random variables

We will use the rank correlations Spearman's rho and Kendall's tau for integer valued random variables as defined in [11]. Below we will state these and rewrite them in terms of the functions \mathcal{F} and \mathcal{H} , defined in (1) and (2) respectively. We will then proceed, defining estimators for these correlations and prove that, under natural conditions, these converge to the correct value.

3.1 Spearman's rho

Given two integer valued random variables X and Y , Spearman's rho $\rho(X, Y)$ is defined as, c.f. [11]

$$\begin{aligned}\rho(X, Y) = 3 & (\mathbb{P}(X < X', Y < Y'') + \mathbb{P}(X \leq X', Y < Y'') \\ & + \mathbb{P}(X < X', Y \leq Y'') + \mathbb{P}(X \leq X', Y \leq Y'') - 1),\end{aligned}$$

where (X', Y') and (X'', Y'') are independent copies of (X, Y) . We will rewrite this expression, starting with a single term:

$$\begin{aligned}\mathbb{P}(X < X', Y < Y'') &= \mathbb{E}[I\{X < X'\} I\{Y < Y''\}] \\ &= 1 - \mathbb{E}[I\{X' \leq X\}] - \mathbb{E}[I\{Y'' \leq Y\}] + \mathbb{E}[I\{X' \leq X\} I\{Y'' \leq Y\}] \\ &= 1 - \mathbb{E}[F_X(X)] - \mathbb{E}[F_Y(Y)] + \mathbb{E}[F_X(X)F_Y(Y)].\end{aligned}$$

If we do the same for the other three terms and use (57) we obtain,

$$\rho(X, Y) = 3\mathbb{E}[\mathcal{F}_X(X)\mathcal{F}_Y(Y)] - 3. \quad (3)$$

Since, given two continuous random variables \mathcal{X} and \mathcal{Y} , Spearman's rho is defined as

$$\rho(\mathcal{X}, \mathcal{Y}) = 12\mathbb{E}[F_{\mathcal{X}}(\mathcal{X})F_{\mathcal{Y}}(\mathcal{Y})] - 3,$$

Lemma A.3 now implies that

$$\rho(X, Y) = \rho(\tilde{X}, \tilde{Y}). \quad (4)$$

3.2 Kendall's tau

For two continuous random variables \mathcal{X} and \mathcal{Y} , Kendall's tau $\tau(\mathcal{X}, \mathcal{Y})$ is defined as

$$\tau(\mathcal{X}, \mathcal{Y}) = 4\mathbb{E}[H_{\mathcal{X}, \mathcal{Y}}(\mathcal{X}, \mathcal{Y})] - 1.$$

Given two discrete random variables X and Y , Kendall's Tau can be written as, c.f. [11] Proposition 2.2,

$$\tau(X, Y) = \mathbb{E}[\mathcal{H}_{X, Y}(X, Y)] - 1. \quad (5)$$

Similar to Spearman's rho we obtain, using Lemma A.3, that

$$\tau(X, Y) = \tau(\tilde{X}, \tilde{Y}). \quad (6)$$

Hence applying the continuization principle from Definition 2.2 on X and Y preserves both rank correlations. We remark that (4) and (6) were obtained for arbitrary discrete random variables, using a different approach, in [11].

3.3 Convergence for Spearman's rho and Kendall's tau

Let $\{X_n\}_{n \in \mathbb{N}}$ and $\{Y_n\}_{n \in \mathbb{N}}$ be sequences of integer valued random variables. If $(X_n, Y_n) \Rightarrow (X, Y)$, for some integer valued random variables X and Y , then $\lim_{n \rightarrow \infty} \mathbb{E}[\mathcal{F}_{X_n}(X_n)\mathcal{F}_{Y_n}(Y_n)] = \mathbb{E}[\mathcal{F}_X(X)\mathcal{F}_Y(Y)]$ which implies that $\lim_{n \rightarrow \infty} \rho(X_n, Y_n) = \rho(X, Y)$. The next theorem generalizes this to the setting of the convergence of $(X_n, Y_n | Z_n)$, of Definition 2.1.

Theorem 3.1. Let $\{X_n\}_{n \in \mathbb{N}}$, $\{Y_n\}_{n \in \mathbb{N}}$ be sequences of integer valued random variables for which there exist a sequence $\{Z_n\}_{n \in \mathbb{N}}$ of random elements and two integer valued random variables X and Y such that

$$(X_n, Y_n | Z_n) \Rightarrow (X, Y) \quad \text{as } n \rightarrow \infty.$$

Then, as $n \rightarrow \infty$,

$$\text{i) } 3\mathbb{E} [\mathcal{F}_{X_n|Z_n}(X_n)\mathcal{F}_{Y_n|Z_n}(Y_n) | Z_n] - 3 \xrightarrow{\mathbb{P}} \rho(X, Y) \text{ and}$$

$$\text{ii) } \mathbb{E} [\mathcal{H}_{X_n, Y_n | Z_n}(X_n, Y_n) | Z_n] - 1 \xrightarrow{\mathbb{P}} \tau(X, Y).$$

Moreover, we also have convergence of the expectations:

$$\text{iii) } \lim_{n \rightarrow \infty} 3\mathbb{E} [\mathcal{F}_{X_n|Z_n}(X_n)\mathcal{F}_{Y_n|Z_n}(Y_n)] - 3 = \rho(X, Y) \text{ and}$$

$$\text{iv) } \lim_{n \rightarrow \infty} \mathbb{E} [\mathcal{H}_{X_n, Y_n | Z_n}(X_n, Y_n)] - 1 = \tau(X, Y).$$

Proof. Observe first that since $(X_n, Y_n | Z_n) \Rightarrow (X, Y)$, it follows that for all $k, l \in \mathbb{Z}$, as $n \rightarrow \infty$,

$$F_{X_n|Z_n}(k) \xrightarrow{\mathbb{P}} F_X(k) \tag{7}$$

$$F_{Y_n|Z_n}(l) \xrightarrow{\mathbb{P}} F_Y(l) \tag{8}$$

$$H_{X_n, Y_n | Z_n}(k, l) \xrightarrow{\mathbb{P}} H_{X, Y}(k, l). \tag{9}$$

Moreover, these convergence hold uniformly, since X and Y are integer valued.

i) Using first (3) and then applying Lemma A.3 and Proposition A.4 we obtain,

$$\begin{aligned} & |3\mathbb{E} [\mathcal{F}_{X_n|Z_n}(X_n)\mathcal{F}_{Y_n|Z_n}(Y_n) | Z_n] - 3 - \rho(X, Y)| \\ &= 3 |\mathbb{E} [\mathcal{F}_{X_n|Z_n}(X_n)\mathcal{F}_{Y_n|Z_n}(Y_n) | Z_n] - \mathbb{E} [\mathcal{F}_X(X)\mathcal{F}_Y(Y)]| \\ &= 12 \left| \mathbb{E} [F_{\tilde{X}_n|Z_n}(\tilde{X}_n)F_{\tilde{Y}_n|Z_n}(\tilde{Y}_n) | Z_n] - \mathbb{E} [F_{\tilde{X}}(\tilde{X})F_{\tilde{Y}}(\tilde{Y})] \right| \\ &\leq 12 \left| \mathbb{E} [F_{\tilde{X}_n|Z_n}(\tilde{X}_n)F_{\tilde{Y}_n|Z_n}(\tilde{Y}_n) | Z_n] - \mathbb{E} [F_{\tilde{X}}(\tilde{X}_n)F_{\tilde{Y}}(\tilde{Y}_n) | Z_n] \right| \\ &\quad + 12 \left| \mathbb{E} [F_{\tilde{X}}(\tilde{X}_n)F_{\tilde{Y}}(\tilde{Y}_n) | Z_n] - \mathbb{E} [F_{\tilde{X}}(\tilde{X})F_{\tilde{Y}}(\tilde{Y})] \right| \\ &\leq 12 \sup_{x, y \in \mathbb{R}} |F_{\tilde{X}_n|Z_n}(x)F_{\tilde{Y}_n|Z_n}(y) - F_{\tilde{X}}(x)F_{\tilde{Y}}(y)| \tag{10} \end{aligned}$$

$$+ 12 \left| \mathbb{E} [F_{\tilde{X}}(\tilde{X}_n)F_{\tilde{Y}}(\tilde{Y}_n) | Z_n] - \mathbb{E} [F_{\tilde{X}}(\tilde{X})F_{\tilde{Y}}(\tilde{Y})] \right|. \tag{11}$$

Because the function $h(x, y) = F_{\tilde{X}}(x)F_{\tilde{Y}}(y)$ is continuous and bounded, (11) converges in probability to 0. For (10) we observe that

$$\begin{aligned} |F_{\tilde{X}_n|Z_n}(x)F_{\tilde{Y}_n|Z_n}(y) - F_{\tilde{X}}(x)F_{\tilde{Y}}(y)| &\leq |F_{\tilde{X}_n|Z_n}(x)F_{\tilde{Y}_n|Z_n}(y) - F_{\tilde{X}_n|Z_n}(x)F_{\tilde{Y}}(y)| \\ &\quad + |F_{\tilde{X}_n|Z_n}(x)F_{\tilde{Y}}(y) - F_{\tilde{X}}(x)F_{\tilde{Y}}(y)| \\ &\leq |F_{\tilde{Y}_n|Z_n}(y) - F_{\tilde{Y}}(y)| + |F_{\tilde{X}_n|Z_n}(x) - F_{\tilde{X}}(x)|. \end{aligned}$$

It now follows that (10) converges in probability to 0, since the convergence (7) and (8) are uniform.

ii) Here we again use Lemma A.3 and Proposition A.4, now combined with (5) to obtain,

$$\begin{aligned} & |\mathbb{E} [\mathcal{H}_{X_n, Y_n | Z_n}(X_n, Y_n) | Z_n] - 1 - \tau(X, Y)| \\ &= |\mathbb{E} [\mathcal{H}_{X_n, Y_n | Z_n}(X_n, Y_n) | Z_n] - \mathbb{E} [\mathcal{H}_{X, Y}(X, Y)]| \\ &= 4 \left| \mathbb{E} [H_{\tilde{X}_n, \tilde{Y}_n | Z_n}(\tilde{X}_n, \tilde{Y}_n) | Z_n] - \mathbb{E} [H_{\tilde{X}, \tilde{Y}}(\tilde{X}, \tilde{Y})] \right| \end{aligned}$$

$$\begin{aligned}
&\leq 4 \left| \mathbb{E} \left[H_{\tilde{X}_n, \tilde{Y}_n | Z_n}(\tilde{X}_n, \tilde{Y}_n) \middle| Z_n \right] - \mathbb{E} \left[H_{\tilde{X}, \tilde{Y}}(\tilde{X}_n, \tilde{Y}_n) \middle| Z_n \right] \right| \\
&\quad + 4 \left| \mathbb{E} \left[H_{\tilde{X}, \tilde{Y}}(\tilde{X}_n, \tilde{Y}_n) \middle| Z_n \right] - \mathbb{E} \left[H_{\tilde{X}, \tilde{Y}}(\tilde{X}, \tilde{Y}) \right] \right| \\
&\leq 4 \sup_{x, y \in \mathbb{R}} \left| H_{\tilde{X}_n, \tilde{Y}_n | Z_n}(x, y) - H_{\tilde{X}, \tilde{Y}}(x, y) \right| + 4 \left| \mathbb{E} \left[H_{\tilde{X}, \tilde{Y}}(\tilde{X}_n, \tilde{Y}_n) \middle| Z_n \right] - \mathbb{E} \left[H_{\tilde{X}, \tilde{Y}}(\tilde{X}, \tilde{Y}) \right] \right|
\end{aligned}$$

The former term converges in probability to 0 because (9) holds uniformly, and for the latter this holds since $h(x, y) = H_{\tilde{X}, \tilde{Y}}(x, y)$ is continuous and bounded.

Since both $\mathbb{E} [\mathcal{F}_{X_n | Z_n}(X_n) \mathcal{F}_{Y_n | Z_n}(Y_n) | Z_n]$ and $\mathbb{E} [\mathcal{H}_{X_n, Y_n | Z_n}(X_n, Y_n) | Z_n]$ are bounded a.e. we obtain iii) and iv) directly from i) and ii), respectively. \square

4 Rank correlations for random graphs

We now turn to the setting of rank correlations for degree-degree dependencies in random directed graphs. We will first introduce some terminology concerning random graphs. Then we will recall the rank correlations given in [17] and prove statistical consistency of these measures.

4.1 Random graphs

Given a directed graph $G = (V, E)$, we denote by $(D^+(v), D^-(v))_{v \in V}$ the degree sequence where D^+ denotes the out-degree and D^- the in-degree. We adopt the convention, introduced in [17], to index the degree type by $\alpha, \beta \in \{+, -\}$. Furthermore, we will use the projections $\pi_*, \pi^* : V^2 \rightarrow V$ to distinguish the source and target of a possible edge. That is, if $(v, w) \in V^2$ then $\pi_*(v, w) = v$ and $\pi^*(v, w) = w$. When both projections are applicable we will use π . For $v, w \in V$ we denote by $E(v, w) = \{e \in E | \pi_* e = v, \pi^* e = w\}$ the set of all edges from v to w . For $e \in V^2$, we write $E(e) = E(\pi_* e, \pi^* e)$.

Given a set V of vertices we call a graph $G = (V, E)$ random, if for each $e \in V^2$, $|E(e)|$ is a random variable. Since $I\{e \in E\} = I\{|E(e)| > 0\}$, it follows that the former is also a random variable, cf. [3] for a similar definition of random graphs using edge indicators. Therefore, when we refer to G as a random element it is understood that we refer to the random variables $|E(e)|$, for $e \in V^2$.

When G is a random graph, the number of edges in the graph and the degrees of the nodes are random variables defined by $I\{e \in E\}$ and $|E(e)|$, $e \in V^2$:

$$\begin{aligned}
|E| &= \sum_{e \in V^2} I\{e \in E\} |E(e)|, \\
D^-(v) &= \sum_{w \in V} I\{(w, v) \in E\} |E(w, v)|, \quad v \in V, \\
D^+(v) &= \sum_{w \in V} I\{(v, w) \in E\} |E(v, w)|, \quad v \in V.
\end{aligned}$$

Given a random graph $G = (V, E)$ we define a uniformly sampled edge \mathcal{E}_G as a two-dimensional random variable on V^2 such that

$$\mathbb{P}(\mathcal{E}_G = e | G) = \frac{|E(e)|}{|E|}.$$

When it is clear which graph we are considering, we will use \mathcal{E} instead of \mathcal{E}_G . Let $\alpha, \beta \in \{+, -\}$, $k, l \in \mathbb{N}$ and π be any of the projections π_* and π^* . Then we define

$$F_G^\alpha(k) = F_{D^\alpha(\pi(\mathcal{E}_G)) | G}(k), \tag{12}$$

$$H_G^{\alpha, \beta}(k, l) = H_{D^\alpha(\pi_*(\mathcal{E}_G)), D^\beta(\pi^*(\mathcal{E}_G)) | G}(k, l). \tag{13}$$

These functions are the empirical distribution of $D^\alpha(\pi(\mathcal{E}_G))$ and the joint empirical distribution of $D^\alpha(\pi_*(\mathcal{E}_G))$ and $D^\beta(\pi^*(\mathcal{E}_G))$, respectively, given the random graph G . The functions \mathcal{F}_G^α and $\mathcal{H}_G^{\alpha,\beta}$ are defined in a similar way as (1) and (2), using (12) and (13), respectively. In order to keep notations clear, we will, when considering both projections π_* and π^* , always use α to index the degree type of the sources and β to index the degree type of targets. Moreover, we will often write $D^\alpha \pi \mathcal{E}_G$ instead of $D^\alpha(\pi(\mathcal{E}_G))$.

Now we will introduce Spearman's rho and Kendall's tau on random directed graphs and write them in terms of the functions (12) and (13). This way we will be in a setting similar to the one of Theorem 3.1 so that we can utilize this theorem to prove statistical consistency of these rank correlations.

4.2 Spearman's Rho

Spearman's rho measure for degree-degree dependencies in directed graphs, introduced in [17], is in fact Pearson's correlation coefficient computed on the ranks of the degrees rather than their actual values. In our setting, this definition is ambiguous because the data has many ties. For example, if the in-degree of node v is d then we will observe $D^- \pi^* e = d$ for at least d edges $e \in E$, plus there will be many more nodes with the same degree. In [17] we consider two possible ways of resolving ties: by assigning a unique rank to each tied value uniformly at random, and by assigning the same, average, rank to all tied values. We denote the ranks resulting from the random and the average resolution of ties by R and \bar{R} , respectively. Formally, for $\alpha, \beta \in \{+, -\}$, we write:

$$R^\alpha \pi_* e = \sum_{f \in E} I \{ D^\alpha \pi_* f + U_f \geq D^\alpha \pi_* e + U_e \}, \quad (14)$$

$$R^\beta \pi^* e = \sum_{f \in E} I \{ D^\beta \pi^* f + W_f \geq D^\beta \pi^* e + W_e \}, \quad (15)$$

where U, W are independent $|V|^2$ vectors of independent uniform random variables on $[0, 1)$, and

$$\bar{R}^\alpha \pi e = \frac{1}{2} + \sum_{f \in E} I \{ D^\alpha \pi f > D^\alpha \pi e \} + \frac{1}{2} I \{ D^\alpha \pi f = D^\alpha \pi e \}. \quad (16)$$

Then the corresponding two versions of Spearman's rho are defined as follows, cf. [17]:

$$\begin{aligned} \rho_\alpha^\beta(G) &= \frac{12 \sum_{e \in E} R^\alpha \pi_*(e) R^\beta \pi^*(e) - 3|E|(|E|+1)^2}{|E|^3 - |E|} \quad \text{and} \\ \bar{\rho}_\alpha^\beta(G) &= \frac{4 \sum_{e \in E} \bar{R}^\alpha \pi_*(e) \bar{R}^\beta \pi^*(e) - |E|(|E|+1)^2}{\text{Var}_*(\bar{R}^\alpha) \text{Var}^*(\bar{R}^\beta)}, \end{aligned}$$

where

$$\begin{aligned} \text{Var}_*(\bar{R}^\alpha) &= \sqrt{4 \sum_{e \in E} \bar{R}^\alpha \pi_*(e)^2 - |E|(|E|+1)^2} \quad \text{and} \\ \text{Var}^*(\bar{R}^\beta) &= \sqrt{4 \sum_{e \in E} \bar{R}^\beta \pi^*(e)^2 - |E|(|E|+1)^2}. \end{aligned}$$

The next proposition relates the random variables $\rho_\alpha^\beta(G)$ and $\bar{\rho}_\alpha^\beta(G)$ to the random variable

$$\mathbb{E} \left[\mathcal{F}_G^\alpha(D^\alpha \pi_* \mathcal{E}) \mathcal{F}_G^\beta(D^\beta \pi^* \mathcal{E}) \mid G \right]. \quad (17)$$

Proposition 4.1. *Let $G = (V, E)$ be a random graph, \mathcal{E} an edge on G sampled uniformly at random and $\alpha, \beta \in \{+, -\}$. Then*

- i) $\frac{1}{|E|} \sum_{e \in E} \frac{\overline{R}^\alpha \pi_* e}{|E|} \frac{\overline{R}^\beta \pi^* e}{|E|} = \frac{1}{4} \mathbb{E} \left[\mathcal{F}_G^\alpha (D^\alpha \pi_* \mathcal{E}) \mathcal{F}_G^\beta (D^\beta \pi^* \mathcal{E}) \middle| G \right] + o_{\mathbb{P}}(|E|^{-1}) \quad \text{and}$
- ii) $\frac{1}{|E|} \sum_{e \in E} \frac{R^\alpha \pi_* e}{|E|} \frac{R^\beta \pi^* e}{|E|} = \frac{1}{4} \mathbb{E} \left[\mathcal{F}_G^\alpha (D^\alpha \pi_* \mathcal{E}) \mathcal{F}_G^\beta (D^\beta \pi^* \mathcal{E}) \middle| G \right] + o_{\mathbb{P}}(|E|^{-1}).$

Proof. i) Let \mathcal{E}' be an independent copy of \mathcal{E} and $e \in V^2$. Then it follows from (16) that

$$\begin{aligned}
\frac{\overline{R}^\alpha \pi e}{|E|} &= \frac{1}{2|E|} + \sum_{f \in E} \frac{1}{|E|} I\{D^\alpha \pi f > D^\alpha \pi e\} + \frac{1}{2|E|} I\{D^\alpha \pi f = D^\alpha \pi e\} \\
&= 1 + \frac{1}{2|E|} - \frac{1}{2|E|} \sum_{f \in E} I\{D^\alpha \pi f \leq D^\alpha \pi e\} + I\{D^\alpha \pi f \leq D^\alpha \pi e - 1\} \\
&= 1 + \frac{1}{2|E|} - \frac{1}{2} \sum_{f \in V^2} (I\{D^\alpha \pi f \leq D^\alpha \pi e\} + I\{D^\alpha \pi f \leq D^\alpha \pi e - 1\}) \frac{|E(f)|}{|E|} \\
&= 1 + \frac{1}{2|E|} - \frac{1}{2} \sum_{f \in V^2} (I\{D^\alpha \pi f \leq D^\alpha \pi e\} + I\{D^\alpha \pi f \leq D^\alpha \pi e - 1\}) \mathbb{P}(\mathcal{E}' = f | G) \\
&= 1 + \frac{1}{2|E|} - \frac{1}{2} (F_G^\alpha (D^\alpha \pi e) + F_G^\alpha (D^\alpha \pi e - 1)) \\
&= 1 + \frac{1}{2|E|} - \frac{1}{2} \mathcal{F}_G^\alpha (D^\alpha \pi e).
\end{aligned} \tag{18}$$

Using a similar expression for $(\overline{R}^\beta \pi_* e) / |E|$ we obtain,

$$\begin{aligned}
\frac{1}{|E|} \sum_{e \in E} \frac{\overline{R}^\alpha \pi_* e}{|E|} \frac{\overline{R}^\beta \pi^* e}{|E|} &= \frac{1}{|E|} \sum_{e \in E} \left(1 + \frac{1}{2|E|} - \frac{1}{2} \mathcal{F}_G^\alpha (D^\alpha \pi_* e) \right) \left(1 + \frac{1}{2|E|} - \frac{1}{2} \mathcal{F}_G^\beta (D^\beta \pi^* e) \right) \\
&= \mathbb{E} \left[\left(1 + \frac{1}{2|E|} - \frac{1}{2} \mathcal{F}_G^\alpha (D^\alpha \pi_* \mathcal{E}) \right) \left(1 + \frac{1}{2|E|} - \frac{1}{2} \mathcal{F}_G^\beta (D^\beta \pi^* \mathcal{E}) \right) \middle| G \right].
\end{aligned}$$

Rearranging the terms yields

$$\begin{aligned}
\frac{1}{|E|} \sum_{e \in E} \frac{\overline{R}^\alpha \pi_* e}{|E|} \frac{\overline{R}^\beta \pi^* e}{|E|} &= \frac{1}{4} \mathbb{E} \left[\mathcal{F}_G^\alpha (D^\alpha \pi_* \mathcal{E}) \mathcal{F}_G^\beta (D^\beta \pi^* \mathcal{E}) \middle| G \right] \\
&\quad + 1 - \frac{1}{2} \mathbb{E} \left[\mathcal{F}_G^\alpha (D^\alpha \pi_* \mathcal{E}) + \mathcal{F}_G^\beta (D^\beta \pi^* \mathcal{E}) \middle| G \right] + o_{\mathbb{P}}(|E|^{-1}).
\end{aligned} \tag{19}$$

Since the sum over all average ranks equals $|E|(|E| + 1)/2$, it follows that

$$\frac{1}{2} + \frac{1}{2|E|} = \frac{1}{|E|} \sum_{e \in E} \frac{\overline{R}^\alpha \pi e}{|E|} = 1 + \frac{1}{2|E|} - \frac{1}{2} \mathbb{E} [\mathcal{F}_G^\alpha (D^\alpha \pi e) | G],$$

from which we deduce that

$$\mathbb{E} [\mathcal{F}_G^\alpha (D^\alpha \pi e) | G] = 1. \tag{20}$$

The result now follows by inserting (20) in (19).

ii) Again, let \mathcal{E}' be an independent copy of \mathcal{E} and $\alpha, \beta \in \{+, -\}$. For $x, y \in \mathbb{R}$, we write $\widetilde{F}_G^\alpha(x) = F_{\widetilde{D^\alpha \pi_* \mathcal{E}}|G}(x)$ and similarly $\widetilde{F}_G^\beta(y) = F_{\widetilde{D^\beta \pi^* \mathcal{E}}|G}(y)$. Then we have,

$$\frac{R^\alpha \pi_* e}{|E|} = \frac{1}{|E|} \sum_{f \in E} I\{D^\alpha \pi_* f + U_f \geq D^\alpha \pi_* e + U_e\}$$

$$\begin{aligned}
&= \frac{1}{|E|} \sum_{f \in E} I \{ D^\alpha \pi_* f + U_f > D^\alpha \pi_* e + U_e \} + I \{ f = e \} \\
&= 1 - \mathbb{E} [I \{ D^\alpha \pi_* \mathcal{E}' + U_{\mathcal{E}'} \leq D^\alpha \pi_* e + U_e \} | G] + \frac{1}{|E|} \\
&= 1 - \tilde{F}_G^\alpha (D^\alpha \pi_* e + U_e) + \frac{1}{|E|}.
\end{aligned} \tag{21}$$

Using similar calculations we get

$$\frac{R^\beta \pi^* e}{|E|} = 1 - \tilde{F}_G^\beta (D^\beta \pi^* e + W_e) + \frac{1}{|E|}. \tag{22}$$

Now, using both (21) and (22), we obtain,

$$\begin{aligned}
\frac{1}{|E|} \sum_{e \in E} \frac{R^\alpha \pi_* e}{|E|} \frac{R^\beta \pi^* e}{|E|} &= 1 + \frac{2}{|E|} + \frac{1}{|E|^2} + \frac{1}{|E|} \sum_{e \in E} \tilde{F}_G^\alpha (D^\alpha \pi_* e + U_e) \tilde{F}_G^\beta (D^\beta \pi^* e + W_e) \\
&\quad - \left(1 + \frac{1}{|E|} \right) \frac{1}{|E|} \sum_{e \in E} \left(\tilde{F}_G^\alpha (D^\alpha \pi_* e + U_e) + \tilde{F}_G^\beta (D^\beta \pi^* e + W_e) \right) \\
&= 1 + \frac{2}{|E|} + \frac{1}{|E|^2} + \mathbb{E} \left[\tilde{F}_G^\alpha \left(\widetilde{D^\alpha \pi_* \mathcal{E}} \right) \tilde{F}_G^\beta \left(\widetilde{D^\beta \pi^* \mathcal{E}} \right) | G \right] \\
&\quad - \left(1 + \frac{1}{|E|} \right) \left(\mathbb{E} \left[\tilde{F}_G^\alpha \left(\widetilde{D^\alpha \pi_* \mathcal{E}} \right) | G \right] + \mathbb{E} \left[\tilde{F}_G^\beta \left(\widetilde{D^\beta \pi^* \mathcal{E}} \right) | G \right] \right) \\
&= \frac{1}{4} \mathbb{E} \left[\mathcal{F}_G^\alpha (D^\alpha \pi_* \mathcal{E}) \mathcal{F}_G^\beta (D^\beta \pi^* \mathcal{E}) | G \right] + \frac{1}{|E|} + \frac{1}{|E|^2}.
\end{aligned}$$

The last line follows by first using Propositions A.2 and A.4 to rewrite the conditional expectations and then applying (20). \square

4.3 Kendall's Tau

The definition for $\tau_\alpha^\beta(G)$ is, cf. [17],

$$\tau_\alpha^\beta(G) = \frac{2(\mathcal{N}_C(G) - \mathcal{N}_D(G))}{|E|(|E| - 1)},$$

where $\mathcal{N}_C(G)$ and $\mathcal{N}_D(G)$ denote the number of concordant and discordant pairs, respectively, among $(D^\alpha \pi_* e, D^\beta \pi^* e)_{e \in E}$. We recall that a pair $(D^\alpha \pi_* e, D^\beta \pi^* e)$ and $(D^\alpha \pi_* f, D^\beta \pi^* f)$, for $e, f \in E$ is called (discordant) concordant if

$$(D^\alpha \pi_* e - D^\alpha \pi_* f) (D^\beta \pi^* e - D^\beta \pi^* f) (< 0) > 0.$$

Therefore we have, for the concordant pairs,

$$\begin{aligned}
\frac{2}{|E|^2} \mathcal{N}_C(G) &= \frac{1}{|E|^2} \sum_{e, f \in E} I \{ D^\alpha \pi_*(f) < D^\alpha \pi_*(e), D^\beta \pi^*(f) < D^\beta \pi^*(e) \} \\
&\quad + \frac{1}{|E|^2} \sum_{e, f \in E} I \{ D^\alpha \pi_*(f) > D^\alpha \pi_*(e), D^\beta \pi^*(f) > D^\beta \pi^*(e) \} \\
&= \mathbb{E} \left[H_G^{\alpha, \beta} (D^\alpha \pi_* \mathcal{E} - 1, D^\beta \pi^* \mathcal{E} - 1) | G \right] \\
&\quad + 1 - \mathbb{E} [F_G^\alpha (D^\alpha \pi_* \mathcal{E}) | G] - \mathbb{E} [F_G^\beta (D^\beta \pi^* \mathcal{E}) | G] \\
&\quad + \mathbb{E} \left[H_G^{\alpha, \beta} (D^\alpha \pi_* \mathcal{E}, D^\beta \pi^* \mathcal{E}) | G \right].
\end{aligned}$$

In a similar fashion we get for the discordant pairs

$$\begin{aligned} \frac{2}{|E|^2} \mathcal{N}_D(G) &= \mathbb{E} [F_G^\alpha (D^\alpha \pi_* \mathcal{E} - 1) | G] + \mathbb{E} [F_G^\beta (D^\beta \pi^* \mathcal{E} - 1) | G] \\ &\quad - \mathbb{E} [H_G^{\alpha, \beta} (D^\alpha \pi_* \mathcal{E} - 1, D^\beta \pi^* \mathcal{E}) | G] - \mathbb{E} [H_G^{\alpha, \beta} (D^\alpha \pi_* \mathcal{E}, D^\beta \pi^* \mathcal{E} - 1) | G]. \end{aligned}$$

Combining the above with (20) we conclude that

$$\tau_\alpha^\beta(G) = \mathbb{E} [H_G^{\alpha, \beta} (D^\alpha \pi_* \mathcal{E}, D^\beta \pi^* \mathcal{E}) | G] - 1 + o_{\mathbb{P}}(|E|^{-1}). \quad (23)$$

4.4 Statistical consistency of rank correlations

We will now prove that the rank correlations defined in the previous two sections are, under natural regularity conditions on the degree sequences, consistent statistical estimators.

For a sequence $\{G_n\}_{n \in \mathbb{N}}$ of random graphs with $|V_n| = n$, it is common in the theory of random graphs to assume convergence of the empirical degree distributions, see for instance Condition 7.5 in [16], Condition 4.1 in [2]. Here, similarly to [7], we impose the following regularity condition on the degrees at the end points of edges.

Condition 4.2. *Given a sequence $\{G_n\}_{n \in \mathbb{N}}$ of random graphs with $|V_n| = n$ and $\alpha, \beta \in \{+, -\}$ there exist integer valued random variables \mathcal{D}^α and \mathcal{D}^β , not concentrated in a single point, such that*

$$(D_n^\alpha \pi_* \mathcal{E}_n, D_n^\beta \pi^* \mathcal{E}_n | G_n) \Rightarrow (\mathcal{D}^\alpha, \mathcal{D}^\beta) \quad \text{as } n \rightarrow \infty,$$

where \mathcal{E}_n is a uniformly sampled edge in G_n .

In the previous two sections it was shown that $\rho_\alpha^\beta(G)$, $\bar{\rho}_\alpha^\beta(G)$ and $\tau_\alpha^\beta(G)$ on a random graph G are related to, respectively,

$$\mathbb{E} [\mathcal{F}_G^\alpha (D^\alpha \pi_* \mathcal{E}) \mathcal{F}_G^\beta (D^\beta \pi^* \mathcal{E}) | G] \quad \text{and} \quad \mathbb{E} [H_G^{\alpha, \beta} (D^\alpha \pi_* \mathcal{E}, D^\beta \pi^* \mathcal{E}) | G].$$

Note that these are in fact empirical versions of the functions appearing in the definitions of Spearman's rho and Kendall's tau, cf. (3) and (5). The following result formalizes these observations and states that under Condition 4.2, $\rho_\alpha^\beta(G_n)$, $\bar{\rho}_\alpha^\beta(G_n)$ and $\tau_\alpha^\beta(G_n)$ are indeed consistent statistical estimators of correlation measures associated with Spearman's rho and Kendall's tau.

Theorem 4.3. *Let $\alpha, \beta \in \{+, -\}$ and $\{G_n\}_{n \in \mathbb{N}}$ be a sequence of graphs satisfying Condition 4.2 such that as $n \rightarrow \infty$, $|E_n| \xrightarrow{\mathbb{P}} \infty$. Then, as $n \rightarrow \infty$,*

- i) $\rho_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} \rho(\mathcal{D}^\alpha, \mathcal{D}^\beta)$,
- ii) $\bar{\rho}_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} \frac{\rho(\mathcal{D}^\alpha, \mathcal{D}^\beta)}{3\sqrt{S_{\mathcal{D}^\alpha}(\mathcal{D}^\alpha) S_{\mathcal{D}^\beta}(\mathcal{D}^\beta)}}$,

where $S_{\mathcal{D}^\alpha}(\mathcal{D}^\alpha) = \mathbb{E} [F_{\mathcal{D}^\alpha}(\mathcal{D}^\alpha) F_{\mathcal{D}^\alpha}(\mathcal{D}^\alpha - 1)]$, and

- iii) $\tau_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} \tau(\mathcal{D}^\alpha, \mathcal{D}^\beta)$.

Moreover, we have convergence of the first moments:

- iv) $\lim_{n \rightarrow \infty} \mathbb{E} [\rho_\alpha^\beta(G_n)] = \rho(\mathcal{D}^\alpha, \mathcal{D}^\beta)$,
- v) $\lim_{n \rightarrow \infty} \mathbb{E} [\bar{\rho}_\alpha^\beta(G_n)] = \frac{\rho(\mathcal{D}^\alpha, \mathcal{D}^\beta)}{3\sqrt{S_{\mathcal{D}^\alpha}(\mathcal{D}^\alpha) S_{\mathcal{D}^\beta}(\mathcal{D}^\beta)}} \quad \text{and}$

$$\text{vi}) \lim_{n \rightarrow \infty} \mathbb{E} [\tau_\alpha^\beta(G_n)] = \tau(\mathcal{D}^\alpha, \mathcal{D}^\beta).$$

Proof. i) By Proposition 4.1 we have that

$$\frac{12}{|E_n|} \sum_{e \in E_n} \frac{R_n^\alpha \pi_* e}{|E_n|} \frac{R_n^\beta \pi^* e}{|E_n|} = 3\mathbb{E} [\mathcal{F}_{G_n}^\alpha (D_n^\alpha \pi_* \mathcal{E}_n) \mathcal{F}_{G_n}^\beta (D_n^\beta \pi^* \mathcal{E}_n) \mid G_n] + o_{\mathbb{P}}(|E_n|^{-1}).$$

From this and the fact that $|E_n| \xrightarrow{\mathbb{P}} \infty$ it follows that,

$$\begin{aligned} \rho_\alpha^\beta(G_n) &= \frac{1}{1 - |E_n|^{-2}} \left(\frac{12}{|E_n|} \sum_{e \in E_n} \frac{R_n^\alpha \pi_* e}{|E_n|} \frac{R_n^\beta \pi^* e}{|E_n|} - 3 \frac{|E_n|(|E_n| + 1)^2}{|E_n|^3} \right) \\ &= 3\mathbb{E} [\mathcal{F}_{G_n}^\alpha (D_n^\alpha \pi_* \mathcal{E}_n) \mathcal{F}_{G_n}^\beta (D_n^\beta \pi^* \mathcal{E}_n) \mid G_n] - 3 + o_{\mathbb{P}}(|E_n|^{-1}) \\ &\xrightarrow{\mathbb{P}} \rho(\mathcal{D}^\alpha, \mathcal{D}^\beta) \quad \text{as } n \rightarrow \infty, \end{aligned}$$

where the last line follows from Theorem 3.1.

ii) From (18) it follows that,

$$\left(\frac{\bar{R}_n^\alpha \pi e}{|E_n|} \right)^2 = \left(1 + \frac{1}{2|E_n|} \right)^2 - \left(1 + \frac{1}{2|E_n|} \right) \mathcal{F}_{G_n}^\alpha (D^\alpha \pi e) + \frac{1}{4} \mathcal{F}_{G_n}^\alpha (D^\alpha \pi e)^2.$$

Therefore,

$$\begin{aligned} \frac{1}{|E_n|} \sum_{e \in E_n} \left(\frac{\bar{R}_n^\alpha \pi e}{|E_n|} \right)^2 &= \left(1 + \frac{1}{2|E_n|} \right)^2 + \frac{1}{4} \mathbb{E} [\mathcal{F}_{G_n}^\alpha (D^\alpha \pi \mathcal{E}_n)^2 \mid G_n] \\ &\quad - \left(1 + \frac{1}{2|E_n|} \right) \mathbb{E} [\mathcal{F}_{G_n}^\alpha (D^\alpha \pi \mathcal{E}_n) \mid G_n] \\ &= 1 + \frac{1}{4} \mathbb{E} [\mathcal{F}_{G_n}^\alpha (D^\alpha \pi \mathcal{E}_n)^2 \mid G_n] - \mathbb{E} [\mathcal{F}_{G_n}^\alpha (D_n^\alpha \pi \mathcal{E}_n) \mid G_n] + o_{\mathbb{P}}(|E_n|^{-1}) \\ &\xrightarrow{\mathbb{P}} 1 + \frac{1}{4} \mathbb{E} [\mathcal{F}_{\mathcal{D}^\alpha} (\mathcal{D}^\alpha)^2] - \mathbb{E} [\mathcal{F}_{\mathcal{D}^\alpha} (\mathcal{D}^\alpha)] \quad \text{as } n \rightarrow \infty \\ &= \frac{1}{4} + \frac{1}{4} \mathbb{E} [F_{\mathcal{D}^\alpha} (\mathcal{D}^\alpha) F_{\mathcal{D}^\alpha} (\mathcal{D}^\alpha - 1)], \end{aligned}$$

where we used Lemma A.1 for the last line. It follows that, as $n \rightarrow \infty$,

$$\frac{4}{|E_n|} \sum_{e \in E_n} \left(\frac{\bar{R}_n^\alpha \pi e}{|E_n|} \right)^2 - \frac{|E_n|(|E_n| + 1)^2}{|E_n|^3} \xrightarrow{\mathbb{P}} \mathbb{E} [F^\alpha (\mathcal{D}^\alpha) F^\alpha (\mathcal{D}^\alpha - 1)].$$

Since \mathcal{D}^α and \mathcal{D}^β are not concentrated in one point the above term is non-zero. Now, combining this with Proposition 4.1 i) and applying Theorem 3.1, we obtain

$$\bar{\rho}_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} \frac{\rho(\mathcal{D}^\alpha, \mathcal{D}^\beta)}{3\sqrt{S_{\mathcal{D}^\alpha}(\mathcal{D}^\alpha) S_{\mathcal{D}^\beta}(\mathcal{D}^\beta)}} \quad \text{as } n \rightarrow \infty.$$

iii) Combining (23) with Theorem 3.1 yields, as $n \rightarrow \infty$,

$$\tau_\alpha^\beta(G_n) = \mathbb{E} [\mathcal{H}_{G_n}^{\alpha, \beta} (D_n^\alpha \pi_* \mathcal{E}_n, D_n^\beta \pi^* \mathcal{E}_n) \mid G_n] - 1 + o_{\mathbb{P}}(|E_n|^{-1}) \xrightarrow{\mathbb{P}} \tau(\mathcal{D}^\alpha, \mathcal{D}^\beta).$$

Finally, iv), v), vi) now follow from, respectively, i), ii) and iii) since $\rho_\alpha^\beta(G_n)$, $\bar{\rho}_\alpha^\beta(G_n)$ and $\tau_\alpha^\beta(G_n)$ are bounded. \square

Comparing results i) and iv) to ii) and v), note that the way in which ties are resolved influences the measure estimated by Spearman's rho on random directed graphs. In particular, resolving ties uniformly at random yields the value corresponding to Spearman's rho for the two limiting integer valued random variables \mathcal{D}^α and \mathcal{D}^β as defined in [11], in the infinite size network limit.

5 Directed Configuration Model

In this section we will analyze degree-degree dependencies for the directed *Configuration Model* (CM), as described and analyzed in [2]. First, in Section 5.1, we analyze the model where in- and out-links are connected at random, which, in general, results in a multi-graph. Then we move on to two other models that produce simple graphs: the *Repeated* and *Erased* Configuration Model (RCM and ECM). By applying Theorem 4.3, in Sections 5.2 and 5.3, we will show that RCM and ECM can be used as null models for the rank correlations ρ , $\bar{\rho}$ and τ .

5.1 General model: multi-graphs

The directed Configuration Model in [2] starts with picking two target distributions F_- , F_+ for the in- and out-degrees, respectively, stochastically bounded from above by regularly varying distributions. We will adopt notations from [2] and let γ and ξ denote random variables with distributions F_- and F_+ , respectively. It is assumed that $\mathbb{E}[\gamma] = \mathbb{E}[\xi] < \infty$. The next step is generating a bi-degree sequence of inbound and outbound stubs. This is done by first taking two independent sequences of n independent copies of γ and ξ , which are then modified into a sequence of in- and outbound stubs

$$\widehat{\mathfrak{D}}(G) = \left(\widehat{D}^+(v), \widehat{D}^-(v) \right)_{v \in V},$$

using the algorithm in [2], Section 2.1. This algorithm ensures that the total number of in- and outbound stubs is the same, $|\widehat{E}| = \sum_{v \in V} \widehat{D}^\alpha(v)$, $\alpha \in \{+, -\}$. Using this bi-degree sequence, a graph is built by randomly pairing the stubs to form edges. We call a graph generated by this model a *Configuration Model graph*, or CM graph for short. We remark that a CM graph in general does not need to be simple.

Given a vertex set V , a bi-degree sequence $\widehat{\mathfrak{D}}(G)$ and $v \in V$, we denote by v_i^+ , v_j^- for $1 \leq i \leq \widehat{D}^+(v)$ and $1 \leq j \leq \widehat{D}^-(v)$, respectively, the outbound and inbound stubs of v . For $v, w \in V$, we denote by $\{v_i^+ \rightarrow w_j^-\}$ the event that the outbound stub v_i^+ is connected to the inbound stub w_j^- and by $\{v_i^+ \rightarrow w\}$ the event that v_i^+ is connected to an inbound stub of w . By definition of CM, it follows that $\mathbb{P}(v_i^+ \rightarrow w_j^- | \widehat{\mathfrak{D}}(G)) = 1/|\widehat{E}|$ and hence $\mathbb{P}(v_i^+ \rightarrow w | \widehat{\mathfrak{D}}(G)) = \widehat{D}^-(w)/|\widehat{E}|$. Furthermore we observe that $|\widehat{E}_n(e)| = \sum_{i=1}^{\widehat{D}_n^+ \pi_* e} I\{(\pi_* e)_i^+ \rightarrow \pi^* e\}$. Given a random graph G , we denote

$$I_e^{\alpha, \beta}(k, l) = I\{D^\alpha \pi_* e = k\} I\{D^\beta \pi^* e = l\},$$

where $\alpha, \beta \in \{+, -\}$, $k, l \in \mathbb{N}$ and $e \in V^2$.

For proper reference we summarize some results from Proposition 2.5, in [2], which we will use in the remainder of this paper.

Proposition 5.1 ([2], Proposition 2.5). *Let $\widehat{\mathfrak{D}}(G_n)$ be the bi-degree sequence on n vertices, as generated in Section 2.1 of [2], and $k, l \in \mathbb{N}$. Then, as $n \rightarrow \infty$,*

$$\begin{aligned} \frac{1}{n} \sum_{v \in V_n} I\{\widehat{D}_n^+ v = k\} I\{\widehat{D}_n^- v = l\} &\xrightarrow{\mathbb{P}} \mathbb{P}(\xi = k) \mathbb{P}(\gamma = l), \\ \frac{1}{n} \sum_{v \in V_n} \widehat{D}_n^+ v &\xrightarrow{\mathbb{P}} \mathbb{E}[\xi] \quad \text{and} \quad \frac{1}{n} \sum_{v \in V_n} \widehat{D}_n^- v \xrightarrow{\mathbb{P}} \mathbb{E}[\gamma]. \end{aligned}$$

Given a random graph $G = (V, E)$, we will use $\mathfrak{D}(G)$ as a short hand notation for its degree sequence $(D^-(v), D^+(v))_{v \in V}$. We emphasize that for a graph generated using an initial bi-degree sequence, the eventual degree sequence $\mathfrak{D}(G)$ can be different from $\widehat{\mathfrak{D}}(G)$. This, for example, is true for the ECM, Section 5.3, where, after the random pairing of the stubs, self-loops are removed and multiple edges are merged.

In order to apply Theorem 4.3 to a sequence of (multi-)graphs $\{G_n\}_{n \in \mathbb{N}}$ generated by CM, we need to prove that

$$(D_n^\alpha \pi_* \mathcal{E}_n, D_n^\beta \pi^* \mathcal{E}_n | G_n) \Rightarrow (\mathcal{D}^\alpha, \mathcal{D}^\beta),$$

for some integer valued random variables \mathcal{D}^α and \mathcal{D}^β . For this, it suffices to show that, as $n \rightarrow \infty$,

$$H_{G_n}^{\alpha, \beta}(k, l) \xrightarrow{\mathbb{P}} H_{\mathcal{D}^\alpha, \mathcal{D}^\beta}(k, l),$$

for all $k, l \in \mathbb{N}$. We will prove this by showing that

$$\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \xrightarrow{\mathbb{P}} \mathbb{P}(\mathcal{D}^\alpha = k, \mathcal{D}^\beta = l),$$

as $n \rightarrow \infty$, using a second moment argument as follows. Given a sequence $\{G_n\}_{n \in \mathbb{N}}$ of graphs, $\alpha, \beta \in \{+, -\}$ and $k, l \in \mathbb{N}$, we will show that the empirical joint probability $\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \right]$ converges to $\mathbb{P}(\mathcal{D}^\alpha = k, \mathcal{D}^\beta = l)$. Then we will prove that the variance of $\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right]$ converges to zero.

We start with expressing the first and second moment of $\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right]$, for CM graphs, conditioned on the bi-degree sequence $\widehat{\mathfrak{D}}(G_n)$ in terms of the degrees. We observe that, for $\alpha, \beta \in \{+, -\}$, $e \in V_n^2$ and $k, l \in \mathbb{N}$, the events $\{D_n^\alpha \pi_* e = k\}$ and $\{D_n^\beta \pi^* e = l\}$ are completely defined by $\widehat{\mathfrak{D}}(G_n)$, hence so is $I_e^{\alpha, \beta}(k, l)$. We remark that, since CM leaves the number of inbound and outbound stubs intact, we have $\mathfrak{D}(G_n) = \widehat{\mathfrak{D}}(G_n)$. However, in this section we will keep using hats, e.g. \widehat{D}_n instead of D_n , to emphasize that G_n can be a multi-graph.

Lemma 5.2. *Let $\{G_n\}_{n \in \mathbb{N}}$ be a sequence of CM graphs with $|V_n| = n$ and $\alpha, \beta \in \{+, -\}$. Then, for each $k, l \in \mathbb{N}$,*

$$\begin{aligned} \text{i) } & \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right] = \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{\widehat{D}_n^+ \pi_* e \widehat{D}_n^- \pi^* e}{|\widehat{E}_n|^2} \quad \text{and} \\ \text{ii) } & \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right]^2 \middle| \widehat{\mathfrak{D}}(G_n) \right] = \left(\sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{\widehat{D}_n^+ \pi_* e \widehat{D}_n^- \pi^* e}{|\widehat{E}_n|^2} \right)^2 + o_{\mathbb{P}}(1). \end{aligned}$$

Proof.

$$\begin{aligned} \text{i) } & \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right] = \mathbb{E} \left[\sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{|\widehat{E}_n(e)|}{|\widehat{E}_n|} \middle| \widehat{\mathfrak{D}}(G_n) \right] \tag{24} \\ & = \frac{1}{|\widehat{E}_n|} \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \mathbb{E} \left[|\widehat{E}_n(e)| \middle| \widehat{\mathfrak{D}}(G_n) \right] \\ & = \frac{1}{|\widehat{E}_n|} \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \mathbb{E} \left[\sum_{i=1}^{\widehat{D}_n^+ \pi_* e} I \left\{ (\pi_* e)_i^+ \rightarrow \pi^* e \right\} \middle| \widehat{\mathfrak{D}}(G_n) \right] \\ & = \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{\left(\widehat{D}_n^+ \pi_* e \right) \left(\widehat{D}_n^- \pi^* e \right)}{|\widehat{E}_n|^2}. \end{aligned}$$

ii) Following similar calculations as above we get,

$$\begin{aligned} & \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right]^2 \middle| \widehat{\mathfrak{D}}(G_n) \right] \\ & = \mathbb{E} \left[\sum_{e, f \in V_n^2} I_e^{\alpha, \beta}(k, l) I_f^{\alpha, \beta}(k, l) \frac{|\widehat{E}_n(e)| |\widehat{E}_n(f)|}{|\widehat{E}_n|^2} \middle| \widehat{\mathfrak{D}}(G_n) \right] \tag{25} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{|\widehat{E}_n|^2} \sum_{e,f \in V_n^2} \left(I_e^{\alpha,\beta}(k,l) I_f^{\alpha,\beta}(k,l) \right. \\
&\quad \left. \sum_{i=1}^{\widehat{D}_n^+ \pi_* e} \sum_{s=1}^{\widehat{D}_n^+ \pi_* f} \mathbb{E} \left[I \{(\pi_* e)_i^+ \rightarrow \pi^* e\} I \{(\pi_* f)_s^+ \rightarrow \pi^* f\} \mid \widehat{\mathfrak{D}}(G_n) \right] \right). \quad (26)
\end{aligned}$$

We will, for $e, f \in V_n^2$, analyze

$$\frac{1}{|\widehat{E}_n|^2} \sum_{i=1}^{\widehat{D}_n^+ \pi_* e} \sum_{s=1}^{\widehat{D}_n^+ \pi_* f} \mathbb{E} \left[I \{(\pi_* e)_i^+ \rightarrow \pi^* e\} I \{(\pi_* f)_s^+ \rightarrow \pi^* f\} \mid \widehat{\mathfrak{D}}(G_n) \right] \quad (27)$$

for all different cases, $e = f$, $e \cap f = \emptyset$, $e_* = f_*$ and $e^* = f^*$. First, suppose that $e = f$. Then (27) equals

$$\frac{1}{|\widehat{E}_n|^2} \sum_{i,s=1}^{\widehat{D}_n^+ \pi_* e} \sum_{j,t=1}^{\widehat{D}_n^- \pi^* e} \frac{I \{i = s\} I \{j = t\}}{|\widehat{E}_n|} + \frac{I \{i \neq s\} I \{j \neq t\}}{|\widehat{E}_n|(|\widehat{E}_n| - 1)}.$$

Writing out the sums and using that $e = f$ we obtain,

$$(27) = \frac{\widehat{D}_n^+ \pi_* e \widehat{D}_n^- \pi^* e \widehat{D}_n^+ \pi_* f \widehat{D}_n^- \pi^* f}{|\widehat{E}_n|^3 (|\widehat{E}_n| - 1)} \quad (28)$$

$$+ \frac{(\widehat{D}_n^+ \pi_* e) (\widehat{D}_n^- \pi^* e)}{|\widehat{E}_n|^3} + \frac{(\widehat{D}_n^+ \pi_* e) (\widehat{D}_n^- \pi^* e)}{|\widehat{E}_n|^3 (|\widehat{E}_n| - 1)} \quad (29)$$

$$- \frac{(\widehat{D}_n^- \pi^* e)^2 (\widehat{D}_n^+ \pi_* e)}{|\widehat{E}_n|^3 (|\widehat{E}_n| - 1)} - \frac{(\widehat{D}_n^+ \pi_* e)^2 (\widehat{D}_n^- \pi^* e)}{|\widehat{E}_n|^3 (|\widehat{E}_n| - 1)} \quad (30)$$

Since for all $k \geq 0$ and $\kappa \in \{+,-\}$ it holds that

$$\frac{1}{|\widehat{E}_n|^{k+1}} \sum_{v \in V_n} (\widehat{D}_n^\kappa v)^k \leq \frac{1}{|\widehat{E}_n|^{k+1}} \left(\sum_{v \in V_n} \widehat{D}_n^\kappa v \right)^k = \frac{1}{|\widehat{E}_n|},$$

we deduce that the terms in (29) and (30) contribute as $o_{\mathbb{P}}(1)$ in (26), from which the result for $e = f$ follows. The calculations for the other three cases for $e, f \in V_n^2$ are similar and are hence omitted. \square

As a direct consequence we have the following

Proposition 5.3. *Let $\{G_n\}_{n \in \mathbb{N}}$ be a sequence of CM graphs with $|V_n| = n$ and $\alpha, \beta \in \{+,-\}$. Then, for each $k, l \in \mathbb{N}$, as $n \rightarrow \infty$,*

$$\left| \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha,\beta}(k,l) \mid G_n \right]^2 \mid \widehat{\mathfrak{D}}(G_n) \right] - \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha,\beta}(k,l) \mid G_n \right] \mid \widehat{\mathfrak{D}}(G_n) \right]^2 \right| \xrightarrow{\mathbb{P}} 0.$$

Now, using the convergence results from [2], summarized in Proposition 5.1, we are able to determine the limiting random variables \mathcal{D}^α and \mathcal{D}^β .

Proposition 5.4. *Let $\{G_n\}_{n \in \mathbb{N}}$ be a sequence of CM graphs with $|V_n| = n$ and $\alpha, \beta \in \{+,-\}$. Then there exist integer valued random variables \mathcal{D}^α and \mathcal{D}^β such that for each $k, l \in \mathbb{N}$, as $n \rightarrow \infty$,*

$$\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}}^{\alpha,\beta}(k,l) \mid G_n \right] \mid \widehat{\mathfrak{D}}(G_n) \right] \xrightarrow{\mathbb{P}} \mathbb{P}(\mathcal{D}^\alpha = k) \mathbb{P}(\mathcal{D}^\beta = l).$$

Proof. First let $(\alpha, \beta) = (+, -)$. Then it follows from Lemma 5.2 i) that

$$\begin{aligned}
\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right] \mid \widehat{\mathcal{D}}(G_n) \right] &= \sum_{v, w \in V_n} I \left\{ \widehat{D}_n^+ v = k \right\} I \left\{ \widehat{D}_n^- w = l \right\} \frac{\widehat{D}_n^+ v \widehat{D}_n^- w}{|\widehat{E}_n|^2} \\
&= \left(\sum_{v \in V_n} I \left\{ \widehat{D}_n^+ v = k \right\} \frac{\widehat{D}_n^+ v}{|\widehat{E}_n|} \right) \left(\sum_{w \in V_n} I \left\{ \widehat{D}_n^- w = l \right\} \frac{\widehat{D}_n^- w}{|\widehat{E}_n|} \right) \\
&= \left(k \sum_{v \in V_n} \frac{I \left\{ \widehat{D}_n^+ v = k \right\}}{|\widehat{E}_n|} \right) \left(l \sum_{w \in V_n} \frac{I \left\{ \widehat{D}_n^- w = l \right\}}{|\widehat{E}_n|} \right) \\
&\xrightarrow{\mathbb{P}} \frac{k \mathbb{P}(\xi = k)}{\mathbb{E}[\xi]} \frac{l \mathbb{P}(\gamma = l)}{\mathbb{E}[\gamma]} \quad \text{as } n \rightarrow \infty,
\end{aligned}$$

where the convergence in the last line is by Proposition 5.1. The other three cases are slightly more involved. Consider, for example, $(\alpha, \beta) = (-, +)$. Then we have,

$$\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right] \mid \widehat{\mathcal{D}}(G_n) \right] = \sum_{v \in V_n} I \left\{ \widehat{D}_n^- v = k \right\} \frac{\widehat{D}_n^+ v}{|\widehat{E}_n|} \sum_{w \in V_n} I \left\{ \widehat{D}_n^+ w = l \right\} \frac{\widehat{D}_n^- w}{|\widehat{E}_n|} \quad (31)$$

We will first analyze the last summation.

$$\begin{aligned}
\frac{1}{|\widehat{E}_n|} \sum_{w \in V_n} \widehat{D}_n^-(w) I \left\{ \widehat{D}_n^+ w = l \right\} &= \frac{1}{|\widehat{E}_n|} \sum_{i \in \mathbb{N}} i \sum_{w \in V_n} I \left\{ \widehat{D}_n^- w = i \right\} I \left\{ \widehat{D}_n^+ w = l \right\} \\
&\xrightarrow{\mathbb{P}} \frac{\mathbb{P}(\xi = l)}{\mathbb{E}[\xi]} \sum_{i \in \mathbb{N}} i \mathbb{P}(\gamma = i) \quad \text{as } n \rightarrow \infty \\
&= \frac{\mathbb{P}(\xi = l) \mathbb{E}[\gamma]}{\mathbb{E}[\xi]} = \mathbb{P}(\xi = l), \quad (32)
\end{aligned}$$

where we again used Proposition 5.1 and $\mathbb{E}[\gamma] = \mathbb{E}[\xi]$. In a similar way we obtain that, as $n \rightarrow \infty$,

$$\frac{1}{|\widehat{E}_n|} \sum_{v \in V_n} \widehat{D}_n^+(v) I \left\{ \widehat{D}_n^-(v) = k \right\} \xrightarrow{\mathbb{P}} \mathbb{P}(\gamma = k). \quad (33)$$

Applying (32) and (33) to (31) we get

$$\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{-, +}(k, l) \mid G_n \right] \mid \widehat{\mathcal{D}}(G_n) \right] \xrightarrow{\mathbb{P}} \mathbb{P}(\gamma = k) \mathbb{P}(\xi = l).$$

For the other two cases we obtain, as $n \rightarrow \infty$,

$$\begin{aligned}
\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{+, +}(k, l) \mid G_n \right] \mid \widehat{\mathcal{D}}(G_n) \right] &\xrightarrow{\mathbb{P}} \frac{k \mathbb{P}(\xi = k) \mathbb{P}(\xi = l)}{\mathbb{E}[\xi]} \\
\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{-, -}(k, l) \mid G_n \right] \mid \widehat{\mathcal{D}}(G_n) \right] &\xrightarrow{\mathbb{P}} \frac{l \mathbb{P}(\gamma = k) \mathbb{P}(\gamma = l)}{\mathbb{E}[\gamma]}
\end{aligned}$$

The results now holds if we define \mathcal{D}^α and \mathcal{D}^β by their probabilities summarized in Table 1. \square

We end this section with a convergence result for first and second moment of $\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right]$.

Proposition 5.5. *Let $\{G_n\}_{n \in \mathbb{N}}$ be a sequence of CM graphs with $|V_n| = n$ and $\alpha, \beta \in \{+, -\}$. Then, for each $k, l \in \mathbb{N}$,*

$$\text{i) } \lim_{n \rightarrow \infty} \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right] \right] = \mathbb{P}(\mathcal{D}^\alpha = k) \mathbb{P}(\mathcal{D}^\beta = l),$$

α	β	$\mathbb{P}(\mathcal{D}^\alpha = k)$	$\mathbb{P}(\mathcal{D}^\beta = l)$
+	-	$k\mathbb{P}(\xi = k)/\mathbb{E}[\xi]$	$l\mathbb{P}(\gamma = l)/\mathbb{E}[\gamma]$
-	+	$\mathbb{P}(\gamma = k)$	$\mathbb{P}(\xi = l)$
+	+	$k\mathbb{P}(\xi = k)/\mathbb{E}[\xi]$	$\mathbb{P}(\xi = l)$
-	-	$\mathbb{P}(\gamma = k)$	$l\mathbb{P}(\gamma = l)/\mathbb{E}[\gamma]$

Table 1: Distributions of \mathcal{D}^α and \mathcal{D}^β for $\alpha, \beta \in \{+, -\}$.

$$\text{ii) } \lim_{n \rightarrow \infty} \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right]^2 \right] = \mathbb{P}(\mathcal{D}^\alpha = k)^2 \mathbb{P}(\mathcal{D}^\beta = l)^2,$$

and hence, as $n \rightarrow \infty$, $\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right] \xrightarrow{\mathbb{P}} \mathbb{P}(\mathcal{D}^\alpha = k) \mathbb{P}(\mathcal{D}^\beta = l)$.

Proof.

i) Let $k, l \in \mathbb{N}$, then, since

$$\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}}^{\alpha, \beta}(k, l) \mid G_n \right] \mid \widehat{\mathfrak{D}}(G_n) \right] \leq 1, \quad (34)$$

it follows, using Proposition 5.4 and dominated convergence, that for each pair $\alpha, \beta \in \{+, -\}$, we have

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right] \right] = \mathbb{P}(\mathcal{D}^\alpha = k) \mathbb{P}(\mathcal{D}^\beta = l),$$

where $\mathcal{D}^\alpha, \mathcal{D}^\beta$ have distributions defined in Table 1.

ii) For the second moment we get, using conditioning on $\widehat{\mathfrak{D}}(G_n)$,

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right]^2 \right] &= \lim_{n \rightarrow \infty} \mathbb{E} \left[\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right]^2 \mid \widehat{\mathfrak{D}}(G_n) \right] \right] \\ &= \lim_{n \rightarrow \infty} \mathbb{E} \left[\left(\sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{\widehat{D}_n^+ \pi^* e \widehat{D}_n^- \pi^* e}{|\widehat{E}_n|^2} \right)^2 + o_{\mathbb{P}}(1) \right] \quad (35) \end{aligned}$$

$$= \lim_{n \rightarrow \infty} \mathbb{E} \left[\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}}^{\alpha, \beta}(k, l) \mid G_n \right] \mid \widehat{\mathfrak{D}}(G_n) \right]^2 + o_{\mathbb{P}}(1) \right] \quad (36)$$

$$= (\mathbb{P}(\mathcal{D}^\alpha = k) \mathbb{P}(\mathcal{D}^\beta = l))^2. \quad (37)$$

Here (35) follows from Lemma 5.2 ii), (36) is by Lemma 5.2 i), and (37) is due to Proposition 5.4, continuous mapping theorem, (34) and the fact that the $o_{\mathbb{P}}(1)$ terms are uniformly bounded, see proof Lemma 5.2. The distributions of $\mathcal{D}^\alpha, \mathcal{D}^\beta$ are again given in Table 1.

The last result now follows by a second moment argument. \square

5.2 Repeated Configuration Model

Described in Section 4.1 of [2], RCM connects inbound and outbound stubs uniformly at random and then the resulting graph is checked to be simple. If not, one repeats the connection step until the resulting graph is simple. If the distributions F_- and F_+ have finite variances, then the probability of the graph being simple converges to a non-zero number, see [2], Theorem 4.3. Therefore, throughout this section, we will assume that $\mathbb{E}[\gamma^2], \mathbb{E}[\xi^2] < \infty$.

Let $\{G_n\}_{n \in \mathbb{N}}$ be again a sequence of CM graphs, and let S_n denote the event that G_n is simple. We will prove, in Theorem 5.7 below, that for a sequence of RCM graphs of growing size, our three rank correlation measures converge to zero, by showing that for all $\alpha, \beta \in \{+, -\}$ and $k, l \in \mathbb{N}$,

$$\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n, S_n \right] \xrightarrow{\mathbb{P}} \mathbb{P}(\mathcal{D}^\alpha = k) \mathbb{P}(\mathcal{D}^\beta = l),$$

as $n \rightarrow \infty$, where \mathcal{D}^α and \mathcal{D}^β are random variables whose distributions are defined in Table 1.

First we show that, asymptotically, conditioning on the graph being simple does not effect the conditional expectation $\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right]$.

Lemma 5.6. *Let $\{G_n\}_{n \in \mathbb{N}}$ be a sequence of CM graphs with $|V_n| = n$ and $\alpha, \beta \in \{+, -\}$ and denote by S_n the event that G_n is simple. Then, for each $k, l \in \mathbb{N}$, as $n \rightarrow \infty$,*

$$\left| \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n, S_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right] - \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right] \right| \xrightarrow{\mathbb{P}} 0.$$

Proof. First, we write

$$\begin{aligned} & \left| \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n, S_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right] - \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right] \right| \\ &= \left| \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \left(\frac{\mathbb{P}(S_n)}{\mathbb{P}(S_n)} - 1 \right) \middle| \widehat{\mathfrak{D}}(G_n) \right] \right|. \end{aligned} \quad (38)$$

Next, denote by

$$\text{Var} \left(\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right) \quad \text{and} \quad \text{Var} \left(I \{S_n\} \middle| \widehat{\mathfrak{D}}(G_n) \right)$$

the variance of, respectively $\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right]$ and $I \{S_n\}$, conditioned on $\widehat{\mathfrak{D}}(G_n)$. Then, by adding and subtracting in (38) the product of the conditional expectations

$$\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right] \left(\frac{\mathbb{P}(S_n \mid \widehat{\mathfrak{D}}(G_n))}{\mathbb{P}(S_n)} - 1 \right),$$

we get

$$\begin{aligned} (38) &\leq \frac{1}{\mathbb{P}(S_n)} \sqrt{\text{Var} \left(I \{S_n\} \middle| \widehat{\mathfrak{D}}(G_n) \right)} \sqrt{\text{Var} \left(\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right)} \\ &\quad + \left| \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right] \left(\frac{\mathbb{P}(S_n \mid \widehat{\mathfrak{D}}(G_n))}{\mathbb{P}(S_n)} - 1 \right) \right| \\ &\leq \frac{1}{\mathbb{P}(S_n)} \sqrt{\text{Var} \left(\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right)} + \left| \frac{\mathbb{P}(S_n \mid \widehat{\mathfrak{D}}(G_n))}{\mathbb{P}(S_n)} - 1 \right|. \end{aligned} \quad (39)$$

Following the argument in the first part of the proof of Proposition 4.4 from [2] we conclude that, $\mathbb{P}(\mathcal{S}_n \mid \widehat{\mathfrak{D}}(G_n))$ and $\mathbb{P}(\mathcal{S}_n)$ converge to the same positive limit, hence the latter expression in (39) is $o_{\mathbb{P}}(1)$. The result now follows, since by Proposition 5.3

$$\text{Var} \left(\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n) \right) = o_{\mathbb{P}}(1).$$

□

In the next theorem we show that the conditions of Theorem 4.3 hold for a sequence of RCM graphs, and thus obtain the desired convergence of the three rank correlations, using a second moment argument.

Theorem 5.7. *Let $\{G_n\}_{n \in \mathbb{N}}$ be a sequence of RCM graphs with $|V_n| = n$ and $\alpha, \beta \in \{+, -\}$. Then, as $n \rightarrow \infty$,*

$$\rho_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} 0, \quad \bar{\rho}_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} 0 \quad \text{and} \quad \tau_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} 0.$$

Proof. Instead of conditioning on RCM graphs we condition on CM graphs G_n and the event that it is simple, S_n . Let $k, l \in \mathbb{N}$ and let $\mathcal{D}^\alpha, \mathcal{D}^\beta$ have distributions defined in Table 1. Then, for each pair $\alpha, \beta \in \{+, -\}$, we have

$$\begin{aligned} & \left| \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n, S_n \right] \mid \widehat{\mathfrak{D}}(G_n) \right] - \mathbb{P}(\mathcal{D}^\alpha = k) \mathbb{P}(\mathcal{D}^\beta = l) \right| \\ & \leq \left| \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n, S_n \right] \mid \widehat{\mathfrak{D}}(G_n) \right] - \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right] \mid \widehat{\mathfrak{D}}(G_n) \right] \right| \\ & \quad + \left| \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right] \mid \widehat{\mathfrak{D}}(G_n) \right] - \mathbb{P}(\mathcal{D}^\alpha = k) \mathbb{P}(\mathcal{D}^\beta = l) \right|. \end{aligned}$$

Hence by Lemma 5.6 and Proposition 5.4 it follows that, as $n \rightarrow \infty$,

$$\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n, S_n \right] \mid \widehat{\mathfrak{D}}(G_n) \right] \xrightarrow{\mathbb{P}} \mathbb{P}(\mathcal{D}^\alpha = k) \mathbb{P}(\mathcal{D}^\beta = l).$$

Since $\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n, S_n \right] \mid \widehat{\mathfrak{D}}(G_n) \right] \leq 1$, dominated convergence and the above imply that

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n, S_n \right] \right] = \mathbb{P}(\mathcal{D}^\alpha = k) \mathbb{P}(\mathcal{D}^\beta = l). \quad (40)$$

For the second moment we have

$$\begin{aligned} & \left| \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n, S_n \right]^2 \mid \widehat{\mathfrak{D}}(G_n) \right] - \mathbb{P}(\mathcal{D}^\alpha = k)^2 \mathbb{P}(\mathcal{D}^\beta = l)^2 \right| \\ & \leq \left| \mathbb{E} \left[\left(\left(\frac{I\{S_n\}}{\mathbb{P}(S_n)} \right)^2 - 1 \right) \mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right]^2 \mid \widehat{\mathfrak{D}}(G_n) \right] \right| \end{aligned} \quad (41)$$

$$+ \left| \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right]^2 \mid \widehat{\mathfrak{D}}(G_n) \right] - \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right] \mid \widehat{\mathfrak{D}}(G_n) \right]^2 \right| \quad (42)$$

$$+ \left| \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right] \mid \widehat{\mathfrak{D}}(G_n) \right]^2 - \mathbb{P}(\mathcal{D}^\alpha = k)^2 \mathbb{P}(\mathcal{D}^\beta = l)^2 \right| \quad (43)$$

From Proposition 5.3 it follows that (42) converges to zero, while this holds for (43) because of Proposition 5.4 and the continuous mapping theorem. Finally, since

$$\left(\left(\frac{I\{S_n\}}{\mathbb{P}(S_n)} \right)^2 - 1 \right) \leq \left(\frac{I\{S_n\}}{\mathbb{P}(S_n)} - 1 \right) \left(1 + \mathbb{P}(S_n)^{-1} \right) \quad \text{and} \quad \mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right] \leq 1,$$

it follows that

$$(41) \leq \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right] \left(\frac{I\{S_n\}}{\mathbb{P}(S_n)} - 1 \right) \mid \widehat{\mathfrak{D}}(G_n) \right] \left(1 + \mathbb{P}(S_n)^{-1} \right) \xrightarrow{\mathbb{P}} 0 \quad \text{as } n \rightarrow \infty,$$

by (38), Lemma 5.6 and Proposition 4.4 from [2]. Therefore, using (34) and dominated convergence, we get

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n, S_n \right]^2 \right] = \mathbb{P}(\mathcal{D}^\alpha = k)^2 \mathbb{P}(\mathcal{D}^\beta = l)^2. \quad (44)$$

Combining (40) and (44), a second moment argument now yields that,

$$\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n, S_n \right] \xrightarrow{\mathbb{P}} \mathbb{P}(\mathcal{D}^\alpha = k) \mathbb{P}(\mathcal{D}^\beta = l) \quad \text{as } n \rightarrow \infty.$$

The result now follows from Theorem 4.3 by observing that the random variables \mathcal{D}^α and \mathcal{D}^β are independent and not concentrated in a single point. The latter is needed so that in case of average ranking we have $S_{\mathcal{D}^\alpha}(\mathcal{D}^\alpha) \neq 0$, see Theorem 4.3. \square

5.3 Erased Configuration Model

When the variances of the degree distributions are infinite, the probability of getting a simple graph using RCM converges to zero as the graph size increases. To remedy this we use ECM, described in Section 4.2 of [2]. In ECM stubs are connected at random, and then self-loops are removed and multiple edges are merged. We emphasize that for this model the actual degree sequence $\mathfrak{D}(G)$ may differ from the bi-degree sequence, $\widehat{\mathfrak{D}}(G)$, used to do the pairing.

We will often use results from Proposition 4.5 of [2], which we state below for reference.

Proposition 5.8 ([2], Proposition 4.5). *Let $G_n = (V_n, E_n)$ be a sequence of ECM graphs with $|V_n| = n$ and $k, l \in \mathbb{N}$. Then, as $n \rightarrow \infty$,*

$$\frac{1}{n} \sum_{v \in V_n} I\{D^+v = k\} \xrightarrow{\mathbb{P}} \mathbb{P}(\xi = k) \quad \text{and} \quad \frac{1}{n} \sum_{v \in V_n} I\{D^-v = l\} \xrightarrow{\mathbb{P}} \mathbb{P}(\gamma = l).$$

We will follow the same second moment argument approach as in the previous section to prove that all three rank correlations, ρ , $\bar{\rho}$ and τ converge to zero in ECM. First we will establish a convergence result for the total number of erased in- and outbound stubs.

For $v, w \in V$ and $\alpha \in \{+, -\}$, we denote by $E^{c, \alpha}(v)$ and $E^c(v, w)$, respectively, the set of erased α -stubs from v and erased edges between v and w . For $e \in V^2$, we write $E^c(e) = E^c(\pi_* e, \pi^* e)$.

Lemma 5.9. *Let $\{G_n\}_{n \in \mathbb{N}}$ be a sequence of ECM graphs with $|V_n| = n$ and $\alpha \in \{+, -\}$. Then*

$$\frac{1}{n} \sum_{v \in V_n} |E_n^{c, \alpha}(v)| \xrightarrow{\mathbb{P}} 0 \quad \text{as } n \rightarrow \infty.$$

Proof. Let $N \in \mathbb{N}$ and fix a $v \in V_N$, then for all $n \geq N$, $|E_n^{c, \alpha}(v)| \leq \gamma_n + 1$ where all γ_n are i.i.d. copies of γ . Since by Lemma 5.2 from [2] we have $E_n^{c, \alpha}(v) \rightarrow 0$ almost surely and furthermore $\mathbb{E}[\gamma] < \infty$, dominated convergence implies that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{v \in V_n} \mathbb{E}[|E_n^{c, \alpha}(v)|] = 0.$$

Applying the Markov inequality then yields, for arbitrary $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{1}{n} \sum_{v \in V_n} |E_n^{c, \alpha}(v)| \geq \varepsilon \right) \leq \lim_{n \rightarrow \infty} \frac{\sum_{v \in V_n} \mathbb{E}[|E_n^{c, \alpha}(v)|]}{n\varepsilon} = 0.$$

□

Since

$$|E| = |\widehat{E}| - \sum_{v \in V} |E^{c, \alpha}(v)| \quad \text{for } \alpha \in \{+, -\},$$

the above lemma combined with Proposition 5.1 implies that

$$\frac{|E_n|}{n} \xrightarrow{\mathbb{P}} \mathbb{E}[\gamma] \quad \text{as } n \rightarrow \infty. \quad (45)$$

We proceed with the next lemma, which is an adjustment of Lemma 5.2, where we now condition on both the bi-degree sequence of stubs as well as the eventual degree sequence. We remark that $I_e^{\alpha, \beta}(k, l)$ is completely determined by the latter while $\sum_{e \in V^2} |E^c(e)|$ is completely determined by the combination of the two sequences. Recall that for $e \in V^2$, $|\widehat{E}(e)|$ denotes the number of edges $f \in E$ with $f = e$ before removal of self-loops and merging multiple edges and observe that $|E(e)| = |\widehat{E}(e)| - |E^c(e)|$.

Lemma 5.10. *Let $\{G_n\}_{n \in \mathbb{N}}$ be a sequence of ECM graphs with $|V_n| = n$. Then, for each $k, l \in \mathbb{N}$ and $\alpha, \beta \in \{+, -\}$,*

- i) $\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] = \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{D_n^+ \pi_* e D_n^- \pi^* e}{|E_n|^2} + o_{\mathbb{P}}(1),$
- ii) $\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right]^2 \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] = \left(\sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{D_n^+ \pi_* e D_n^- \pi^* e}{|E_n|^2} \right)^2 + o_{\mathbb{P}}(1).$

To obtain this result we need the following Lemma.

Lemma 5.11. *Let $\{G_n\}_{n \in \mathbb{N}}$ be a sequence of ECM graphs with $|V_n| = n$. Then, for each $k, l \in \mathbb{N}$ and $\alpha, \beta \in \{+, -\}$,*

$$\sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{\widehat{D}_n^+ \pi_* e \widehat{D}_n^- \pi^* e}{|\widehat{E}_n|^2} = \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{D_n^+ \pi_* e D_n^- \pi^* e}{|E_n|^2} + o_{\mathbb{P}}(1).$$

Proof. Since $\widehat{D}_n^\alpha \pi e = D_n^\alpha \pi e + |E_n^{c, \alpha}(\pi e)|$, we have

$$\begin{aligned} \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{\widehat{D}_n^+ \pi_* e \widehat{D}_n^- \pi^* e}{|\widehat{E}_n|^2} &= \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{D_n^+ \pi_* e D_n^- \pi^* e}{|E_n|^2} \\ &\quad + \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{\widehat{D}_n^+ \pi_* e |E^{c, -}(\pi^* e)|}{|\widehat{E}_n|^2} \end{aligned} \quad (46)$$

$$\begin{aligned} &\quad + \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{\widehat{D}_n^- \pi_* e |E^{c, +}(\pi_* e)|}{|\widehat{E}_n|^2} \end{aligned} \quad (47)$$

$$\begin{aligned} &\quad + \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{|E^{c, +}(\pi_* e)| |E^{c, -}(\pi^* e)|}{|\widehat{E}_n|^2}. \end{aligned} \quad (48)$$

By Lemma 5.9 and Proposition 5.1 it follows that (48) is $o_{\mathbb{P}}(1)$. For (46) we have

$$\begin{aligned} \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{\widehat{D}_n^+ \pi_* e |E^{c, -}(\pi^* e)|}{|\widehat{E}_n|^2} &\leq \sum_{v \in V_n} \frac{\widehat{D}_n^+ v}{|\widehat{E}_n|} \sum_{w \in V_n} \frac{|E_n^{c, -}(w)|}{|\widehat{E}_n|} \\ &\leq \sum_{w \in V_n} \frac{|E_n^{c, -}(w)|}{|\widehat{E}_n|} = o_{\mathbb{P}}(1), \end{aligned}$$

where the last line is due to $\sum_{v \in V_n} \widehat{D}_n^+ v = |\widehat{E}_n|$. The last equation then follows from Lemma 5.9 and Proposition 5.1. This holds similarly for (47) and hence the result follows. \square

Proof of Lemma 5.10. i) By splitting $|E_n(e)|$ we obtain,

$$\begin{aligned} \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \middle| G_n \right] \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] &= \mathbb{E} \left[\sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{|E_n(e)|}{|E_n|} \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] \\ &= \frac{|\widehat{E}_n|}{|E_n|} \mathbb{E} \left[\sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \frac{|\widehat{E}_n(e)|}{|\widehat{E}_n|} \middle| \widehat{\mathfrak{D}}(G_n) \right] \end{aligned} \quad (49)$$

$$- \frac{1}{|E_n|} \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \mathbb{E} \left[|E_n^c(e)| \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] \quad (50)$$

For (50) we have,

$$\frac{1}{|E_n|} \sum_{e \in V_n^2} I_e^{\alpha, \beta}(k, l) \mathbb{E} \left[|E_n^c(e)| \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] \leq \frac{1}{|E_n|} \sum_{e \in V_n^2} \mathbb{E} \left[|E_n^c(e)| \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right]$$

$$= \frac{1}{|E_n|} \sum_{v \in V_n} |E_n^{c,+}(v)|,$$

which is $o_{\mathbb{P}}(1)$ by Lemma 5.9 and (45). Now, since the conditional expectation in (49) equals (24), it follows from Lemma 5.2 i), Lemma 5.11 and (45) that

$$(49) = \sum_{e \in V_n^2} I_e^{\alpha,\beta}(k, l) \frac{D_n^+ \pi_* e D_n^- \pi^* e}{|E_n|^2} + o_{\mathbb{P}}(1).$$

ii) Splitting both terms $|E_n(e)|$ and $|E_n(f)|$ for $e, f \in V_n^2$ yields,

$$\begin{aligned} & \mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha,\beta}(k, l) \middle| G_n \right]^2 \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] \\ &= \mathbb{E} \left[\sum_{e, f \in V_n^2} I_e^{\alpha,\beta}(k, l) I_f^{\alpha,\beta}(k, l) \frac{|E_n(e)| |E_n(f)|}{|E_n|^2} \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] \\ &= \frac{|\widehat{E}_n|^2}{|E_n|^2} \mathbb{E} \left[\sum_{e, f \in V_n^2} I_e^{\alpha,\beta}(k, l) I_f^{\alpha,\beta}(k, l) \frac{|\widehat{E}(e)| |\widehat{E}(f)|}{|\widehat{E}_n|} \middle| \widehat{\mathfrak{D}}(G_n) \right] \end{aligned} \quad (51)$$

$$+ \sum_{e, f \in V_n^2} I_e^{\alpha,\beta}(k, l) I_f^{\alpha,\beta}(k, l) \mathbb{E} \left[\frac{|E_n^c(e)| |E_n^c(f)|}{|E_n|^2} \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] \quad (52)$$

$$- \sum_{e, f \in V_n^2} I_e^{\alpha,\beta}(k, l) I_f^{\alpha,\beta}(k, l) \mathbb{E} \left[\frac{|E_n^c(e)| |\widehat{E}_n(f)|}{|E_n|^2} \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] \quad (53)$$

$$- \sum_{e, f \in V_n^2} I_e^{\alpha,\beta}(k, l) I_f^{\alpha,\beta}(k, l) \mathbb{E} \left[\frac{|E_n^c(f)| |\widehat{E}_n(e)|}{|E_n|^2} \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] \quad (54)$$

Recognizing the conditional expectation in (51) as (25), then using first Lemma 5.2 ii) and then Lemma 5.11 and (45), it follows that (51) equals

$$\left(\sum_{e \in V_n^2} I_e^{\alpha,\beta}(k, l) \frac{D_n^+ \pi_* e D_n^- \pi^* e}{|E_n|^2} \right)^2 + o_{\mathbb{P}}(1).$$

It remains to show that (52)-(54) are $o_{\mathbb{P}}(1)$. For (52) we have

$$\sum_{e, f \in V_n^2} I_e^{\alpha,\beta}(k, l) I_f^{\alpha,\beta}(k, l) \mathbb{E} \left[\frac{|E_n^c(e)| |E_n^c(f)|}{|E_n|^2} \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] \leq \left(\frac{1}{|E_n|} \sum_{v \in V_n} |E_n^{c,+}(v)| \right)^2 = o_{\mathbb{P}}(1)$$

by Lemma 5.9 and (45). Since (53) and (54) are symmetric we will only consider the latter:

$$\begin{aligned} & \sum_{e, f \in V_n^2} I_e^{\alpha,\beta}(k, l) I_f^{\alpha,\beta}(k, l) \mathbb{E} \left[\frac{|E_n^c(f)| |\widehat{E}_n(e)|}{|E_n|^2} \middle| \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] \\ & \leq \left(\sum_{f \in V_n^2} \frac{|E_n^c(f)|}{|E_n|} \right) \frac{1}{|E_n|} \sum_{e \in V_n^2} \mathbb{E} \left[|\widehat{E}_n(e)| \middle| \widehat{\mathfrak{D}}(G_n) \right] \\ &= \left(\sum_{v \in V_n} \frac{|E_n^+(v)|}{|E_n|} \right) \frac{|\widehat{E}_n|}{|E_n|} = o_{\mathbb{P}}(1). \end{aligned}$$

Here, for the last line, we used $\sum_{e \in V_n^2} \mathbb{E} \left[|\widehat{E}_n(e)| \middle| \widehat{\mathfrak{D}}(G_n) \right] = |\widehat{E}_n|$, and then Lemma 5.9 and (45). \square

A straightforward adaptation of the proof of Proposition 5.4, using Lemma 5.10 instead of Lemma 5.2, yields the following result.

Proposition 5.12. *Let $\{G_n\}_{n \in \mathbb{N}}$ be a sequence of ECM graphs with $|V_n| = n$ and $\alpha, \beta \in \{+, -\}$. Then there exist integer valued random variables \mathcal{D}^α and \mathcal{D}^β such that for each $k, l \in \mathbb{N}$, as $n \rightarrow \infty$,*

$$\mathbb{E} \left[\mathbb{E} \left[I_{\mathcal{E}_n}^{\alpha, \beta}(k, l) \mid G_n \right] \mid \widehat{\mathfrak{D}}(G_n), \mathfrak{D}(G_n) \right] \xrightarrow{\mathbb{P}} \mathbb{P}(\mathcal{D}^\alpha = k) \mathbb{P}(\mathcal{D}^\beta = l),$$

where the distributions of \mathcal{D}^α and \mathcal{D}^β are given in Table 1.

We can now again use a second moment argument to get the convergence result for the three rank correlations in the Erased Configuration Model. We omit the proof since the computation of the variance follows the exact same steps as those in Proposition 5.5, where now, instead of only conditioning on $\widehat{\mathfrak{D}}(G_n)$, we also condition on $\mathfrak{D}(G_n)$ and use Lemma 5.10.

Theorem 5.13. *Let $\{G_n\}_{n \in \mathbb{N}}$ be a sequence of ECM graphs with $|V_n| = n$ and $\alpha, \beta \in \{+, -\}$. Then, as $n \rightarrow \infty$,*

$$\rho_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} 0, \quad \bar{\rho}_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} 0 \quad \text{and} \quad \tau_\alpha^\beta(G_n) \xrightarrow{\mathbb{P}} 0.$$

This theorem shows that even when the variance of the degree sequences is infinite, one can construct a random graph for which the degree-degree dependencies, measured by rank correlations, converge to zero in the infinite graph size limit. Therefore this model can be used as a null model for such dependencies.

Acknowledgments:

We like to thank an anonymous referee for thoroughly reading our manuscript and giving constructive comments and suggestions for improvement.

This work is supported by the EU-FET Open grant NADINE (288956).

Appendix A Continuization

In this appendix we will establish several relations between the distribution functions of integer valued random variables and their continuations, using the functions \mathcal{F} and \mathcal{H} defined in (1) and (2), respectively.

Let $\tilde{X} = X + U$ be as in Definition 2.2, take $k \in \mathbb{Z}$ and define $I_k = [k, k+1)$. Then for $x \in I_k$,

$$F_{\tilde{X}}(x) = (x - k)F_X(k) + (k + 1 - x)F_X(k - 1). \quad (55)$$

As a consequence, it follows that for $x \in I_k$,

$$dF_{\tilde{X}}(x) = (F_X(k) - F_X(k - 1)) dx = \mathbb{P}(X = k) dx. \quad (56)$$

These identities capture the essential relations between X and its continuization \tilde{X} . As a first result we have the following.

Lemma A.1. *Let X be an integer valued random variable and $m \in \mathbb{N}$. Then,*

$$\mathbb{E} \left[F_{\tilde{X}}(\tilde{X})^m \right] = \frac{1}{m+1} \sum_{i=0}^m \mathbb{E} \left[F_X(X)^i F_X(X-1)^{m-i} \right].$$

Proof. Using (55) we obtain,

$$\begin{aligned} \int_{I_k} F_{\tilde{X}}(x)^m dx &= \int_{I_k} ((x - k)F_X(k) + (k + 1 - x)F_X(k - 1))^m dx \\ &= \sum_{i=0}^m \binom{m}{i} F_X(k)^i F_X(k - 1)^{m-i} \int_0^1 (y)^i (1 - y)^{m-i} dy \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=0}^m \frac{m!}{i!(m-i)!} F_X(k)^i F_X(k-1)^{m-i} \frac{\Gamma(i+1)\Gamma(m-i+1)}{\Gamma(m+2)} \\
&= \frac{1}{m+1} \sum_{i=0}^m F_X(k)^i F_X(k-1)^{m-i},
\end{aligned}$$

Combining this with (56), we get

$$\begin{aligned}
\mathbb{E} [F_{\tilde{X}}(\tilde{X})^m] &= \sum_{k \in \mathbb{Z}} \int_{I_k} F_{\tilde{X}}(x)^m dF_{\tilde{X}}(x) \\
&= \sum_{k \in \mathbb{Z}} \int_{I_k} F_{\tilde{X}}(x)^m \mathbb{P}(X=k) dx \\
&= \frac{1}{m+1} \sum_{i=0}^m \mathbb{E} [F_X(X)^i F_X(X-1)^{m-i}].
\end{aligned}$$

□

As a direct consequence of Lemma A.1 we get

$$\frac{1}{2} = \mathbb{E} [F_{\tilde{X}}(\tilde{X})] = \frac{1}{2} \mathbb{E} [\mathcal{F}_X(X)], \quad (57)$$

relating $F_{\tilde{X}}$ to \mathcal{F}_X . Similar to (55), if Z is a random element independent of X , we get for $x \in I_k$,

$$F_{\tilde{X}|Z}(x) = (x-k)F_{X|Z}(k) + (k+1-x)F_{X|Z}(k-1). \quad (58)$$

Applying (58) in a similar way as (55) we arrive at an extension of Lemma A.1. The proof is elementary, hence omitted.

Proposition A.2. *Let X be an integer valued random variable and Z a random element independent of the continuous part of \tilde{X} . Then*

- i) $\mathbb{E} [F_{\tilde{X}}(\tilde{X}) | Z] = \frac{1}{2} \mathbb{E} [\mathcal{F}_X(X) | Z]$, a.s.;
- ii) $F_{\tilde{X}|Z}(\tilde{X}) = \frac{1}{2} \mathcal{F}_{X|Z}(X)$, a.s.

The following results are extensions of the previous ones to the case of two integer valued random variables X and Y . We will state these without proofs, since these are either straightforward extensions of those for the case of a single random variable or follow from elementary calculations and the previous results.

Lemma A.3. *Let X, Y be integer valued random variables. Then,*

- i) $\mathbb{E} [F_{\tilde{X}}(\tilde{X}) F_{\tilde{Y}}(\tilde{Y})] = \frac{1}{4} \mathbb{E} [\mathcal{F}_X(X) \mathcal{F}_Y(Y)]$,
- ii) $\mathbb{E} [H_{\tilde{X}, \tilde{Y}}(\tilde{X}, \tilde{Y})] = \frac{1}{4} \mathbb{E} [\mathcal{H}_{X, Y}(X, Y)]$.

Proposition A.4. *Let X, Y be integer valued random variables and let Z be a random variable independent of the uniform parts of \tilde{X} and \tilde{Y} . Then*

- i) $\mathbb{E} [\tilde{F}_{\tilde{X}}(\tilde{X}) \tilde{F}_{\tilde{Y}}(\tilde{Y}) | Z] = \frac{1}{4} \mathbb{E} [\mathcal{F}_X(X) \mathcal{F}_Y(Y) | Z]$ a.s.;
- ii) $H_{\tilde{X}, \tilde{Y}|Z}(\tilde{X}, \tilde{Y}) = \frac{1}{4} \mathcal{H}_{X, Y|Z}(X, Y)$ a.s.

References

- [1] Marián Boguná, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic spreading in complex networks with degree correlations. *arXiv preprint cond-mat/0301149*, 2003.
- [2] Ningyuan Chen and Mariana Olvera-Cravioto. Directed random graphs with given degree distributions. *Stochastic Systems*, 3(1):147–186, 2013.
- [3] Fan Chung, Linyuan Lu, and Van Vu. Spectra of random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 100(11):6313–6318, 2003.
- [4] Michel Denuit and Philippe Lambert. Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, 93(1):40–57, 2005.
- [5] SN Dorogovtsev, AL Ferreira, AV Goltsev, and JFF Mendes. Zero pearson coefficient for strongly correlated growing trees. *Physical Review E*, 81(3):031135, 2010.
- [6] Takehisa Hasegawa, Taro Takaguchi, and Naoki Masuda. Observability transitions in correlated networks. *Physical Review E*, 88(4):042809, 2013.
- [7] Nelly Litvak and Remco van der Hofstad. Degree-degree correlations in random graphs with heavy-tailed degrees. *arXiv preprint arXiv:1202.3071*, 2012.
- [8] Nelly Litvak and Remco van der Hofstad. Uncovering disassortativity in large scale-free networks. *Physical Review E*, 87(2):022801, 2013.
- [9] Xiao Fan Liu and Chi Kong Tse. Impact of degree mixing pattern on consensus formation in social networks. *Physica A: Statistical Mechanics and its Applications*, 407:1–6, 2014.
- [10] Sergei Maslov, Kim Sneppen, and Alexei Zaliznyak. Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A: Statistical Mechanics and its Applications*, 333:529–540, 2004.
- [11] Mhamed Mesfioui and Abdelouahid Tajar. On the properties of some nonparametric concordance measures in the discrete case. *Nonparametric Statistics*, 17(5):541–554, 2005.
- [12] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.
- [13] Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.
- [14] Juyong Park and Mark EJ Newman. Origin of degree correlations in the internet and other networks. *Physical Review E*, 68(2):026112, 2003.
- [15] Animesh Srivastava, Bivas Mitra, Fernando Peruani, and Niloy Ganguly. Attacks on correlated peer-to-peer networks: An analytical study. pages 1076–1081, 2011.
- [16] Remco Van Der Hofstad. Random graphs and complex networks. *Unpublished manuscript*, 2007.
- [17] Pim van der Hoorn and Nelly Litvak. Degree-degree correlations in directed networks with heavy-tailed degrees. *arXiv preprint arXiv:1310.6528*, 2013.

Phase transitions for scaling of structural correlations in directed networks.

Pim van der Hoorn* Nelly Litvak†

April 8, 2015

Abstract

Analysis of degree-degree dependencies in complex networks, and their impact on processes on networks requires null models, i.e. models that generate uncorrelated scale-free networks. Most models to date however show structural negative dependencies, caused by finite size effects. We analyze the behavior of these structural negative degree-degree dependencies, using rank based correlation measures, in the directed Erased Configuration Model. We obtain expressions for the scaling as a function of the exponents of the distributions. Moreover, we show that this scaling undergoes a phase transition, where one region exhibits scaling related to the natural cut-off of the network while another region has scaling similar to the structural cut-off for uncorrelated networks. By establishing the speed of convergence of these structural dependencies we are able to asses statistical significance of degree-degree dependencies on finite complex networks when compared to networks generated by the directed Erased Configuration Model.

1 Introduction

The tendency of nodes in a network to be connected to nodes of similar large or small degree, called network assortativity, degree mixing or degree-degree dependency, is an important characterization of the topology of the network, influencing many processes on the network. It has received significant attention in the literature, for instance in the field of network stability [29], attacks on P2P networks [25] and epidemics [2, 3].

An important method to analyze these degree-degree dependencies or their influence on other network properties or processes on the network, is to compare results to an average over several instances of similar networks with neutral mixing. These null models often come in two flavors. The first approach is to sample from graphs with the same degree sequence but neutral mixing. A widely accepted methodology for such sampling is through the local rewiring model, [17], which takes the original network and randomly swaps edges until a randomized version is attained. The disadvantage of these methods is that they have no theoretical performance guarantees. The second approach is to generate a random graph with neutral mixing, which preserves basic features, such as the degree distribution. A well known model of this type is the Configuration Model [6, 19, 21]. Here the degrees of vertices are drawn independently from the given distribution, under the restriction that the total sum of degrees is even. Then the stubs are paired uniformly at random to form edges. If we want to obtain a simple graph in this way, we can either rewire till a simple graph is generated (Repeated Configuration Model), or we remove the excess edges and self loops (Erased Configuration Model).

We note that there are many other methods, that generate simple random graphs and have theoretically established performance guarantees, for example, the sequential algorithm in [1] that creates a random graph with given degrees, or a grand-canonical model in [24] that generates a

*University of Twente, w.l.f.vanderhoorn@utwente.nl

†University of Twente, n.litvak@utwente.nl

Wikipedia	N	$N^{1/2}$	γ_+	γ_-	$\max D^+$	$\max D^-$
DE	1,532,978	1,238	1.80	1.05	5,032	118,064
EN	4,212,493	2,052	2.14	1.20	8,104	432,629
IT	1,017,953	1,009	1.96	1.05	5,212	91,588
NL	1,144,615	1,070	1.82	1.10	10,175	102,450
PL	949,153	974	1.90	1.04	4,100	112,537

Table 1: Basic degree characteristics of Wikipedia networks. The exponents of the degree distributions are estimated using the implementation of the techniques from [9] by Peter Bloem, <http://github.com/Data2Semantics/powerlaws>.

graph with given average degrees using a maximum-entropy method. However, to the best of our knowledge, none of these methods has an efficient implementation, that is feasible for a truly large network, such as Wikipedia or Twitter.

Although for both local rewiring and the Configuration Model neutral mixing is expected, since there is no preference in connecting two vertices, negative correlations are observed, [7, 18, 22], for scale-free networks with infinite variance of degrees, i.e. where the degree distribution satisfies

$$P(k) \sim k^{-(\gamma+1)}, \quad 1 < \gamma \leq 2. \quad (1)$$

In [18] this phenomenon is explained by observing that if one allows at most one edge between two vertices, nodes with large degree must connect to nodes of small degree because there are simply not enough distinct large nodes to connect to. A similar explanation is given in [7]. Here, however, this is then related to the difference in scaling between the *natural* and *structural cut-off* of the network. The former is defined [10] as the degree value k_c , of which, on average, only one instance is observed:

$$N \int_{k_c}^{\infty} P(k) dk \sim 1. \quad (2)$$

The structural cut-off is defined as the value k_s for which the ratio between the average number of edges that connect any two vertices of degree k_s , and the maximum possible number of such edges in a simple graph, is 1. For networks with degree distribution (1) it follows from (2) that the natural cut-off scales as $N^{1/\gamma}$, while the structural cut-off for uncorrelated networks scales as, see [4], $N^{1/2}$. Therefore, when $\gamma < 2$, the natural cut-off scales at a slower rate which in turn gives rise to structural negative correlations.

To remedy these finite size effects the authors of [7] propose an Uncorrelated Configuration Model. This model follows the same procedure as the regular Configuration Model, with the addition that the sampled degrees are bounded, $m \leq k_i \leq N^{1/2}$. Experiments in [7] indeed show that these networks are uncorrelated. However, many scale-free networks, for instance Twitter, have nodes who's degree is of larger order than $N^{1/2}$, which is a characteristic property of scale-free graphs. For example, Table 1 displays the characteristics of Wikipedia networks for different languages. Here we see that the maximum out-degree could be considered to be of order $N^{1/2}$, while the maximum in-degree is definitely of a much larger scale. Therefore, randomized versions of these networks, generated by the Uncorrelated Configuration Model, do not have the same basic degree characteristics as the original network, since the maximum degree is restricted. Hence, they are less suitable for comparison of the degree-degree dependencies.

In this paper we consider the directed Erased Configuration Model, [8], where after the pairing self loops are removed and multiple edges are merged. This model was shown to have neutral mixing in the infinite network size limit, see [27] Section 5. Therefore, from a purely mathematical point of view, it is a null model for degree-degree dependencies in the limit. Moreover, asymptotically, the degree distributions are preserved and hence, all basic degree characteristics. Still, for finite sizes, structural dependencies are present.

Rather than trying to control these correlations, our goal is to evaluate their magnitude and investigate their size dependence. We obtain the scaling for the structural correlations in the

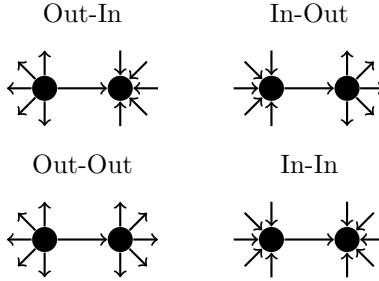


Figure 1: The four different degree-degree dependency types in directed networks.

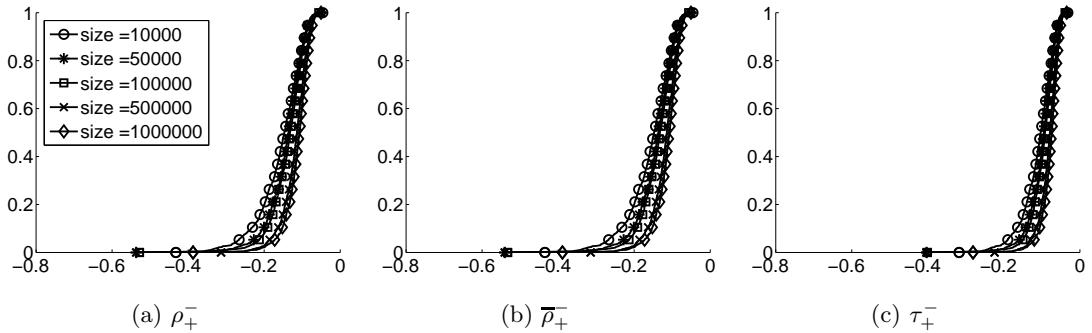


Figure 2: Plots of the empirical cumulative distribution of ρ_+^- , $\bar{\rho}_+^-$ and τ_+^- for ECM graphs of different sizes with $\gamma_\pm = 1.2$. Each plot is based on 10^3 realizations of the model.

ECM, in terms of the power law exponents of the in- and out-degrees. In particular, we show that this scaling undergoes an interesting phase transition, and can be dominated by terms related to either the structural or the natural cut-off of the network. To the best of our knowledge, this is the first study that provides a systematic mathematical characterization for the magnitude of negative correlations in a simple graph with neutral mixing.

By determining the scaling of the structural correlations we can assess the significance of measured correlations as well as their influence on network processes, on real world networks of finite size, by comparing them to the directed Erased Configuration Model. This approach has the advantage of preserving the degree characteristics of the original network, it can be easily implemented and applied to all networks with scale free-degree distributions and finite expectation.

2 Degree-degree dependencies in random directed networks

We analyze degree-degree dependencies in random directed networks of size N , where the distribution of the out- and in-degree (D^+ , D^-) follow, respectively,

$$P^+(k) \sim k^{-(1+\gamma_+)} \text{ and } P^-(\ell) \sim \ell^{-(1+\gamma_-)}, \quad \gamma_\pm > 1. \quad (3)$$

In directed networks one can consider four types of degree-degree dependencies, depending on the choice of the degree type on both sides of an edge, see Figure 1. For the remainder of this paper we denote by E the number of edges and adopt the notation style from [12, 28] to index the degree types by $\alpha, \beta \in \{+, -\}$.

A common measure for degree-degree dependencies, introduced in [20], computes Pearson's correlation coefficients on the joint data $(D_i^\alpha, D_j^\beta)_{i \rightarrow j}$, where the indices run over all i, j for which there is an edge $i \rightarrow j$.

However, Pearson's correlation coefficients are unable to measure strong negative degree-degree dependencies in large networks where the variance of the degrees is infinite, as was shown for

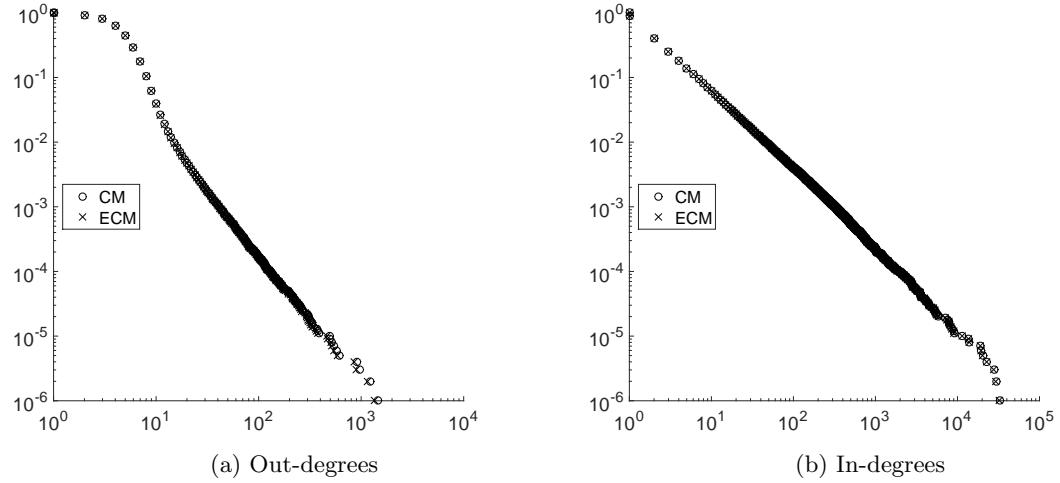


Figure 3: Plots of the out- and in-degree distribution, on log-log scale, for a graph generated by the ECM, of size 10^6 with $\gamma_+ = 1.9$ and $\gamma_- = 1.2$, before (CM) and after (ECM) the removing of edges.

undirected networks in [16, 11] and for directed networks in [28]. Since our interest is mainly in networks in the infinite variance domain, i.e. $1 < \gamma_{\pm} \leq 2$, we need different measures. In [28] it was suggested to use rank correlations, related to Spearman's rho [23] and Kendall's tau [14], to measure degree-degree dependencies.

Spearman's rho computes Pearson's correlation coefficient on the ranks of $(D_i^\alpha, D_j^\beta)_{i \rightarrow j}$ rather than their actual values. Since this data will contain many ties, one needs to use ranking schemes that deal with these ties. In [28] two such schemes are considered, resolving ties at random and assigning an average rank to tied values, which give two correlation measures denoted by ρ_α^β and $\bar{\rho}_\alpha^\beta$, respectively. Here, the subscript index denotes the degree type of the source, while the superscript index denotes the degree type of the target of a directed edge. For instance, ρ_+^- denotes Spearman's rho for the Out-In dependency. The second rank correlation measure, Kendall's tau τ_α^β , calculates the normalized number of swaps needed to match the ranks of the joint data.

Exact formulas for these three measures, in terms of the degrees, are given in [28]. In [27] formulas are given in terms of the empirical distributions of D^α and D^β and their joint distribution, evaluated at (D_i^α, D_j^β) for an edge $i \rightarrow j$ selected uniformly at random. From these it follows that if the network has neutral mixing, then ρ_α^β and τ_α^β are similar, while ρ_α^β and $\bar{\rho}_\alpha^\beta$ differ by a term of $O(1)$, which does not influence the scaling. To illustrate this we plotted the empirical cdf's of ρ_+^- , $\bar{\rho}_+^-$ and τ_+^- for a collection of ECM graphs in Figure 2; where we clearly observe the similar behavior of the three measures. Therefore, for the analysis of degree-degree dependencies, we will only consider ρ_α^β , which corresponds to Spearman's rho where ties are resolved uniformly at random.

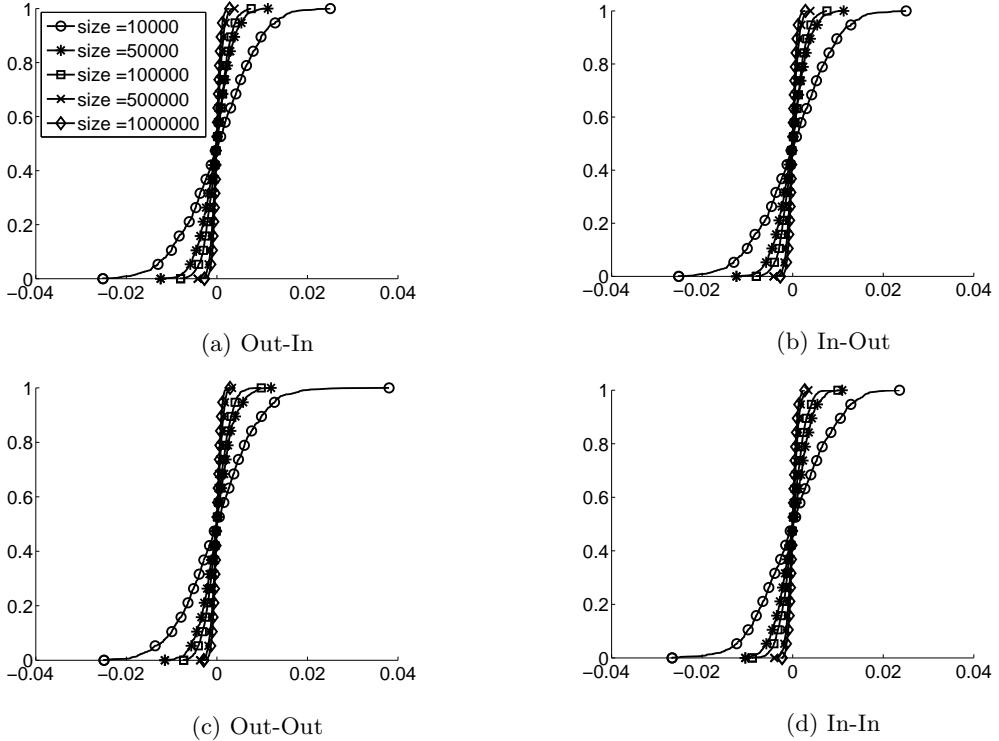


Figure 4: Plots of the empirical cumulative distribution of ρ_α^β for all four degree-degree dependency types for ECM graphs of different sizes with $\gamma_\pm = 2.1$. Each plot is based on 10^3 realizations of the model.

3 The directed Erased Configuration Model

The directed Configuration Model (CM) starts with degree sequences $(D_i^+, D_i^-)_{1 \leq i \leq N}$ that satisfy,

$$E = \sum_{i=1}^N D_i^\pm \sim \mu N \quad (4a)$$

$$\sum_{i=1}^N D_i^+ D_i^- \sim \mu^2 N \quad (4b)$$

$$\sum_{i=1}^N (D_i^\pm)^p \sim N^{p/\gamma_\pm}, \quad p > \gamma_\pm, \quad (4c)$$

for some $\mu > 0$. The stubs are then paired at random to form edges. This will in general constitute a graph with self-loops and multiple edges between nodes. If the degree variance is finite, then the probability of generating a simple graph is bounded away from zero and thus, by repeating the pairing step until such a graph is generated, we get a network randomly sampled from all networks of given size and degree sequences. This is called the Repeated Configuration Model (RCM).

When the variance of the degrees is infinite, the probability of generating a simple graph converges to zero as the graph size increases, and therefore we need to enforce that the resulting graph is simple. For this we use the Erased Configuration Model (ECM), where, during the pairing, a new edge is removed if it already exists or if it is a self loop. Although this seems to be a strong alteration of the initial degree sequence, asymptotically, the degrees of the resulting

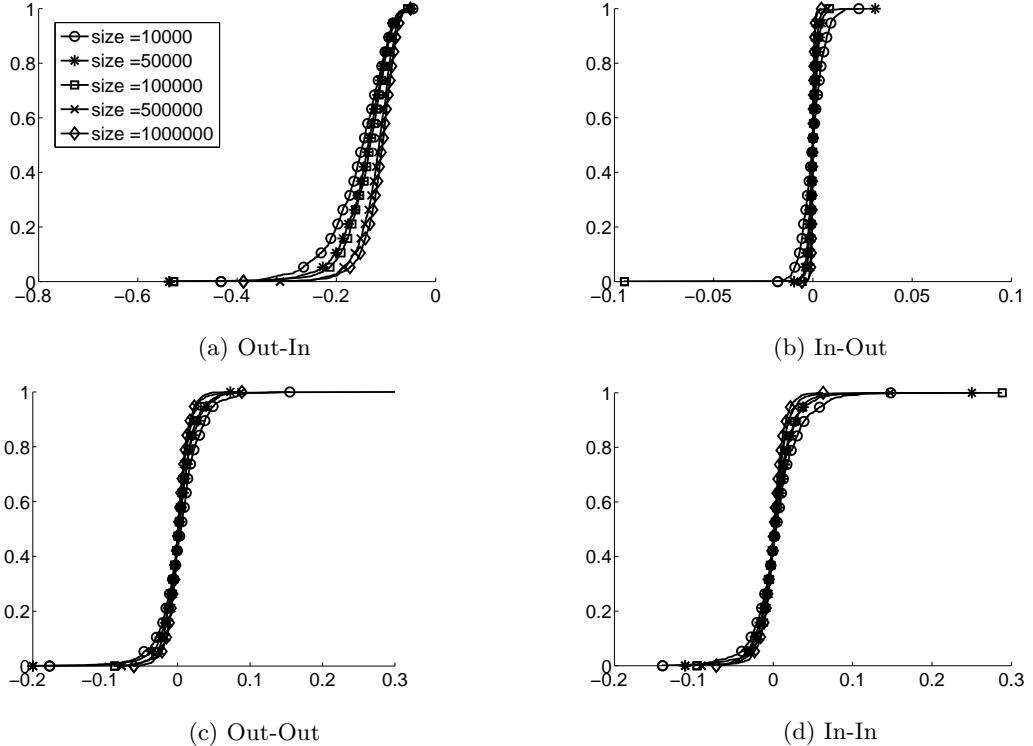


Figure 5: Plots of the empirical cumulative distribution of ρ_α^β for all four degree-degree dependency types for ECM graphs of different sizes with $\gamma_\pm = 1.2$. Each plot is based on 10^3 realizations of the model.

network still follow the same distribution, see [8]. For illustration, in Figure 3, we plotted the degree distributions of an ECM graph of size 10^6 before and after the removing of edges. Clearly there is hardly any difference between the two distributions. In particular the degree sequences of ECM graphs still satisfy (4). Unlike many other methods, random pairing of the stubs can be implemented very efficiently for even billions of nodes. Moreover, the ECM is computationally less expensive than RCM, since we do not need to repeat the pairing. Therefore we suggest to use the ECM as a standard null-model. In the rest of the paper we will characterize the structural dependencies in the ECM.

4 Degree-degree dependencies in the ECM

It is clear that when we use the CM, i.e. allow for multiple edges and self loops, then our graphs will have neutral mixing since all stubs are connected completely at random. For the ECM however, we remove edges to make the graph simple, which has been shown [18, 7] to give rise to negative correlations. Nevertheless, the ECM has asymptotically neutral mixing, which can be shown as follows.

Let E_{ij} be the matrix counting the number of edges between i and j after the pairing and let E_{ij}^c denote the matrix counting the number of removed edges between i and j by the ECM. Then for the CM it holds that $D_i^+ = \sum_{j=1}^N E_{ij}$ while for the ECM we have $D_i^{++} = \sum_{j=1}^N (E_{ij} - E_{ij}^c)$. Therefore, the difference between the empirical distributions of D_i^α and D_j^β , for an edge $i \rightarrow j$ sampled at random, in the CM and ECM, will be of the order $\sum_{i,j=1}^N E_{ij}^c / E$, whose average, with respect to the degree sequences, converges to zero [27],

N	$\langle \rho_+^- \rangle$	$\langle \rho_-^+ \rangle$	$\langle \rho_+^+ \rangle$	$\langle \rho_-^- \rangle$
10000	-0.1568	-0.0001	0.0039	0.0048
50000	-0.1439	0.0001	0.0014	0.0029
100000	-0.1388	-0.0001	0.0026	0.0028
500000	-0.1198	0.0001	0.0011	0.0017
1000000	-0.1131	0.0000	0.0009	0.0002

Table 2: The average values for ρ for all four degree-degree dependencies types, for ECM graphs of different sizes, with $\gamma_{\pm} = 1.2$, based on 10^3 realizations of the model.

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i,j=1}^N \langle E_{ij}^c \rangle = 0. \quad (5)$$

This implies that the values of ρ_{α}^{β} for an ECM graph will converge to that of a CM graph, hence, asymptotically, $\rho_{\alpha}^{\beta} = 0$ and also $\overline{\rho}_{\alpha}^{\beta} = 0 = \tau_{\alpha}^{\beta}$, for the ECM.

However, for finite realizations in the infinite variance regime, negative correlations are still observed. To illustrate this we plotted the empirical cumulative distribution functions of ρ_{α}^{β} for graphs generated by the ECM with, both, finite and infinite degree variance, see Figure 4 and Figure 5, respectively. In addition, Table 2 contains the average values for all four correlation types in the infinite variance regime. One immediately observes that the Out-In dependency in ECM graphs with infinite variance, Figure 5a, displays strong structural negative correlations which decrease as the network grows, while for other three dependency types the values are concentrated around zero. Moreover, we see, Figure 4, that all four dependency types behave similar when the variance of the degrees is finite.

These negative Out-In correlations (ρ_+^-) can be explained by first observing that multiple edges are more likely to start in a node of large out-degree and end in a node of large in-degree, since these are more likely to be sampled. Now, consider the algorithm as first connecting all stubs at random and then removing self loops and merging multiple edges. By construction, immediately after the pairing the network will have neutral mixing. When merging multiple edges we will often delete connections from nodes of large out-degree to nodes of large in-degree. Such edges have contributed positively into ρ_+^- , thus, deleting them will shift ρ_+^- from zero in the CM to a negative value in the ECM. The other three dependency types are not effected since the out- and in-degree of a node in the ECM are independent.

Motivated by the analysis in this section, we will further focus on the behavior of ρ_+^- in the infinite-variance case, $1 < \gamma_+, \gamma_- \leq 2$, as the only scenario where we observe prominent structural correlations. We will discuss other scenarios in Section 6.

5 Scaling of the Out-In degree-degree dependency in the ECM

We will determine the scaling of ρ_+^- as a function of the exponents γ_{\pm} . That is, we will find coefficients $f(\gamma_+, \gamma_-)$ such that

$$\frac{\rho_+^- - \langle \rho_+^- \rangle}{N^{f(\gamma_+, \gamma_-)}}$$

converges to some limiting distribution. Here the expectation $\langle \rho_+^- \rangle$ is taken over all possible graphs of size N , generated by the ECM, with degree sequences satisfying (4). We note that although $\langle \rho_+^- \rangle$ is of similar order as the typical spreading of ρ_+^- , the latter, which we are going to evaluate, will define the magnitude of the structural negative correlations.

We obtain the scaling exponents $f(\gamma_+, \gamma_-)$ by establishing upper bounds on the scaling, and then show empirically that these bounds are tight. The scaling is an important quantity, characterizing the spread around the sample mean of ρ_+^- as a function of N . Roughly, this tells us how

much the measured values on a ECM graph of size N can deviate from the average and therefore enable us to asses the significance of the measured correlations of the corresponding real world networks.

5.1 Scaling of the erased number of edges

As we discussed in the previous section, the structural negative correlations appear after multiple edges and self-loops are erased. Hence, part of the scaling of ρ_+^- comes from the scaling of the average total number of erased edges. The latter scaling has a phase transition, which we will show by establishing two different upper bounds.

For the first upper bound, observe that

$$\sum_{i,j=1}^N E_{ij}^c = \sum_{i=1}^N S_{ii} + \sum_{i,j=1}^N M_{ij}, \quad (6)$$

where S is the diagonal matrix counting the number of self loops and M is the zero diagonal matrix that counts the excess edges, so $M_{ij} = k > 0$ means that $E_{ij} = k + 1$. For the self loops it holds that

$$\langle S_{ii} \rangle = \frac{D_i^+ D_i^-}{E}. \quad (7)$$

If we now take the total number of pairs of edges between i and j as an upper bound for the M_{ij} , then

$$\langle M_{ij} \rangle \leq \frac{(D_i^+)^2 (D_j^-)^2}{E^2}. \quad (8)$$

Applying this to (6) we get

$$\sum_{i,j=1}^N \frac{\langle E_{ij}^c \rangle}{E} \leq \frac{\sum_{i,j=1}^N (D_i^+)^2 (D_j^-)^2}{E^3} + \frac{\sum_{i=1}^N D_i^+ D_i^-}{E^2}. \quad (9)$$

We remark that if the second moment of both the out- and in-degree exists, then this upper bound scales as N^{-1} . When this is not the case, we get the scaling from (4) as

$$\frac{1}{E} \sum_{i,j=1}^N \langle E_{ij}^c \rangle = O\left(N^{(2/\gamma_+) + (2/\gamma_-) - 3}\right). \quad (10)$$

The upper bound (10) is rather crude in the sense that for certain $1 < \gamma_{\pm} \leq 2$, we have $(2/\gamma_+) + (2/\gamma_-) > 3$ so that the right-hand side of (10) becomes infinite as $N \rightarrow \infty$.

To get a more precise upper bound let $p(n, m, L)$ denote the probability that none of the outbound stubs from a set of size n connect to an inbound stub from a set of size m , given that the total number of available stubs is L . We will establish a recursive relation for $p(D_i^+, D_j^-, E)$ by adopting the analysis from [26], Section 4. Similarly we get, by conditioning on whether we pick an inbound stub of i or not,

$$p(D_i^+, D_j^-, E) \leq \left(1 - \frac{D_j^-}{E}\right) p(D_i^+ - 1, D_j^-, E - 1),$$

where the upper bound comes from neglecting the event $D_i^+ + D_j^- > E$, in which case $p(D_i^+, D_j^-, E) = 0$. Continuing the recursion yields

$$p(D_i^+, D_j^-, E) \leq \prod_{k=0}^{D_i^+-1} \left(1 - \frac{D_j^-}{E - i}\right),$$

and a first order Taylor expansion then gives

$$p(D_i^+, D_j^-, E) \leq e^{-D_i^+ D_j^- / E}. \quad (11)$$

Now, recall that E_{ij} denotes the total number of edges between i and j in the CM, before the removal step. Therefore,

$$\langle E_{ij}^c \rangle = \langle E_{ij} \rangle - (1 - p(D_i^+, D_j^-, E)).$$

Since $E = \sum_{i,j=1}^N \langle E_{ij} \rangle$ it follows that

$$\frac{1}{E} \sum_{i,j=1}^N \langle E_{ij}^c \rangle = 1 - \frac{N^2}{E} + \frac{1}{E} \sum_{i,j=1}^N p(D_i^+, D_j^-, E) \quad (12)$$

Hence, by plugging (11) into (12) we arrive at the following upper bound for the total average number of erased edges,

$$\frac{1}{E} \sum_{i,j=1}^N \langle E_{ij}^c \rangle \leq 1 - \frac{N^2}{E} + \frac{1}{E} \sum_{i,j=1}^N e^{-D_i^+ D_j^- / E}. \quad (13)$$

The right hand side of (13) can be slightly rewritten to obtain the more informative expression

$$\frac{N^2}{E} \left(\frac{1}{E} \sum_{i,j=1}^N \frac{D_i^+ D_j^-}{N^2} - 1 + \sum_{i,j=1}^N \frac{e^{-(D_i^+ D_j^-)/E}}{N^2} \right). \quad (14)$$

Next, we note that (14) can be seen as an empirical form of

$$\frac{N}{\mu} \left(\frac{1}{N\mu} \langle \xi \rangle - 1 + \left\langle e^{-\xi/(N\mu)} \right\rangle \right), \quad (15)$$

where, using $a \wedge b$ to denote the minimum, ξ has distribution

$$P_\xi(k) \sim k^{-(\gamma_+ \wedge \gamma_-) - 1}.$$

From a classical Tauberian Theorem for regularly varying random variables, see for instance [15], it follows that (15) scales as $N^{1-(\gamma_+ \wedge \gamma_-)}$. However, by the Central limit Theorem for stable random variables, see [30], we have

$$\left| \frac{1}{E} \sum_{i,j=1}^N \frac{D_i^+ D_j^-}{N^2} - \frac{\mu}{N} \right| = O\left(n^{-2+1/(\gamma_+ \wedge \gamma_-)}\right).$$

Therefore, since $E \sim \mu N$, the difference between (14) and (15) scales as $N^{-1+1/(\gamma_+ \wedge \gamma_-)}$. This last term dominates (15) when $1 < \gamma_\pm \leq 2$, hence it follows that

$$\frac{1}{E} \sum_{i,j=1}^N \langle E_{ij}^c \rangle = O(N^{-1+1/(\gamma_+ \wedge \gamma_-)}) \quad (16)$$

The scaling in (16) is related to that of the structural cut off described in [4], adjusted to the setting of directed networks with degree distributions (3). Moreover, comparing (16) to (10) we observe a phase transition, with respect to the tail exponents γ_\pm of the degree distributions, in the scaling of the average total number of removed edges in the ECM, which will induce a phase transition in the scaling of the Out-In degree-degree dependency.

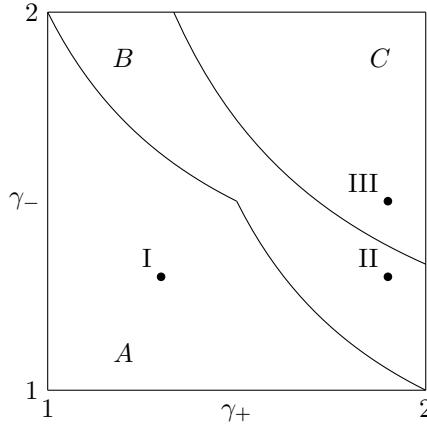


Figure 6: Plot of the different scaling regimes for ρ_+^- . The scaling terms for each of the three regions can be found in Table 3. The Roman numerals indicate the three different choices of γ_+ and γ_- , used in Figure 7 and 8, to illustrate the different regimes.

Region	$f(\gamma_+, \gamma_-)$
A	$1/(\gamma_+ \wedge \gamma_-) - 1$
B	$(2/\gamma_+) + (2/\gamma_-) - 3$
C	$-1/2$

Table 3: The three scaling terms for ρ_+^- for each of the three regions, displayed in Figure 6

5.2 Phase transitions for the Out-In degree-degree dependency

First we remark that for the CM, the empirical distribution of the degrees on both sides of a randomly sampled edge converges to the distribution of two independent random variables as N^{-1} , see [27]. Because Spearman's rho and Kendall's tau on independent joint measurements are normal statistics [13], the scaling of their average is $N^{-1/2}$. Hence ρ_α^β for CM graphs scales as $N^{-1/2}$. Since an ECM graph is basically a CM graph where multiple edges are merged and self-loops are removed, it follows that the distributions for the degrees on both side of a randomly chosen edge differ from those of the CM by terms of the order $\sum_{i,j=1}^N E_{ij}^c/E$. Therefore, the scaling of ρ_+^- is determined by the largest term out of $N^{-1/2}$ and the scaling of $\sum_{i,j=1}^N E_{ij}^c/E$. Since the latter undergoes a phase transition, we actually have a three stage phase transition for the scaling of ρ_+^- in the ECM. The first stage has scaling $N^{-1+1/(\gamma_+ \wedge \gamma_-)}$ and holds for all γ_\pm for which

$$\frac{1}{\gamma_+ \wedge \gamma_-} - 1 \leq \frac{2}{\gamma_+} + \frac{2}{\gamma_-} - 3,$$

since both correspond to upper bounds. The next region, γ_\pm such that $2/\gamma_+ + 2/\gamma_- - 3 \geq -1/2$, has scaling $N^{2/\gamma_+ + 2/\gamma_- - 3}$. Outside this region we have normal scaling, $N^{-1/2}$. The different regions are displayed in Figure 6, while Table 3 shows the three scaling terms. We remark that the phase transitions of the scaling are smooth since they are induced by inequalities on the terms.

5.3 Simulations

In order to show the phase transitions we plotted the empirical cumulative distribution function of ρ_+^- for the specific choices of γ_\pm , corresponding to the points I, II and III in Figure 6. For each of the three points we shifted the empirical data by its average and multiplied it by $N^{-f(\gamma_+, \gamma_-)}$, for any of the three coefficients from Table 3, corresponding to the different scaling areas A, B and C. The results are shown in Figure 7. When the correct scaling is applied, the corresponding cdf plots should almost completely overlap and resemble the cdf of some limiting distribution. We

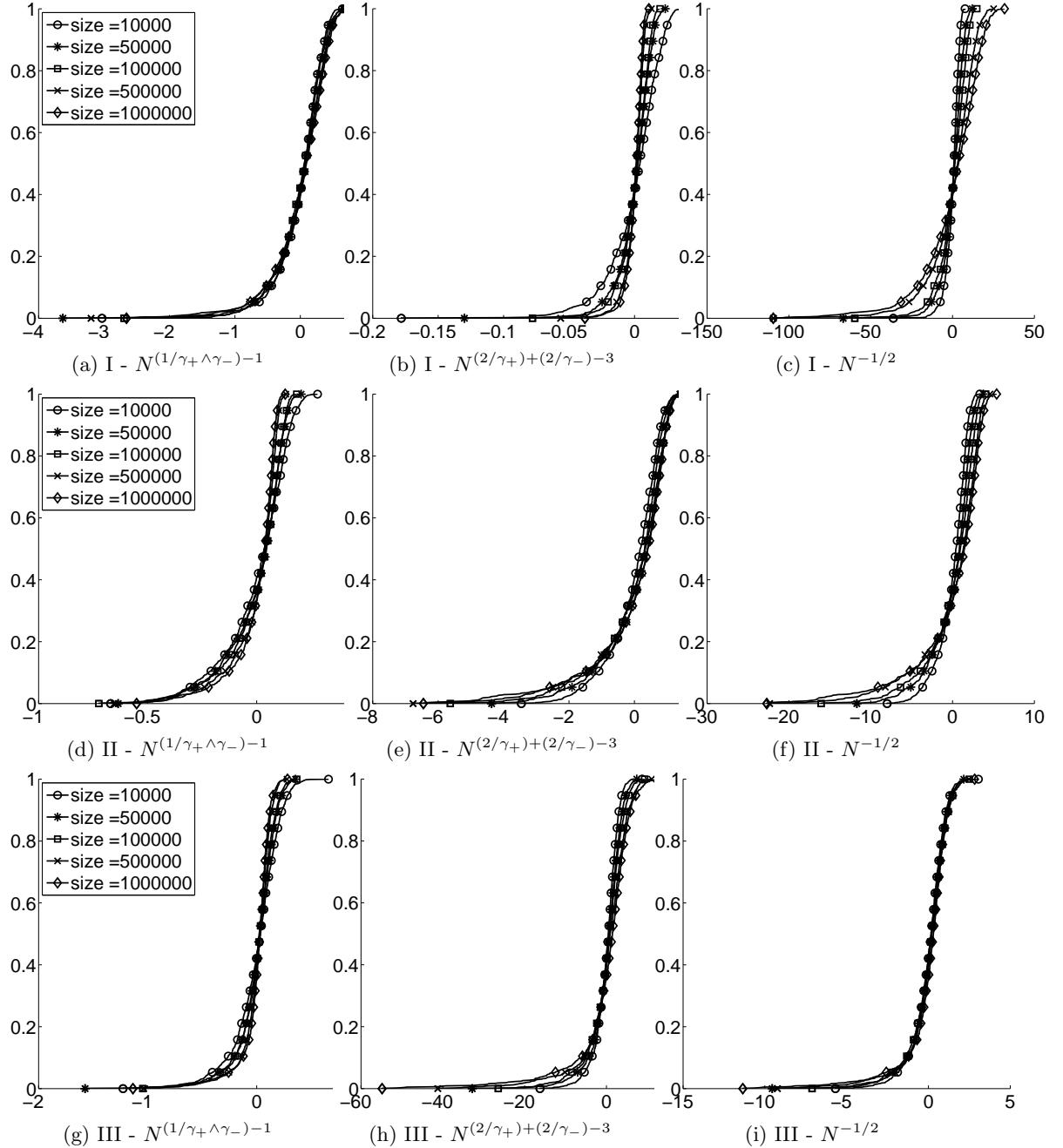


Figure 7: Plots of the empirical cumulative distribution function of ρ_+^- using different scaling and for different choices of γ_{\pm} . The left column is scaled by $N^{1/\min(\gamma_+, \gamma_-)}$, the center column by $N^{2/\gamma_+ + 2/\gamma_- - 3}$ and the right column by $N^{-1/2}$. The first row is for ECM graphs with $\gamma_{\pm} = 1.3$, the second for $\gamma_+ = 1.9, \gamma_- = 1.3$ and the third for $\gamma_+ = 1.9, \gamma_- = 1.5$, corresponding to points I, II and III, respectively in Figure 6.

observe that for each of the three choices I, II and III, this is the case when the corresponding scaling from its area, respectively A, B and C, is chosen.

6 Scaling of degree-degree dependencies for the other cases

In the previous section we completely characterized the scaling behavior of ρ_+^- for ECM graphs with infinite variance of the degrees. Here, we first discuss the remaining correlation types, ρ_+^+ , ρ_-^- and ρ_-^+ in the infinite variance regime and lastly, we consider all four types in the finite variance regime.

The intuition behind the structural negative Out-In dependencies was that multiple edges are more likely to exist between nodes of large out- and in-degree. The other three types do not show negative correlations, see Figure 5b-5d, which we argued was due to the fact that the in- and out-degree of a node in the ECM are independent. Nevertheless, the spread of both the Out-Out and In-In degree-degree dependency exhibits scaling with the same functions as the Out-In dependency. This is illustrated in Figure 8, where we plotted the empirical cumulative distribution of the Out-Out dependency for ECM graphs, for values of γ_\pm corresponding to points I, II and III from Figure 6, scaled by the correct term for each of these points. This is because ρ_+^+ again depends on the number of erased edges, through the out-degree of their target nodes. However, the out-degree of the source node of a removed edge can be both large or small, thus ρ_+^+ in the ECM remains zero on average. By symmetry, the scaling for the In-In dependency is similar.

This non-trivial scaling is typical for the ECM. Recall that in the CM, ρ_α^β is a normal statistic and scales as $N^{-1/2}$, for any, α, β because all degrees are independent random variables. This is exactly what we observe for the In-Out degree-degree dependency, which, in contrast to the other three, is not biased towards removed edges. As we expect, here we have normal, square root, scaling for ECM graphs for any choice of γ_\pm . This can clearly be observed in Figure 9, where we plotted the empirical cumulative distributions of ρ_-^+ scaled by $N^{-1/2}$.

For the degree-degree dependencies in the finite variance regime we plotted the empirical cumulative distributions of ρ_α^β , scaled by $N^{-1/2}$, in Figure 10. Since these are all completely similar, we took the plot for ρ_+^- for a ECM graph of size 10^6 and compared it to a fitted normal distribution with $\mu = 0$ and $\sigma^2 = 0.8$, see Figure 11. These plots strongly overlap, enforcing the claim that for ECM graphs with finite degree variance all four correlations are normal statistics.

7 Conclusion and Discussion

In this paper we analyzed degree-degree dependencies in the directed Erased Configuration Model. We showed, Figure 5, that in the infinite variance regime only the Out-In dependency exhibits structural negative values, while all correlations behave similar when both degrees have finite variance, Figure 4. We investigated the scaling of the structural negative Out-In correlations. These undergo a phase transition in terms of the exponents γ_\pm of the degree distributions (3), which we showed by establishing two upper bounds, (10) and (16), on the total, average, removed number of edges, both of which scale at different rates. Combining this with the square root scaling of Spearman's rho and Kendall's tau, we identified three regions, depending on γ_\pm , with different scaling, Figure 6, and illustrated their phase transitions in Figure 7. Next, we considered the remaining three dependency types for the infinite variance regime. We showed, Figure 8, that the scaling of the Out-Out and In-In correlations behaves similarly to the Out-In, even though they do not exhibit structural negative values, while the In-Out degree-degree dependency has square root scaling, Figure 9. Finally we investigated the scaling for correlations when the degrees have finite variance. In this case all four types have square root scaling and the plots of the cumulative distributions are very similar, Figure 10. This was confirmed when we compared the plot of ρ_+^- for ECM graphs of size 10^6 , with $\gamma_\pm = 2.1$, with that of a fitted normal distribution in Figure 11.

Our analysis shows that degree-degree dependencies in directed networks display non-trivial behavior in terms of scaling when the degrees have infinite variance. This scaling is important

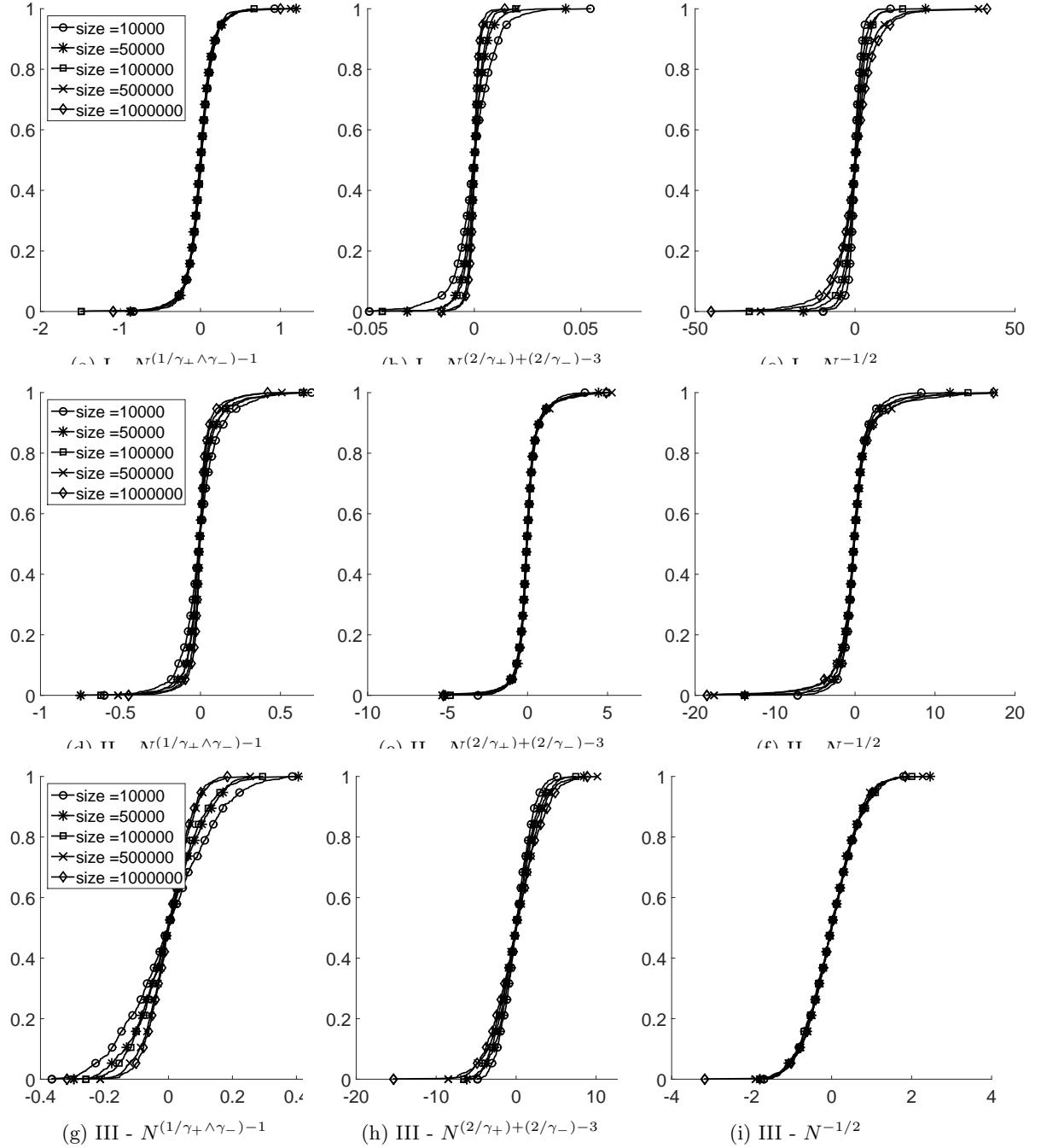


Figure 8: Plots of the empirical cumulative distribution function of ρ_+^+ for choices of γ_\pm corresponding to points I, II and III from Figure 6, using the corresponding scaling.

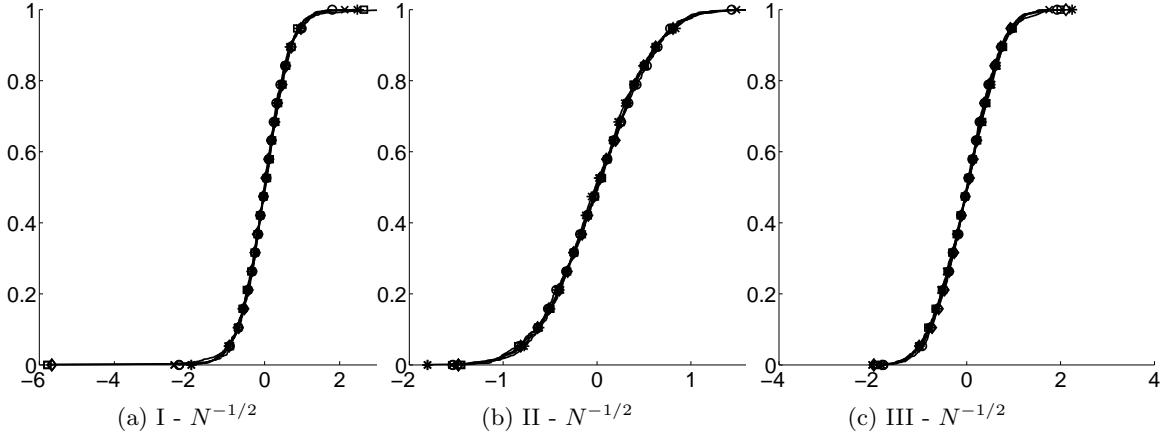


Figure 9: Plots of the empirical cumulative distribution function of ρ_+^+ for choices of γ_\pm corresponding to points I, II and III from Figure 6, using square root scaling.

when doing statistical analysis of these measures or their impact on other processes on networks, for it determines their spread and hence enables to asses the significance of measurements.

We showed that degree-degree dependencies for degrees with finite variance, scaled by $N^{-1/2}$, converge to a normal distribution with zero mean. We have not yet been able to determine the variance of these distributions as a function of the tail exponents γ_\pm which would completely characterize their behavior.

For three of the four correlation types in the infinite variance regime, we did not determine the limiting distributions. This is mainly due to the fact that we expect these to be *stable distributions*, since one of the three scaling regions is due to the Central Limit Theorem for stable random variables. Although these distributions have a well defined characteristic function, their density function, in general, does not have an analytical expression. Moreover, we are dealing with discrete data and simulation of such distributions is a field of it's own. Nevertheless, we do expect that Central Limit Theorems for degree-degree dependencies can be formulated and proven, which would fully complete their statistical analysis.

Finally, our empirical results clearly show the, analytically derived, phase transitions. However, the region with the $N^{(2/\gamma_+)+(2/\gamma_-)-3}$ scaling is less distinct than the other two. One of the possible reasons for this is that within the area where this scaling applies, the difference in value with the other two terms is small. We therefore picked point II in Figure 6 such that this difference was large enough to distinctly show this scaling visually in the plots.

We close by strongly suggesting to use the ECM as a null model for analysis of degree-degree dependencies, both for determining their impact on processes as well as significance. Although for the latter, values are often compared to averages, using the rewiring model [17], we emphasize that fixing the degrees imposes strong constraints on the possible simple graphs that can be generated. Moreover, in real-life networks, not only wiring but also the degrees of the nodes, are a result of a random process. Therefore, in a null-model, it seems more natural to fix only general properties of the network, such as degree distributions.

Acknowledgments:

All computations in this paper were done using the *fastutil* package and the WebGraph framework, [5], from the Laboratory for Web Algorithmics <http://law.di.unimi.it/software.php>. This work is supported by the EU-FET Open grant NADINE (288956).

References

- [1] Joseph Blitzstein and Persi Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Mathematics*, 6(4):489–522, 2011.

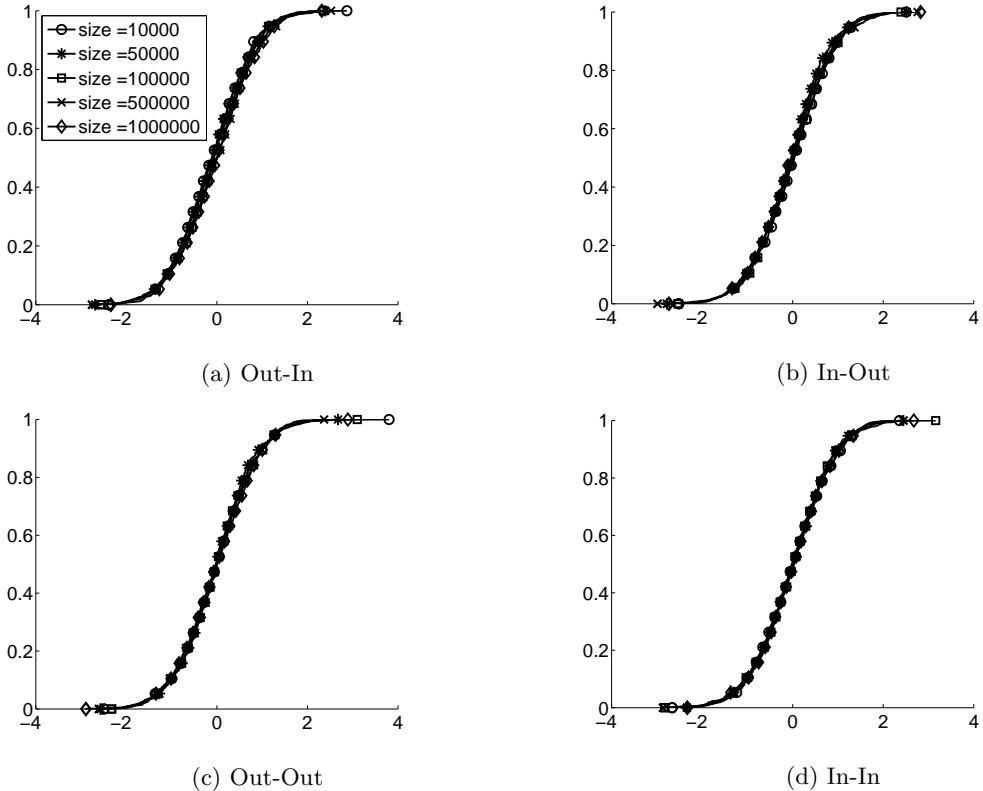


Figure 10: Plots of the empirical cumulative distribution of ρ_α^β for all four degree-degree dependency types for ECM graphs with $\gamma_\pm = 2.1$ of different sizes, scaled by $N^{-1/2}$. Each plot is based on 10^3 realizations of the model.

- [2] Marián Boguná and Romualdo Pastor-Satorras. Epidemic spreading in correlated complex networks. *Physical Review E*, 66(4):047104, 2002.
- [3] Marián Boguná, Romualdo Pastor-Satorras, and Alessandro Vespignani. Absence of epidemic threshold in scale-free networks with degree correlations. *Physical review letters*, 90(2):028701, 2003.
- [4] Marián Boguná, Romualdo Pastor-Satorras, and Alessandro Vespignani. Cut-offs and finite size effects in scale-free networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):205–209, 2004.
- [5] Paolo Boldi and Sebastiano Vigna. The webgraph framework i: compression techniques. In *Proceedings of the 13th international conference on World Wide Web*, pages 595–602. ACM, 2004.
- [6] Béla Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.
- [7] Michele Catanzaro, Marián Boguná, and Romualdo Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Physical Review E*, 71(2):027103, 2005.
- [8] Ningyuan Chen and Mariana Olvera-Cravioto. Directed random graphs with given degree distributions. *Stochastic Systems*, 3(1):147–186, 2013.
- [9] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

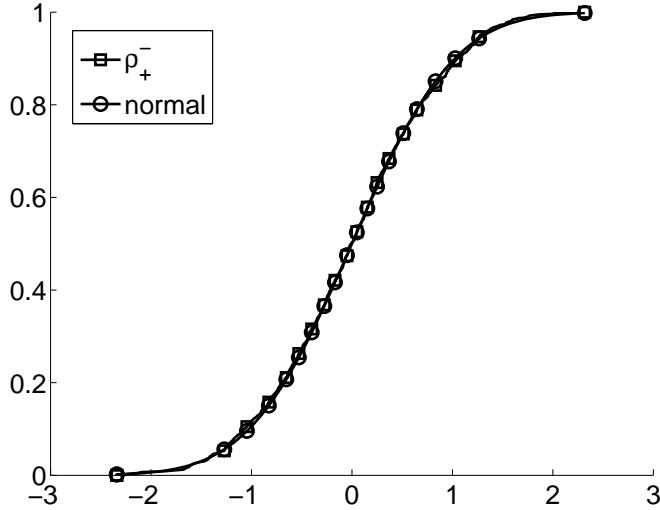


Figure 11: Plot of the empirical cumulative distribution function of ρ_+^- for ECM graphs of size 10^6 with $\gamma_\pm = 2.1$ and a normal cumulative distribution with $\mu = 0$ and $\sigma^2 = 0.8$.

- [10] Sergey N Dorogovtsev and Jose FF Mendes. Evolution of networks. *Advances in physics*, 51(4):1079–1187, 2002.
- [11] SN Dorogovtsev, AL Ferreira, AV Goltsev, and JFF Mendes. Zero pearson coefficient for strongly correlated growing trees. *Physical Review E*, 81(3):031135, 2010.
- [12] Jacob G Foster, David V Foster, Peter Grassberger, and Maya Paczuski. Edge direction and the structure of networks. *Proceedings of the National Academy of Sciences*, 107(24):10815–10820, 2010.
- [13] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The annals of mathematical statistics*, pages 293–325, 1948.
- [14] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
- [15] Nelly Litvak, Werner RW Scheinhardt, and Yana Volkovich. In-degree and pagerank: Why do they follow similar power laws? *Internet Mathematics*, 4(2-3):175–198, 2007.
- [16] Nelly Litvak and Remco van der Hofstad. Uncovering disassortativity in large scale-free networks. *Physical Review E*, 87(2):022801, 2013.
- [17] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.
- [18] Sergei Maslov, Kim Sneppen, and Alexei Zaliznyak. Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A: Statistical Mechanics and its Applications*, 333:529–540, 2004.
- [19] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995.
- [20] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.
- [21] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.

- [22] Juyong Park and Mark EJ Newman. Origin of degree correlations in the internet and other networks. *Physical Review E*, 68(2):026112, 2003.
- [23] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.
- [24] Tiziano Squartini and Diego Garlaschelli. Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics*, 13(8):083001, 2011.
- [25] Animesh Srivastava, Bivas Mitra, Fernando Peruani, and Niloy Ganguly. Attacks on correlated peer-to-peer networks: An analytical study. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 1076–1081. IEEE, 2011.
- [26] Remco van der Hofstad, Gerard Hooghiemstra, and Piet Van Mieghem. Distances in random graphs with finite variance degrees. *Random Structures & Algorithms*, 27(1):76–123, 2005.
- [27] Pim van der Hoorn and Nelly Litvak. Convergence of rank based degree-degree correlations in random directed networks. *arXiv preprint arXiv:1407.7662*, 2014.
- [28] Pim van der Hoorn and Nelly Litvak. Degree-degree dependencies in directed networks with heavy-tailed degrees. *Internet Mathematics*, (just-accepted):00–00, 2014.
- [29] Alexei Vázquez and Yamir Moreno. Resilience to damage of graphs with degree correlations. *Phys. Rev. E*, 67:015101, Jan 2003.
- [30] Ward Whitt. *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer, 2002.

Modelling of trends in Twitter using retweet graph dynamics

Marijn ten Thij¹, Tanneke Ouboter², Daniël Worm², Nelly Litvak^{3*}, Hans van den Berg^{2,3}, and Sandjai Bhulai¹

¹ VU University Amsterdam, Faculty of Sciences, the Netherlands,

{m.c.ten.thij,s.bhulai}@vu.nl,

² TNO, Delft, the Netherlands,

{tanneke.ouboter,daniel.worm,j.l.vandenberg}@tno.nl

³ University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science, the Netherlands

n.litvak@utwente.nl

Abstract. In this paper we model user behaviour in *Twitter* to capture the emergence of trending topics. For this purpose, we first extensively analyse tweet datasets of several different events. In particular, for these datasets, we construct and investigate the retweet graphs. We find that the retweet graph for a trending topic has a relatively dense largest connected component (LCC). Next, based on the insights obtained from the analyses of the datasets, we design a mathematical model that describes the evolution of a retweet graph by three main parameters. We then quantify, analytically and by simulation, the influence of the model parameters on the basic characteristics of the retweet graph, such as the density of edges and the size and density of the LCC. Finally, we put the model in practice, estimate its parameters and compare the resulting behavior of the model to our datasets.

Keywords: Retweet graph, Twitter, graph dynamics, random graph model

1 Introduction

Nowadays, social media play an important role in our society. The topics people discuss on-line are an image of what interests the community. Such trends may have various origins and consequences: from reaction to real-world events and naturally arising discussions to the trends manipulated e.g. by companies and organisations [14]. Trending topics on *Twitter* are ‘ongoing’ topics that become suddenly extremely popular⁴. In our study, we want to reveal differences in the retweet graph structure for different trends and model how these differences arise.

* The work of Nelly Litvak is partially supported the EU-FET Open grant NADINE (288956)

⁴ <https://support.twitter.com/articles/101125-about-trending-topics>

In *Twitter*⁵ users can post messages that consist of a maximum of 140 characters. These messages are called tweets. One can “follow” a user in *Twitter*, which places their messages in the message display, called the timeline. Social ties are directed in *Twitter*, thus if user A follows user B, it does not imply that B follows A. People that “follow” a user are called “friends” of this user. We refer to the network of social ties in *Twitter* as the friend-follower network. Further, one can forward a tweet of a user, which is called a retweet.

There have been many studies on detecting different types of trends, for instance detecting emergencies [9], earthquakes [18], diseases [13] or important events in sports [11]. In many current studies into trend behaviour, the focus is mainly on content of the messages that are part of the trend, see e.g. [12]. Our work focuses instead on the underlying networks describing the social ties between users of *Twitter*. Specifically, we consider a graph of users, where an edge means that one of the users has retweeted a message of a different user.

In this study we use several datasets of tweets on multiple topics. First we analyse the datasets, described in Section 3, by constructing the retweet graphs and obtaining their properties as discussed in Section 4. Next, we design a mathematical model, presented in Section 5, that describes the growth of the retweet graph. The model involves two attachment mechanisms. The first mechanism is the preferential attachment mechanism that causes more popular messages to be retweeted with a higher probability. The second mechanism is the superstar mechanism which ensures that a user that starts a new discussion receives a finite fraction of all retweets in that discussion [2]. We quantify, analytically and with simulations, the influence of the model parameters on its basic characteristics, such as the density of edges, the size and the density of the largest connected component. In Section 6 we put the model in practice, estimate its parameters and compare it to our datasets. We find that what our model captures, is promising for describing the retweet graphs of trending topics. We close with conclusions and discussion in Section 7.

2 Related work

The amount of literature regarding trend detection in *Twitter* is vast. The overview we provide here is by no means complete. Many studies have been performed to determine basic properties of the so-called “Twitterverse”. Kwak et al. [10] analysed the follower distribution and found a non-power-law distribution with a short effective diameter and a low reciprocity. Furthermore they found that ranking by the number of followers and PageRank both induce similar rankings. They also report that *Twitter* is mainly used for News (85% of the content). Huberman et al. [8] found that the network of interactions within *Twitter* is not equal to the follower network, it is a lot smaller.

An important part of trending behaviour in social media is the way these trends progress through the network. Many studies have been performed on

⁵ www.twitter.com

Twitter data. For instance, [3] studies the diffusion of news items in *Twitter* for several well-known news media and finds that these cascades follow a star-like structure. Also, [20] investigates the diffusion of information on *Twitter* using tweets on the Iranian election in 2009, and finds that cascades tend to be wide, not too deep and follow a power law-distribution in their size.

Bhamidi et al. [2] proposed and validated on the data a so-called superstar random graph model for a giant component of a retweet graph. Their model is based on the well-known preferential attachment idea, where users with many retweets have a higher chance to be retweeted [1], however, there is also a superstar node that receives a new retweet at each step with a positive probability. We build on this idea to develop our model for the progression of a trend through the *Twitter* network.

Another perspective on the diffusion of information in social media is obtained through analysing content of messages. For example, [17] finds that on *Twitter*, tags tend to travel to more distant parts of the network and URLs travel shorter distances. Romero et al. [16] analyse the spread mechanics of content through hashtag use and derive probabilities that users adopt a hashtag.

Classification of trends on *Twitter* has attracted considerable attention in the literature. Zubia et al. [21] derive four different types of trends, using 15 features to make their distinction. They distinguish trends triggered by news, current events, memes or commemorative tweets. Lehmann et al. [12] study different patterns of hashtag trends in *Twitter*. They also observe four different classes of hashtag trends. Rattanaritnont et al. [15] propose to distinguish topics based on four factors, which are cascade ratio, tweet ratio, time of tweet and patterns in topic-sensitive hashtags.

We extend the model of [2] by mathematically describing the growth of a complete retweet graph. Our proposed model has two more parameters that define the shape of the resulting graph, in particular, the size and the density of its largest connected component. To the best of our knowledge, this is the first attempt to classify trends using a random graph model rather than algorithmic techniques or machine learning. The advantage of this approach is that it gives insight in emergence of the trend, which, in turn, is important for understanding and predicting the potential impact of social media on real world events.

3 Datasets

We use datasets containing tweets that have been acquired either using the *Twitter* Streaming API⁶ or the *Twitter* REST API⁷. Using the REST API one can obtain tweets or users from *Twitter*'s databases. The Streaming API filters tweets that *Twitter* parses during a day, for example, based on users, locations, hashtags, or keywords.

Most of the datasets used in this study were scraped by RTreporter, a company that uses an incoming stream of Dutch tweets to detect news for news

⁶ <https://dev.twitter.com/docs/streaming-apis>

⁷ <https://dev.twitter.com/docs/api/1.1>

agencies in the Netherlands. These tweets are scraped based on keywords, using the Streaming API. For this research, we selected several events that happened in the period of data collection, based on the wikipedia overviews of 2013 and 2014⁸. We have also used two datasets scraped by TNO - Netherlands Organisation for Applied Scientific Research. The *Project X* dataset contains tweets related to large riots in Haren, the Netherlands. This dataset is acquired by *Twitincident*⁹. For this study, we have filtered this dataset on two most important hashtags: `#projectx` and `#projectxharen`. The *Turkish-Kurdish* dataset is described in more detail in Bouma et al. [4]. A complete overview of the datasets, including the events and the keywords, is given in Table 1. The size and the timespans for each dataset are given in Table 2.

	dataset	keywords
PX	Project X Haren	projectx, projectxharen
TK	Demonstrations in Amsterdam related to the Turkish-Kurdish conflict	koerden, turken, rellen, museumplein, amsterdam
WCS	World cup speedskating single distanced 2013	wkafstanden, sochi, sotsji
W-A	Crowning of His Majesty King Willem-Alexander in the Netherlands	troonwisseling, troon, Willem-Alexander, Wim-Lex, Beatrix, koning, koningin
ESF	Eurovision Song Festival	esf, Eurovisie Songfestival, ESF, songfestival, eurovisie
CL	Champions League final 2013	Bayern Munchen, Borussia Dortmund, dorbay, borussia, bayern, borbay, CL
Morsi	Morsi deposited as Egypt president	Morsi, afgezet, Egypte
Train	Train crash in Santiago, Spain	Treincrash, treincrash, Santiago, Spanje, Santiago de Compostella, trein
Heat	Heat wave in the Netherlands	hittegolf, Nederland
Damascus	Sarin attack in Damascus	Sarin, Damascus, Syrië, syrië
Peshawar	Bombing in Peshawar	Peshawar, kerk, zelfmoordaanslag, Pakistan
Hawk	Hawk spotted in the Netherlands	sperweruil, Zwolle
Pile-up	Multiple pile-ups in Belgium on the A19	A19, Ieper, Kortrijk, kettingbotsing
Schumi	Michael Schumacher has a skiing accident	Michael Schumacher, ski-ongeval
UKR	Rebellion in Ukraine	Azarov, Euromaidan, Euromajdan, Oekraïne, opstand
NAM	Treaty between NAM and Dutch government	Loppersum, gasakkoord, NAM, Groningen
WCD	Michael van Gerwen wins PDC WC Darts	van Gerwen, PDC, WK Darts
NSS	Nuclear Security Summit 2014	NSS2014, NSS, Nuclear Security Summit 2014, Den Haag
MH730	Flight MH730 disappears	MH730, Malaysia Airlines
Crimea	Crimea referendum for independence	Krim, referendum, onafhankelijkheid
Kingsday	First Kingsday in the Netherlands	koningendag, kingsday, koningsdag
Volkert	Volkert van der Graaf released from prison	Volkert, volkertvandergraaf, Volkert van der Graaf

Table 1. Datasets: events and keywords (some keywords are in Dutch).

For each dataset we have observed there is at least one large peak in the progression of the number of tweets. For example, Figure 1 shows such peak in *Twitter* activity for the *Project X* dataset.

⁸ <http://nl.wikipedia.org/wiki/2014> & <http://nl.wikipedia.org/wiki/2013>

⁹ www.twitincident.com

dataset	year	first tweet	last tweet	# tweets	# retweets
PX	2012	Sep 17 09:37:18	Sep 26 02:31:15	31,144	15,357
TK	2011	Oct 19 14:03:23	Oct 27 08:42:18	6,099	999
WCS	2013	Mar 21 09:19:06	Mar 25 08:45:50	2,182	311
W-A	2013	Apr 27 22:59:59	May 02 22:59:25	352,157	88,594
ESF	2013	May 13 23:00:08	May 18 22:59:59	318,652	82,968
CL	2013	May 22 23:00:04	May 26 22:59:54	163,612	54,471
Morsi	2013	Jun 30 23:00:00	Jul 04 22:59:23	40,737	13,098
Train	2013	Jul 23 23:00:02	Jul 30 22:59:41	113,375	26,534
Heat	2013	Jul 10 19:44:35	Jul 29 22:59:58	173,286	42,835
Damascus	2013	Aug 20 23:01:57	Aug 31 22:59:54	39,377	11,492
Peshawar	2013	Sep 21 23:00:00	Sep 24 22:59:59	18,242	5,323
Hawk	2013	Nov 11 23:00:07	Nov 30 22:58:59	54,970	19,817
Pile-up	2013	Dec 02 23:00:15	Dec 04 22:59:57	6,157	2,254
Schumi	2013-14	Dec 29 02:43:16	Jan 01 22:54:50	13,011	5,661
UKR	2014	Jan 26 23:00:36	Jan 31 22:57:12	4,249	1,724
NAM	2014	Jan 16 23:00:22	Jan 20 22:59:49	41,486	14,699
WCD	2013-14	Dec 31 23:03:48	Jan 02 22:59:05	15,268	5,900
NSS	2014	Mar 23 23:00:06	Mar 24 22:59:56	29,175	13,042
MH730	2014	Mar 08 00:18:32	Mar 28 22:40:44	36,765	17,940
Crimea	2014	Mar 13 23:02:22	Mar 17 22:59:57	18,750	5,881
Kingsday	2014	Apr 26 23:00:00	Apr 29 22:53:00	7,576	2,144
Volkert	2014	Apr 30 23:08:14	May 04 22:57:06	9,659	4,214

Table 2. Characteristics of the datasets.

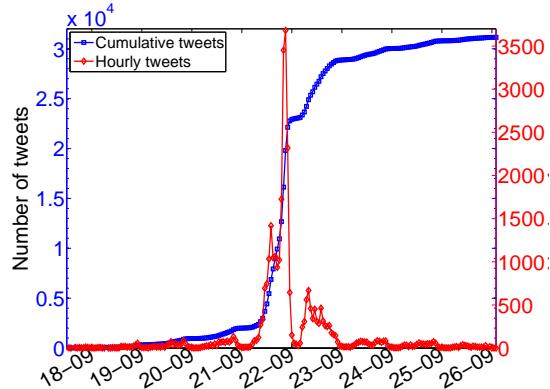


Fig. 1. Project X Number of tweets and cumulative number of tweets per hour.

When a retweet is placed on *Twitter*, the Streaming API returns the retweet together with the message that has been retweeted. We use this information to construct the retweet trees of every message and the user IDs for each posted message. The tweet and graph analysis is done using *Python* and its modules *Tweepy*¹⁰ and *NetworkX*¹¹. In this paper, we investigate the dynamics of retweet graphs with the goal to predict peaks in *Twitter* activity and classify the nature of trends.

¹⁰ <http://www.tweepy.org/>

¹¹ <http://networkx.github.io/>

4 Retweet graphs

Our main object of study is the retweet graph $G = (V, E)$, which is a graph of users that have participated in the discussion on a specific topic. A directed edge $e = (u, v)$ indicates that user v has retweeted a tweet of u . We observe the retweet graph at the time instances $t = 0, 1, 2, \dots$, where either a new node or a new edge was added to the graph, and we denote by $G_t = (V_t, E_t)$ the retweet graph at time t . As usual, the out- (in-) degree of node u is the number of directed edges with source (destination) in u . In what follows, we model and analyse the properties of G_t . For every new message initiated by a new user u a tree T_u is formed. Then, \mathcal{T}_t denotes the forest of message trees. Note that in our model a new message from an already existing user u (that is, $u \in \mathcal{T}_t$) does not initiate a new message tree. We define $|\mathcal{T}_t|$ as the number of new users that have started a message tree up to time t .

After analyzing multiple characteristics of the retweet graphs for every hour of their progression, we found that the size of the largest (weakly) connected component (LCC) and its density are the most informative characteristics for predicting the peak in *Twitter*. In Figure 2 we show the development of these characteristics in the *Project X* dataset. One day before the actual event, we observe a very interesting phenomenon in the development of the edge density of the LCC in Figure 2a. Namely, at some point the edge density of the LCC exceeds 1 (indicated by the dash-dotted gray lines), i.e. there is more than one retweet per user on average. We shall refer to this as the *densification* (or *dens.*) of the LCC. Furthermore, the relative size of the LCC increases from 18% to 25% as well, see Figure 2b.

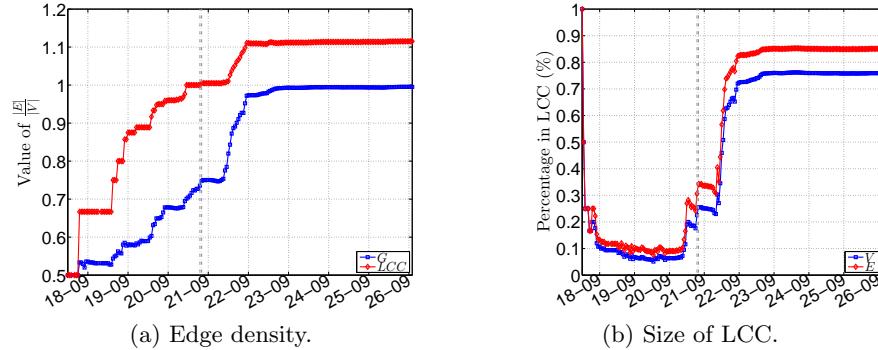


Fig. 2. Progression for the edge density (a) and the size of the LCC (b) in the *Project X* dataset.

We have observed a densification of the LCC in each dataset that we have studied. Indeed, when the LCC grows its density must become at least one (each node is added to the LCC together with at least one edge). However, we have also observed that in each dataset the densification occurs before the main peak, but

the scale of densification is different. For example, in the *Project X* dataset the densification already occurs one day before the peak activity. Plausibly, in this discussion, that ended up in riots, a group of people was actively participating before the event. On the other hand, in the *WCS* dataset, which tweets about an ongoing sport event, the densification of the LCC occurs during the largest peak. This is the third peak in the progression. Hence, our experiments suggest that the time of densification has predictive value for trend progression and classification. See Table 5 for the density of the LCC in each dataset at the end of the progression.

5 Model

Our goal is to design a model that captures the development of trending behaviour. In particular, we need to capture the phenomenon that disjoint components of the retweet graph join together forming the largest component, of which the density of edges may become larger than one. To this end, we employ the superstar model of Bhamidi et al. [2] for modelling distinct components of the retweet graph, and add the mechanism for new components to arrive and the existing components to merge. For the sake of simplicity of the model we neglect the friend-follower network of *Twitter*. Note that in *Twitter* every user can retweet any message sent by any public user, which supports our simplification.

At the start of the progression, we have the graph G_0 . In the analysis of this section, we assume that G_0 consists of a single node. Note that in reality, this does not need to be the case: any directed graph can be used as an input graph G_0 . In fact, in Section 6 we start with the actual retweet graph at a given point in time, and then use the model to build the graph further to its final size.

We consider the evolution of the retweet graph in time $(G_t)_{t \geq 0}$. We use a subscript t to indicate G_t and related notions at time t . We omit the index t when referring to the graph at the end of the progression.

Recall that G_t is a graph of *users*, and an edge (u, v) means that v has retweeted a tweet of u . We consider time instances $t = 1, 2, \dots$ when either a new node or a new edge is added to the graph G_{t-1} . We distinguish three types of changes in the retweet graph:

- $T1$: a new user u has posted a new message on the topic, node u is added to G_{t-1} ;
- $T2$: a new user v has retweeted an existing user u , node v and edge (u, v) are added to G_{t-1} ;
- $T3$: an existing user v has retweeted another existing user u , edge (u, v) is added to G_{t-1} .

The initial node is equivalent to a $T1$ arrival at time $t = 0$. Assume that each change in G_t at $t = 1, 2, \dots$ is $T1$ with probability $\lambda/(1 + \lambda)$, independently of the past. Also, assume that a new edge (retweet) is coming from a new user with probability p . Then the probabilities of $T1$, $T2$ and $T3$ arrivals are, respectively

$\frac{\lambda}{\lambda+1}, \frac{p}{\lambda+1}, \frac{1-p}{\lambda+1}$. The parameter p is governing the process of components merging together, while λ is governing the arrival of new components in the graph.

For both $T2$ and $T3$ arrivals we define the same mechanism for choosing the source of the new edge (u, v) as follows.

Let u_0, u_1, \dots be the users that have been added to the graph as $T1$ arrivals, where u_0 is the initial node. Denote by $T_{i,t}$ the subgraph of G_t that includes u_i and all users that have retweeted the message of u_i in the interval $(0, t]$. We call such a subgraph a message tree with root u_i . We assume that the probability that a $T2$ or $T3$ arrival at time t will attach an edge to one of the nodes in $T_{i,t-1}$ with probability $p_{T_{i,t-1}}$, proportional to the size of the message tree:

$$p_{T_{i,t-1}} = \frac{|T_{i,t-1}|}{\sum_{T_{j,t-1} \subset T_{t-1}} |T_{j,t-1}|}.$$

This creates a preferential attachment mechanism in the formation of the message trees. Next, a node in the selected message tree $T_{i,t-1}$ is chosen as the source node following the superstar attachment scheme [2]: with probability q , the new retweet is attached to u_i , and with probability $1-q$, the new retweet is attached to any other vertex, proportional to the preferential attachment function of the node, that we choose to be the number of children of the node plus one.

Thus we employ the superstar-model, which was suggested in [2] for modelling the largest connected component of the retweet graph on a given topic, in order to describe a progression mechanism for a single retweet tree. Our extensions compared to [2] are that we allow new message trees to appear ($T1$ arrivals), and that different message trees may either remain disconnected or get connected by a $T3$ arrival.

For a $T3$ arrival, the target of the new edge (u, v) is chosen uniformly at random from V_{t-1} , with the exception of the earlier chosen source node u , to prevent self-loops. That is, any user is equally likely to retweet a message from another existing user.

Note that, in our setting, it is easy to introduce a different superstar parameter q_{T_i} for every message tree T_i . This way one could easily implement specific properties of the user that starts the message tree, e.g. his/her number of followers. For the sake of simplicity, we choose the same value of q for all message trees. Also note that we do not include tweets and retweets that do not result in new nodes or edges in a retweet graph. This could be done, for example, by introducing dynamic weights of vertices and edges, that increase with new tweets and retweets. Here we consider only an unweighted model.

5.1 Growth of the graph

The average degree, or edge density, is one of the aspects through which we give insight to the growth of the graph. The essential properties of this characteristic are presented in Theorem 1. The proof is given in the Appendix.

Theorem 1 Let τ_n be the time when node n is added to the graph. Then

$$\mathbb{E} \left[\frac{|E_{\tau_n}|}{|V_{\tau_n}|} \right] = \frac{1}{\lambda + p} - \frac{1}{n(\lambda + p)}, \quad (1)$$

$$\text{var} \left(\frac{|E_{\tau_n}|}{|V_{\tau_n}|} \right) = \frac{(n-1)(\lambda + 1 - p)}{n^2(\lambda + p)^2}. \quad (2)$$

Note that the variance of the average degree in (2) converges to zero as $n \rightarrow \infty$ at rate $\frac{1}{n}$.

The next theorem studies the observed ratio between $T2$ and $T3$ arrivals (new edges) and $T1$ arrivals (new nodes with a new message). As we see from the theorem, this ratio can be used for estimating the parameter λ . The proof is given in the Appendix.

Theorem 2 Let $G_t = (V_t, E_t)$ be the retweet graph at time t , let \mathcal{T}_t be the set of all message trees in G_t . Then

$$\mathbb{E} \left[\frac{|E_t|}{|\mathcal{T}_t|} \right] = \lambda^{-1} \cdot \left(1 - \left(\frac{1}{\lambda + 1} \right)^t \right), \quad (3)$$

$$\lim_{t \rightarrow \infty} \frac{\lambda^3 t}{(1 + \lambda)^2} \text{var} \left(\frac{|E_t|}{|\mathcal{T}_t|} \right) = 1, \quad (4)$$

Furthermore,

$$\frac{\lambda^{3/2} \sqrt{t}}{\lambda + 1} \left(\frac{|E_t|}{|\mathcal{T}_t|} - \frac{1}{\lambda} \right) \xrightarrow{D} Z, \quad (5)$$

where Z is a standard normal $N(0, 1)$ random variable, and \xrightarrow{D} denotes convergence in distribution.

Note that, as expected from the definition of λ ,

$$\lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{|E_t|}{|\mathcal{T}_t|} \right] = \lambda^{-1}. \quad (6)$$

This will be used in Section 6 for estimating λ .

5.2 Component size distribution

In the following, we assume that G_t consists of m connected components (C_1, C_2, \dots, C_m) with known respective sizes $(|C_1|, \dots, |C_m|)$. We aim to derive expressions for the distribution of the component sizes in G_{t+1} .

Lemma 3 The distribution of the sizes of the components of G_{t+1} , given G_t is as follows,

$$\begin{cases} |C_1|, \dots, |C_i|, |C_j|, \dots, |C_m|, 1 & \text{w.p. } \frac{\lambda}{\lambda + 1} \\ |C_1|, \dots, |C_i| + 1, |C_j|, \dots, |C_m| & \text{w.p. } \frac{p}{\lambda + 1} \cdot \frac{|C_i|}{|V|} \\ |C_1|, \dots, |C_i| + |C_j|, \dots, |C_m| & \text{w.p. } \frac{1-p}{\lambda + 1} \cdot \frac{2 \cdot |C_i| \cdot |C_j|}{|V|^2 - |V|} \\ |C_1|, \dots, |C_i|, |C_j|, \dots, |C_m| & \text{w.p. } \frac{1-p}{\lambda + 1} \cdot \frac{\sum_{k=1}^m |C_k|^2 - |C_i|}{|V|^2 - |V|} \end{cases} \quad (7)$$

The proof of Lemma 3 is given in the Appendix. Lemma 3 provides a recursion for computing the distribution of component sizes. However, the computations are highly demanding if not infeasible. Also, deriving an exact expression of the distribution of the component sizes at time t is very cumbersome because they are hard and they strongly depend on the events that occurred at $t = 0, \dots, t-1$. Note that if $p = 1$, there is a direct correspondence between our model and the infinite generalized Pólya process [5]. However, this case is uninformative as there are no $T3$ arrivals. Therefore, in the next section we resort to simulations to investigate the sensitivity of the graph characteristics to the model parameters.

5.3 Influence of q , p and λ

We analyze the influence of the model parameters λ , p and q on the characteristics of the resulting graph numerically using simulations. To this end, we fix two out of three parameters and execute multiple simulation runs of the model, varying the values for the third parameter. We start simulations with graph G_0 , consisting of one node. We perform 50 simulation runs for every parameter setting and obtain the average values over the individual runs for given parameters.

Parameter q affects the degree distribution [2] and the overall structure of the graph. If $q = 0$, then the graph contains less nodes that have many retweets. If $q = 1$ each edge is connected to a superstar, and the graph consists of star-like sub graphs, some of which are connected to each other. In the *Project X* dataset, which is our main case study, $q \approx 0.9$ results in a degree distribution that closely approximates the data. Since degree distributions are not in the scope of this paper, we omit these results for brevity.

We compare the results for two measures that produced especially important characteristics of the *Project X* dataset: $\frac{|E_{LCC}|}{|V_{LCC}|}$ and $\frac{|V_{LCC}|}{|V|}$. These characteristics do not depend on q . In simulations, we set $t = 1,000$, $q = 0.9$ and vary the values for p and λ . The results are given in Figure 3.

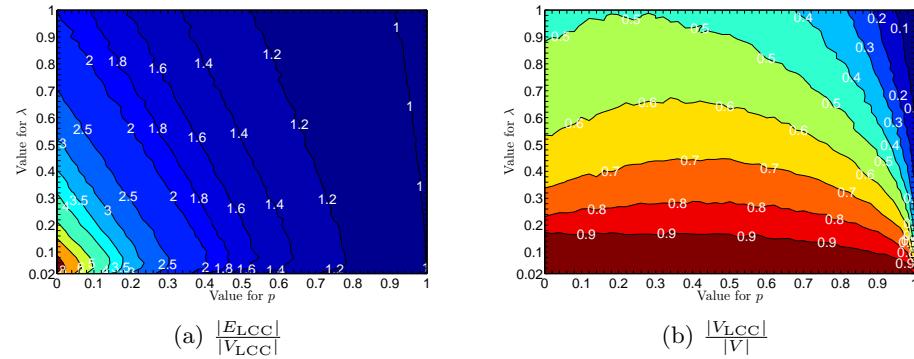


Fig. 3. Numerical results for the model using $q = 0.9$ and $t = 1,000$.

We see that the edge density in the LCC in Figure 3a decreases with λ and p . Note that according to (1), $|E|/|V|$ is well approximated by $1/(\lambda + p)$ when λ or p are large enough. The edge density in LCC shows a similar pattern, but it is slightly higher than in the whole graph. When λ and p are small, there are many $T3$ arrivals, and new nodes are not added frequently enough. This results in an unexpected non-monotonic behaviour of the edge density near the origin. For the fraction of nodes in the LCC, depicted in Figure 3b, we see that the parameter λ is most influential. The parameter p is of considerable influence only when it is large.

6 The model in practice

In this section we obtain parameter estimators for our model and compare the model to the datasets discussed in Section 3.

Using Theorem 2, we know that $\frac{|E_t|}{|\mathcal{T}_t|}$ converges to λ^{-1} as $t \rightarrow \infty$. Thus, we suggest the following estimator for λ at time $t > 0$:

$$\hat{\lambda}_t = \frac{|\mathcal{T}_t|}{|E_t|}. \quad (8)$$

Second, we derive an expression for \hat{p}_t using (1) and substituting (??) for λ :

$$\hat{p}_t = \frac{|V_t| - |\mathcal{T}_t| - 1}{|E_t|}. \quad (9)$$

Since the *Twitter* API only gives back the original message of a retweet and not the level in the progression tree of that retweet, we can not determine q easily from the data. Since this parameter does not have a large influence on the outcomes of the simulations, we choose this parameter to be 0.9 for all datasets.

Notice that we can obtain the numbers ($|E_t|$, $|\mathcal{T}_t|$ and $|V_t|$) directly from a given retweet graph for each $t = 1, 2, \dots$. The computed estimators for our datasets are displayed in Table 5.

Next, we compare 50 simulations of the datasets from the point of densification of the LCC until the graph has reached the same size as the actual dataset. We display the average outcomes of these simulations and compare them to the actual properties of the retweet graphs of each dataset in Table 5.

Here we see diverse results per dataset in the simulations. For the *CL*, *Morsi* and *WCD* datasets, the simulations are very similar to the actual progressions. However, for some datasets, for instance the *ESF* dataset, simulations are far off. In general, the model predicts the density of the LCC quite well for many datasets, but tends to overestimate the size of the LCC. We notice that current random graph models for networks usually capture one or two essential features, such as degree distribution, self-similarity, clustering coefficient or diameter. Our model captures both degree distribution and, in many cases, the density of the LCC. It seems that our model performs better on the datasets that have a singular peak rather than a series of peaks. We have observed on the data that

dataset	$\hat{\lambda}$	\hat{p}	actual progression			simulations (starting at dens.)		
			$ V_{LCC} $ $ V $	$ E $ $ V $	$ E_{LCC} $ $ V_{LCC} $	$ V_{LCC} $ $ V $	$ E $ $ V $	$ E_{LCC} $ $ V_{LCC} $
PX	.23	.78	.76	1.00	1.12	.54	.75	1.08
TK	.42	.85	.25	.79	1.00	.54	.74	1.08
WCS	.49	.73	.20	.81	.99	.49	.95	1.90
W-A	.41	.52	.67	1.07	1.30	.40	.62	1.41
ESF	.38	.43	.73	1.24	1.48	.45	.69	1.42
CL	.40	.72	.44	.90	1.22	.46	.66	1.16
Morsi	.60	.55	.39	.87	1.20	.47	.67	1.17
Train	.54	.78	.28	.76	1.04	.50	.70	1.17
Heat	.42	.59	.60	.99	1.23	.41	.72	1.68
Damascus	.58	.51	.46	.92	1.24	.44	.65	1.30
Peshawar	.54	.68	.31	.82	1.18	.53	.75	1.25
Hawk	.38	.38	.82	1.31	1.45	.49	.76	1.43
Pile-up	.33	.64	.65	1.03	1.24	.58	.93	1.54
Schumi	.38	.83	.33	.82	1.08	.56	.77	1.07
UKR	.72	.37	.53	.91	1.12	.50	.75	1.38
NAM	.44	.48	.50	1.09	1.51	.45	.72	1.51
WCD	.26	.81	.66	.94	1.10	.64	.83	1.07
NSS	.26	.62	.79	1.13	1.26	.23	.35	1.21
MH730	.33	.52	.15	1.18	1.00	.56	.76	1.09
Crimea	.44	.63	.51	.93	1.19	.52	.72	1.12
Kingsday	.47	.92	.07	.72	1.11	.47	.67	1.15
Volkert	.29	.55	.79	1.18	1.31	.64	.87	1.22

Table 5. Estimated parameter values using complete dataset, simulation and progression properties.

each peak activity has a large impact on the parameters estimation. We will strive to adopt the model for incorporating different rules for activity during peaks, and improving results on the size of the LCC.

7 Conclusion and Discussion

We have found that our model performs well in modelling the retweet graph for tweets regarding a singular topic. However, there is a room for improvement when the dataset covers a prolonged discussion with users activity fluctuating over time.

A possible extension of the present work is incorporating more explicitly the time aspect into our model. We could for example add the notion of ‘novelty’, like Gómez et al. in [6], taking into account that e.g. the retweet probability for a user may decrease the longer he/she remains silent after having received a tweet. But also other model parameters may be assumed to vary over time. In addition, we propose to analyse the clustering coefficient of a node in the network model and, in particular, to investigate how it evolves over time. This measure (see [19]) provides more detailed insight in how the graph becomes denser, making it possible to distinguish between local and global density.

References

1. A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

2. S. Bhamidi, J. M. Steele, and T. Zaman. Twitter event networks and the superstar model. *arXiv preprint arXiv:1211.3090*, 2012.
3. D. Bhattacharya and S. Ram. Sharing news articles using 140 characters: A diffusion analysis on Twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 966–971. IEEE, 2012.
4. H. Bouma, O. Rajadell, D. Worm, C. Versloot, and H. Wedemeijer. On the early detection of threats in the real world based on open-source information on the Internet. In *International Conference on Information Technologies and Security (ITSEC)*, 2012.
5. F. Chung, S. Handjani, and D. Jungreis. Generalizations of Polya’s urn problem. *Annals of combinatorics*, 7(2):141–153, 2003.
6. V. Gómez, H. J. Kappen, N. Litvak, and A. Kaltenbrunner. A likelihood-based framework for the analysis of discussion threads. *World Wide Web*, pages 1–31, 2012.
7. P. Hall. On the rate of convergence of moments in the central limit theorem for lattice distributions. *Transactions of the American Mathematical Society*, 278(1):169–181, 1983.
8. B. Huberman, D. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *Available at SSRN 1313405*, 2008.
9. B. Klein, X. Laiseca, D. Casado-Mansilla, D. López-de Ipiña, and A. Nespral. Detection and extracting of emergency knowledge from Twitter streams. *Ubiquitous Computing and Ambient Intelligence*, pages 462–469, 2012.
10. H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.
11. J. Lanagan and A. F. Smeaton. Using Twitter to detect and tag important events in live sports. pages 542–545. AAAI, 2011.
12. J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in Twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 251–260. ACM, 2012.
13. M. J. Paul and M. Dredze. You are what you tweet: Analyzing Twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, 2011.
14. J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252. ACM, 2011.
15. G. Rattanaritnont, M. Toyoda, and M. Kitsuregawa. A study on characteristics of topicspecific information cascade in Twitter. In *Forum on Data Engineering (DE2011)*, pages 65–70, 2011.
16. D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.
17. E. Sadikov and M. M. M. Martinez. Information propagation on Twitter. *CS322 Project Report*, 2009.
18. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

19. D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.
20. Z. Zhou, R. Bandari, J. Kong, H. Qian, and V. Roychowdhury. Information resonance on Twitter: watching iran. In *Proceedings of the First Workshop on Social Media Analytics*, pages 123–131. ACM, 2010.
21. A. Zubiaga, D. Spina, V. Fresno, and R. Martínez. Classifying trending topics: a typology of conversation triggers on Twitter. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2461–2464. ACM, 2011.

Appendix

A1. Proof of Theorem 1

Proof. The proof is based on the fact that the total number of edges $|E_{\tau_n}|$ equals a total number of the $T2$ and $T3$ arrivals on $(0, \tau_n]$. By definition, $(0, \tau_n]$ contains exactly $(n-1)$ of $T1$ or $T2$ arrivals, hence, the number of $T2$ arrivals has a Binomial distribution with number of trials equal to $(n-1)$, and success probability $P(T2 \mid T1 \text{ or } T2) = \frac{p}{\lambda+p}$. Next, the number of $T3$ arrivals on $[\tau_i, \tau_{i+1})$, where $i = 1, \dots, n-1$, has a shifted geometric distribution, namely, the probability of k $T3$ arrivals on $[\tau_i, \tau_{i+1})$ is

$$\left(1 - \frac{1-p}{\lambda+1}\right) \left(\frac{1-p}{\lambda+1}\right)^k, \quad k = 0, 1, \dots$$

Observe that there have been $n-1$ of these transitions from 1 node to n . Hence, the number of $T3$ arrivals on $(0, \tau_n]$ is the sum of $(n-1)$ i.i.d. Geometric random variables with mean $\frac{1-p}{\lambda+p}$. Summarizing the above, we obtain (1). For (2) we also need to observe that the number of $T2$ and $T3$ arrivals on $[0, \tau_n]$ are independent.

A2. Proof of Theorem 2

Proof. Let X_t be the number of $T2$ and $T3$ arrivals by time t . Note that $|E_t| = X_t$, and $|\mathcal{T}_t| = t - X_t + 1$, which is the number of $T1$ arrivals on $[0, t]$, since the first node at time $t = 0$ is by definition a $T1$ arrival. Note that X_t has a binomial distribution with parameters t and $\mathbb{P}(T2 \text{ arrival}) + \mathbb{P}(T3 \text{ arrival}) = \frac{1}{\lambda+1}$. Furthermore, the number of $T1$ arrivals is $t - X_t + 1$ since the first node at time $t = 0$ is by definition a $T1$ arrival. Hence,

$$\begin{aligned} \mathbb{E} \left[\frac{|E_t|}{|\mathcal{T}_t|} \right] &= \sum_{i=1}^t \frac{i}{t-i+1} \binom{t}{i} \left(\frac{1}{\lambda+1}\right)^i \left(\frac{\lambda}{\lambda+1}\right)^{t-i} \\ &= \frac{1}{\lambda} \cdot \sum_{i=1}^t \binom{t}{i-1} \left(\frac{1}{\lambda+1}\right)^{i-1} \left(\frac{\lambda}{\lambda+1}\right)^{t-i+1}, \end{aligned}$$

which proves (3). Next, we write

$$\begin{aligned}
\mathbb{E} \left[\left(\frac{|E_t|}{|\mathcal{T}_t|} \right)^2 \right] &= \sum_{i=0}^t \left(\frac{i}{t-i+1} \right)^2 \binom{t}{i} \left(\frac{1}{\lambda+1} \right)^i \left(\frac{\lambda}{\lambda+1} \right)^{t-i} \\
&= \frac{1}{\lambda} \cdot \sum_{i=1}^t \frac{i}{t-i+1} \binom{t}{i-1} \left(\frac{1}{\lambda+1} \right)^{i-1} \left(\frac{\lambda}{\lambda+1} \right)^{t-i+1} \\
&= \frac{1}{\lambda} \mathbb{E} \left[\frac{t+1}{t-X_t} \mathbb{1}_{\{X_t \leq t-1\}} \right] - \frac{1}{\lambda} \left(1 - \left(\frac{1}{1+\lambda} \right)^t \right), \tag{10}
\end{aligned}$$

where $\mathbb{1}_{\{A\}}$ is an indicator of event A . Denoting

$$Z_t = \frac{X_t - \mathbb{E}[X_t]}{\sqrt{\text{var}(X_t)}} = \frac{(\lambda+1)X_t - t}{\sqrt{\lambda t}}, \tag{11}$$

we further write

$$\mathbb{E} \left[\frac{t+1}{t-X_t} \mathbb{1}_{\{X_t \leq t-1\}} \right] = \mathbb{E} \left[\frac{(t+1)(\lambda+1)}{\lambda t \left(1 - \frac{Z_t}{\sqrt{\lambda t}} \right)} \mathbb{1}_{\{Z_t \leq \sqrt{\lambda t} - \frac{\lambda+1}{\sqrt{\lambda t}}\}} \right]. \tag{12}$$

We now split the indicator above as follows:

$$\mathbb{1}_{\{Z_t \leq -\sqrt{\lambda t}\}} + \mathbb{1}_{\{-\sqrt{\lambda t} < Z_t < \sqrt{\lambda t}/2\}} + \mathbb{1}_{\{\sqrt{\lambda t}/2 \leq Z_t \leq \sqrt{\lambda t} - \frac{\lambda+1}{\sqrt{\lambda t}}\}}. \tag{13}$$

For the first and the third term we use the Chernoff bound:

$$\mathbb{E} \left[\frac{1}{1 - \frac{Z_t}{\sqrt{\lambda t}}} \mathbb{1}_{\{Z_t \leq -\sqrt{\lambda t}\}} \right] \leq 2e^{-\lambda t/4}, \tag{14}$$

$$\mathbb{E} \left[\frac{1}{1 - \frac{Z_t}{\sqrt{\lambda t}}} \mathbb{1}_{\{\sqrt{\lambda t}/2 \leq Z_t \leq \sqrt{\lambda t} - \frac{\lambda+1}{\sqrt{\lambda t}}\}} \right] \leq \frac{\sqrt{\lambda t}}{\lambda+1} 2e^{-\lambda t/16}, \tag{15}$$

and notice that both expressions above converge to zero faster than $1/t$. For the second case, note first that $\mathbb{E}[Z_t] = 0$ and hence it follows from (??) and (??)–(15) that, as $t \rightarrow \infty$,

$$\mathbb{E} \left[Z_t \mathbb{1}_{\{-\sqrt{\lambda t} < Z_t < \sqrt{\lambda t}/2\}} \right] = o \left(\frac{1}{t} \right).$$

Then we use the Taylor expansion to obtain:

$$\begin{aligned}
&\left| \mathbb{E} \left[\frac{1}{1 - \frac{Z_t}{\sqrt{\lambda t}}} \mathbb{1}_{\{-\sqrt{\lambda t} < Z_t < \sqrt{\lambda t}/2\}} \right] - 1 \right| \\
&\leq \mathbb{E} \left[\frac{Z_t^2}{\lambda t} \right] + 2\mathbb{E} \left[\frac{|Z_t|^3}{(\lambda t)^{3/2}} \right] + o \left(\frac{1}{t} \right), \tag{16}
\end{aligned}$$

as $t \rightarrow \infty$. By the central limit theorem, $Z_t \xrightarrow{D} Z$ as $t \rightarrow \infty$. Furthermore, for $r > 0$, the convergence of moments holds [7]: $\lim_{t \rightarrow \infty} \mathbb{E}[|Z_t|^r] = \mathbb{E}[|Z|^r]$. In particular, in (16), $\mathbb{E}[|Z_t|^3]$ converges to a constant, and $\mathbb{E}[Z_t^2]$ converges to 1 as $t \rightarrow \infty$. Thus, using (10)–(12) and (3) we write

$$\begin{aligned} \text{var}\left(\frac{|E_t|}{|\mathcal{T}_t|}\right) &= \mathbb{E}\left[\left(\frac{|E_t|}{|\mathcal{T}_t|}\right)^2\right] - \left(\mathbb{E}\left[\frac{|E_t|}{|\mathcal{T}_t|}\right]\right)^2 \\ &= \mathbb{E}\left[\frac{(t+1)(\lambda+1)}{\lambda t(1 - \frac{Z_t}{\sqrt{\lambda t}})} \mathbb{1}_{\{Z_t \leq \sqrt{\lambda t} - \frac{\lambda+1}{\sqrt{\lambda t}}\}}\right] - \frac{1}{\lambda} - \frac{1}{\lambda^2} + o\left(\frac{1}{t}\right). \end{aligned}$$

Now, subsequently using (??) – (16), we get

$$\begin{aligned} \text{var}\left(\frac{|E_t|}{|\mathcal{T}_t|}\right) &= \frac{1}{\lambda} \frac{(t+1)(\lambda+1)}{\lambda t} \left(1 + \frac{1}{\lambda t} + o\left(\frac{1}{t}\right)\right) \\ &\quad - \frac{1}{\lambda} - \frac{1}{\lambda^2} + o\left(\frac{1}{t}\right), \end{aligned}$$

which results in (4). Statement (??) is proved along similar lines: we apply the expansion directly to the random variable

$$\frac{X_t}{t - X_t + 1} = \frac{(t+1)(\lambda+1)}{(\lambda t + \lambda + 1)(1 - Z_t \frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1})} \mathbb{1}_{\{Z_t \leq \sqrt{\lambda t}\}} - 1,$$

and then use the Chernoff bounds and the CLT to obtain the result.

A3. Proof of Lemma 3

Proof. Assume the arrival at time $t+1$ is of type T1. This occurs w.p. $\frac{\lambda}{\lambda+1}$, and then a new component consisting of size one is created in G_{t+1} , corresponding to the first case in (3).

Next, consider a T2 arrival, which occurs w.p. $\frac{p}{\lambda+1}$. We now add a node to an existing component C_i w.p. $\frac{|C_i|}{|V|}$. Thus the probability that we add the new node to C_i is $\frac{p}{\lambda+1} \cdot \frac{|C_i|}{|V|}$.

Last, we consider a T3 arrival. In this case we have two options. The new edge can either join two components, or join two nodes that are already in one component. For the first case, we derive the probability that C_i and C_j join as

$$\mathbb{P}(C_i \text{ and } C_j \text{ merge}) = \frac{1-p}{\lambda+1} \cdot \frac{2 \cdot |C_i| \cdot |C_j|}{|V|^2 - |V|}.$$

Then for the second case, the number of ways a T3 arrival links two nodes that are already connected in a component, say C_i , is $|C_i|(|C_i| - 1)$. Therefore with probability $\frac{\sum_{k=1}^m |C_k|^2 - |C_i|}{|V|^2 - |V|}$ the component size does not change.

Graph Structure in the Web — Revisited

or A Trick of the Heavy Tail

Robert Meusel
Data and Web Science Group
University of Mannheim
Germany
robert@informatik.uni-mannheim.de

Sebastiano Vigna
Laboratory for Web
Algorithmics
Università degli Studi di Milano
Italy
vigna@acm.org

Oliver Lehmburg
Data and Web Science Group
University of Mannheim
Germany
oli@informatik.uni-mannheim.de

Christian Bizer
Data and Web Science Group
University of Mannheim
Germany
chris@informatik.uni-mannheim.de

ABSTRACT

Knowledge about the general graph structure of the World Wide Web is important for understanding the social mechanisms that govern its growth, for designing ranking methods, for devising better crawling algorithms, and for creating accurate models of its structure. In this paper, we describe and analyse a large, publicly accessible crawl of the web that was gathered by the Common Crawl Foundation in 2012 and that contains over 3.5 billion web pages and 128.7 billion links. This crawl makes it possible to observe the evolution of the underlying structure of the World Wide Web within the last 10 years: we analyse and compare, among other features, degree distributions, connectivity, average distances, and the structure of weakly/strongly connected components.

Our analysis shows that, as evidenced by previous research [17], some of the features previously observed by Broder *et al.* [10] are very dependent on artefacts of the crawling process, whereas other appear to be more structural. We confirm the existence of a giant strongly connected component; we however find, as observed by other researchers [12, 5, 3], very different proportions of nodes that can reach or that can be reached from the giant component, suggesting that the “bow-tie structure” as described in [10] is strongly dependent on the crawling process, and to the best of our current knowledge is not a structural property of the web.

More importantly, statistical testing and visual inspection of size-rank plots show that the distributions of indegree, outdegree and sizes of strongly connected components are not power laws, contrarily to what was previously reported for much smaller crawls, although they might be heavy tailed. We also provide for the first time accurate measurement of distance-based features, using recently introduced algorithms that scale to the size of our crawl [8].

Categories and Subject Descriptors

H.3.4 [Information storage and retrieval]: Systems and software—World Wide Web (WWW)

Keywords

World Wide Web; Web Graph; Network Analysis; Graph Analysis; Web Mining

1. INTRODUCTION

The evolution of the World Wide Web is summarized by Hall and Tiropinis as the development from “the web of documents” in the very beginning, to “the web of people” in the early 2000’s, to the present “web of data and social networks” [13]. With the evolution of the World Wide Web (WWW), the corresponding web graph has grown and evolved as well.

Knowledge about the general graph structure of the web graph is important for a number of purposes. From the structure of the web graph, we can provide evidence for the social phenomena governing the growth of the web [13]. Moreover, the design of *exogenous* ranking mechanisms (i.e., based on the links between pages) can benefit from deeper knowledge of the web graph, and the very process of crawling the web can be made more efficient using information about its structure. Finally, studying the web can help to detect rank manipulations such as spam networks, which publish large numbers of “fake” links in order to increase the ranking of a target page.

In spite of the importance of knowledge about the structure of the web graph, the latest publicly accessible analysis of a large global crawl is nearly a decade old. The first, classic work about the structure of the web as a whole was published by Broder *et al.* [10] in 2000 using an AltaVista crawl of 200 million pages and 1.5 billion links.¹ A second similar crawl was used to validate the results.

One of their main findings was a *bow-tie* structure within the web graph: a giant strongly connected component containing 28% of the nodes. In addition, Broder *et al.* show that the indegree distribution, the outdegree distribution and the distribution of the sizes

¹Throughout the paper, we avoid redundant use of the \approx symbol: all reported figures are rounded.

of strongly connected components are heavy tailed. The paper actually claims the distributions to follow power laws, but provides no evidence in this sense except for the fact that the data points in the left part of the plots are gathered around a line. The authors comment also on the fact that the initial part of the distributions displays some concavity on a log-log plot, which requires further analysis.

An important observation that has been made by Serrano *et al.* [17] analysing four crawls gathered between 2001 and 2004 by different crawlers with different parameters is that *several properties of web crawls are dependent on the crawling process*. Maybe a bit optimistically, Broder *et al.* claimed in 2000 that “These results are remarkably consistent across two different, large AltaVista crawls. This suggests that our results are relatively insensitive to the particular crawl we use, provided it is large enough”. We now know that this is not true: several studies [12, 5, 3, 21] using different (possibly regional) crawls gathered by different crawlers provided quite different pictures of the web graph (e.g., that “daisy” of [12] or the “teapot” of [21]).

In particular, recent strong and surprising results [1] have shown that, in principle, most heavy-tailed (and even power-law) distributions observed in web crawls may be just an artefact of the crawling process itself. It is very difficult to predict when and how we will be able to understand fully whether this is true or not.

Subsequent studies confirmed the existence of a large strongly connected component, usually significantly larger than previously found, and heavy-tailed (often, power-law) distributions. However, such studies used even smaller web crawls while the size of the web was approaching the tera scale, and provided the same, weak visual evidence about distribution fitting. While no crawl can claim to represent the web as a whole (even large search engines crawl only a small portion of the web, geographically, socially and economically selected) the increase in scale of the web requires the analysis of crawls an order of magnitude larger. Nonetheless, billion-scale representative crawls have not been available to the scientific community until very recently. Thus, only large companies such as Google, Yahoo!, Yandex, and Microsoft had updated knowledge about the structure of large crawls of the WWW.

A few exceptions exist, but they have significant problems. The AltaVista webpage connectivity dataset, distributed by Yahoo! as part of the WebScope program, has in theory 1.4 billion nodes, but it is extremely disconnected: half of the nodes are isolated (no links incoming or outgoing) and the largest strongly connected component is less than 4% of the whole graph, which makes it entirely unrepresentative. We have no knowledge of the crawling process, and URLs have been anonymised, so no investigation of the causes of these problems is possible.

The ClueWeb09 graph, gathered in 2009 within the U.S. National Science Foundation’s Cluster Exploratory (CluE), has a similar problem due to known mistakes in the link construction, with a largest strongly connected component that is less the 3% of the whole graph. As such, these two crawls cannot be used to infer knowledge about the structure of the web.

The ClueWeb12 crawl, released concurrently with the writing of this paper, has instead an accurate link structure, and contains a largest strongly connected component covering 76% of the graph. The crawl, however, is significantly smaller than the graph used in this paper, as it contains 1.2 billion pages,² and it is focused mostly on English web pages.

²Note that the web graph distributed with ClueWeb09 and ClueWeb12 appears to be much larger because all *frontier* nodes have been included in the graph. The number we report are those of the actually crawled pages.

In this paper, we try to update the original studies on the structure of the web and its current state. We revisit and update the findings of previous research to give an up-to-date view of the web graph today, using a crawl that is significantly larger (3.5 billion pages) than the ones used in previous work.

We repeat previous measurement, observing interesting differences, and provide new, previously unknown data, such as the distance distribution. The crawl³ as well as the hyperlink graph⁴ are publicly available, so to encourage other researchers and analysts to replicate our results and investigate in further interesting topics.

2. DATASET AND METHODOLOGY

The object of study of this paper is a large web crawl gathered by the Common Crawl Foundation⁵ in the first half of 2012. The crawl contains 3.83 billion web documents, of which over 3.53 billion (92%) are of mime-type `text/html`. The crawler used by the Common Crawl (CC) Foundation for the crawl is based on a breath-first visiting strategy, together with heuristics to detect spam pages. In addition heuristics were used to reduce the number of crawled pages with duplicate or no content. Such heuristics, in principle, may cut some of the visiting paths and make the link structure sparser. The crawl was seeded with the list of pay-level-domain names from a previous crawl and a set of URLs from Wikipedia. The list of seeds was ordered by the number of external references. Unfortunately this list is not public accessible, but we estimated that at least 71 million different seeds were used, based on our observations on the ratio between pages and domains. The selected amount of seeds in combination with the methodology are likely to affect the distribution of host sizes, as popular websites were crawled more intensely: for example, `youtube.com` is represented by 93.1 million pages within the crawl [18]. In addition, it is likely that the large number of seeds used in the multiple phases of the crawl caused the large number of pages of indegree zero (20% of the graph) found in the graph.

Associated with the crawl is a *web graph*, in which each node represents a page and each arc between two nodes represents the existence of one or more hypertextual links between the associated pages. We extracted the web graph from the crawl with a 3-step process, using an infrastructure similar to the framework used by Bizer *et al.* to parse the Common Crawl corpus and extract structured data embedded in HTML pages [4]. We first collected for each crawled page its URL, mime-type, links to other pages, type, and, if available, the redirect URL, using 100 parallel `c1.xlarge` Amazon Elastic Compute Cloud (EC2) machine instances. We then filtered the extracted URLs by mime-type `text/html` and kept only links within HTML elements of type `a` and `link`, as we want to focus on HTML pages linking to other HTML pages.⁶ Also redirects contained in HTTP header have been treated as links. Finally, we used a 40-node Amazon Elastic MapReduce cluster to compress the graph, indexing all URLs and remove duplicate links.

Additionally, we built the host graph and the pay-level-domain (PLD) graph. Nodes in such graphs represent sets of pages with the

³<https://commoncrawl.atlassian.net/wiki/display/CRWL/About+the+Data+Set>

⁴<http://webdatacommons.org/hyperlinkgraph/>

⁵<http://commoncrawl.org/>

⁶We remark that this choice might have introduced some sparsity, as in principle the crawling process might have followed further links, such as `src` attributes of `iframe` elements. Keeping perfectly aligned the online (during the crawl) and offline (in a separate pass after the crawl) link extraction process when they are performed by different organisations is, unfortunately, quite difficult, as link and page selection strategies could differ.

same host/pay-level-domain, and there is an arc between nodes x and y if there is at least one arc from a page in the set associated with x to a page in the set associated with y . Table 1 provides basic data about the size of the graphs.

Granularity	# Nodes in millions	# Arcs in millions
Page Graph	3 563	128 736
Host Graph	101	2 043
PLD Graph	43	623

Table 1: Sizes of the graphs

3. ANALYSIS OF THE WEB GRAPH

Most of the analyses presented in the following section have been performed using the “big” version of the WebGraph framework [6], which can handle more than 2^{31} nodes. The BV compression scheme was able to compress the graph *in crawl order* at 3.52 bits per link, which is just 12.6% of the information-theoretical lower bound (under a suitable permutation of the node identifiers it is common to obtain slightly more than one bit per link). The whole graph occupied in compressed form just 57.5 GB, which made it possible to run resource intensive computations such as the computation of the strongly connected components.

3.1 Indegree & Outdegree Distribution

The simplest indicator of density of web graphs is the average degree, that is, the ratio between the number of arcs and the number of nodes in the graph.⁷

Broder *et al.* report an average degree of 7.5 links per page. Similar low values can be found in crawls of the same years—for instance, in the crawls made by the Stanford WebBase project.⁸ In contrast our graph has average degree of 36.8, meaning that the average degree is factor 4.9 larger than in the earlier crawls. Similar values can be found in 2007 .uk crawls performed by the Laboratory for Web Algorithmics, and the ClueWeb12 crawl has average degree 45.1.⁹ A possible explanation for the increase of the average degree is the wide adoption of *content management systems*, which tend to create dense websites.

Figures 1 and 2 show frequency plots of indegrees and outdegrees in log-log scale. For each d , we plot a point with an ordinate equal to the number of pages with that have degree d . Note that we included the data for degree zero, which is omitted in most of the literature. We then aggregate the values using *Fibonacci binning* [19] to show the approximate shape of the distribution.

Finally, we try to fit a power law to a tail of the data. This part is somewhat delicate: previous work in the late 90’s has often claimed to find power laws just by noting an approximate linear shape in log-log plots: unfortunately, almost all distributions (even, sometime, non-monotone ones) look like a line on a log-log plot [20]. Tails exhibiting high variability, in particular, are very noisy (see the typical “clouds of points” in the right part of degree plots) and difficult to interpret.

⁷Technically speaking, the *density* of a graph is the ratio between the *square* of the number of nodes and the number of arcs, but for very sparse graphs one obtains abysmally small numbers that are difficult to interpret.

⁸<http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/>

⁹We remark that all these values are actually an underestimation, as they represent the average number of outgoing arcs *in the web graph built from the crawl*. The average number of links per page can be higher, as several links will point outside the graph.

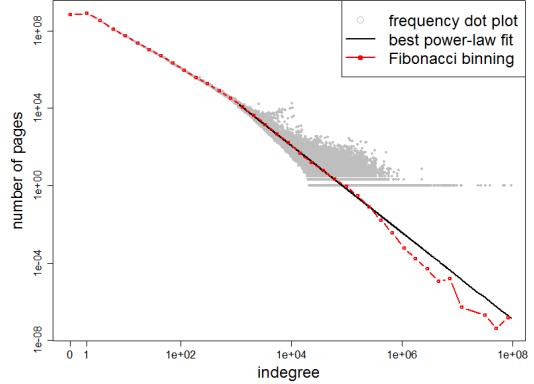


Figure 1: Frequency plot of the indegree distribution

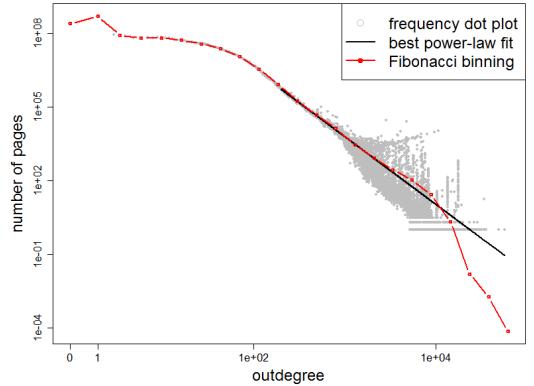


Figure 2: Frequency plot of the outdegree distribution

We thus follow the methodological suggestions of Clauset *et al.* [11]. We use the `plfit`¹⁰ tool to attempt a maximum-likelihood fitting of a power law starting from each possible degree, keeping the starting point and the exponent providing the best likelihood. After that we perform a goodness-of-fit test and estimate a p -value.

The first important fact we report is that *the p-value of the best fits is 0 (± 0.01)*. In other words, from a statistical viewpoint, in spite of some nice graphical overlap the tail of the distribution is *not* a power law. We remark that this paper applies for the first time a sound methodology to a large dataset: it is not surprising that the conclusions diverge significantly from previous literature.

To have some intuition about the possibility of a heavy tail (i.e., that the tail of the distribution is not exponentially bounded) we draw the *size-rank* plot, as suggested in [14]. The size-rank plot is the discrete version of the complementary cumulative distribution function in probability: if the data fits a power law it should display as a line on a log-log scale. Concavity indicates a superpolynomial decay. Size-rank plots are monotonically decreasing functions, and do not suffer the “cloud of points” problem.

Figure 3 shows the size-rank plot of the degree distributions of our graph and the best power-law fit: from what we can ascertain visually, there is a clear concavity, indicating once again that the tail of the distribution is not a power law. The concavity leaves open the possibility of a non-fat heavy tail, such as that of a lognormal distribution.

¹⁰<https://github.com/ntamas/plfit>

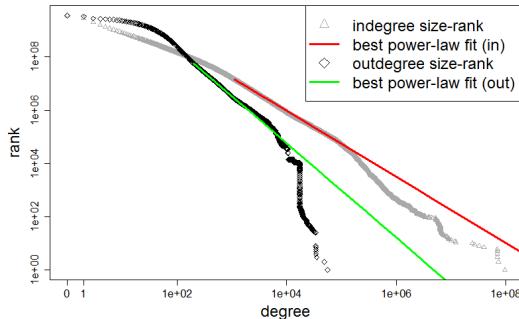


Figure 3: Size-rank plot of degree distributions

In any case, the tails providing the best fit characterize a very small fraction of the probability distribution: for indegrees, we obtain an exponent 2.24 starting at degree 1 129, whereas for outdegrees we obtain an exponent 2.77 starting at 199, corresponding, respectively, to 0.4% and less than 2% of the probability mass (or, equivalently, fraction of nodes). Models replicating this behaviour, thus, explain very little of the process of link formation in the web.

The values we report are slightly different than those of Broder *et al.*, who found 2.09 respectively 2.72 as power-law exponent for the indegree respectively outdegree. But in fact they are incomparable, as our fitting process used different statistical methods.

Finally, the largest outdegree is three magnitudes smaller than the largest indegree. This suggests that the decay of the indegree distribution is significantly slower than that of the outdegree distribution, a fact confirmed by Figure 3.

3.2 High Indegree Pages and Hosts

The three web pages with highest indegree are the starting pages of YouTube, WordPress and Google. Other six pages from YouTube from the privacy, press and copyright sections of this website appear within the top 10 of pages ranked by their indegree. This is an artefact of the large number of pages crawled from YouTube.¹¹

The list of *hosts* with the highest indegree (in the host graph) is more interesting: in Table 2 we show the top 20 hosts by indegree, PageRank [16] and harmonic centrality [9]. While most of the sites are the same, some noise appears because some sites are highly linked for technical or political reasons. In particular, the site `miibeian.gov.cn` must be linked by every Chinese site, hence the very high ranking. PageRank is as usual very correlated to degree, and cannot avoid ranking highly this site, whereas harmonic centrality understands its minor importance and ranks it at position 6146.

3.3 Components

Following the steps of Broder *et al.*, we now analyse the weakly connected components (WCC) of our web graph.

Weakly connected components are difficult to interpret—in theory, unless one has two seed URLs reaching completely disjoint regions of the web (unlikely), one should always find a single weakly connected component. The only other sources of disconnection are crawling and/or parsing artefacts.

Figure 4 shows the distribution of the sizes of the weakly connected components using a visualization similar to the previous figures. The largest component (rightmost grey point) contains about around 94% of the whole graph, and it is slightly larger than the

¹¹The highest ranked pages are listed at http://webdatacommons.org/hyperlinkgraph/top_degree_pages.html.

PageRank	Indegree	Harmonic Centrality
gmpg.org	wordpress.org	youtube.com
wordpress.org	youtube.com	en.wikipedia.org
youtube.com	gmpg.org	twitter.com
livejournal.com	en.wikipedia.org	google.com
tumblr.com	twitter.com	wordpress.org
en.wikipedia.org	google.com	flickr.com
networkadvertising.org	flickr.com	facebook.com
promodj.com	rtalabel.org	apple.com
skriptmail.de	wordpress.com	vimeo.com
parallels.com	mp3shake.com	creativecommons.org
tistory.com	w3schools.com	amazon.com
google.com	domains.lycos.com	adobe.com
miibeian.gov.cn	staff.tumblr.com	myspace.com
phpbb.com	club.tripod.com	w3.org
blog.fc2.com	creativecommons.org	bbc.co.uk
tw.yahoo.com	vimeo.com	nytimes.com
w3schools.com	miibeian.gov.cn	yahoo.com
wordpress.com	facebook.com	microsoft.com
	phpbb.com	guardian.co.uk
		imdb.com

Table 2: The 20 top web hosts by PageRank, indegree and harmonic centrality (boldfaced entries are unique to the list they belong to)

one reported by Broder *et al.* (91.8%). Again, we show the maximum-likelihood power-law fit starting at 14 with exponent 2.22, which however excludes the largest component. The *p*-value is again 0, and the law covers only to 1% of the distribution.

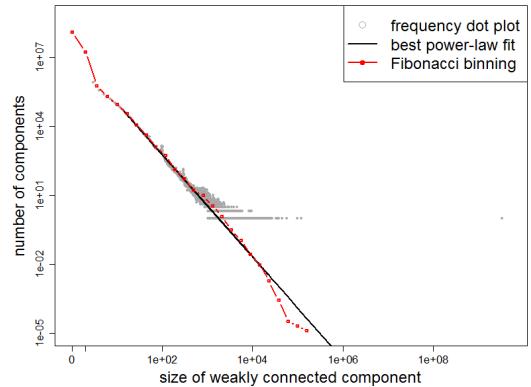


Figure 4: Frequency plot of the distribution of WCCs

More interestingly, we now analyse the strongly connected components (SCC). Computing the strongly connected components of a 3.5 billion node graph was no easy task: it required one terabyte of core memory and, in fact, the computation was only possible because WebGraph [6] uses *lazy* techniques to generate successor lists (i.e., successors lists are never actually stored in memory in uncompressed form).

Figure 5 shows the distribution of the sizes of the strongly connected components. The largest component (rightmost grey point) contains 51.3% of the nodes. Again, we show a fitted power law starting at 22 with exponent 2.20, which however excludes the largest component, and fits only to 8.9% of the distribution. The *p*-value is again 0.

In Figure 6 we show the size-rank plots of both distributions, which confirm again that the apparent fitting in the previous figures is an artefact of the frequency plots (the rightmost grey points are again the giant components).

3.4 The Bow Tie

Having identified the giant strongly connected component, we can determine the so-called *bow tie*, a depiction of the structure of the web suggested by Broder *et al.*. The bow tie is made of six different components:

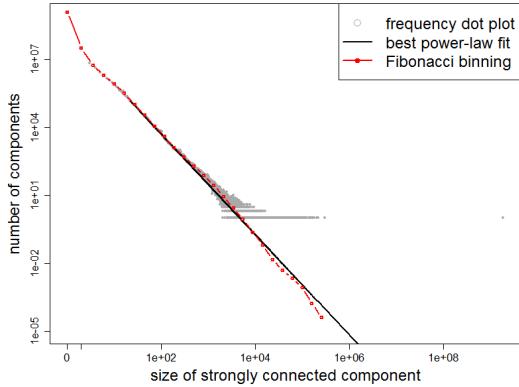


Figure 5: Frequency plot of the distribution of SCCs

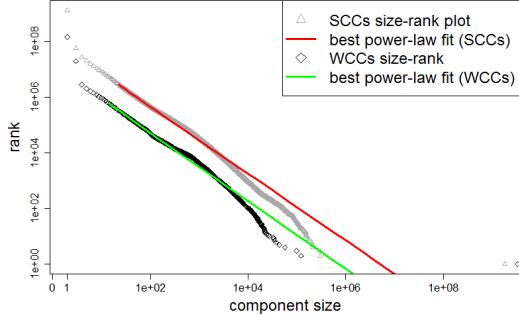


Figure 6: Size-rank plot of the distribution of components

- the core is given by the giant strongly connected component (LSCC);
- the IN component contains non-core pages that can reach the core via a directed path;
- the OUT component contains non-core pages that can be reached from the core;
- the TUBES are formed by non-core pages reachable from IN and that can reach OUT;
- pages reachable from IN, or that can reach OUT, but are not listed above, are called TENDRILS;
- the remaining pages are DISCONNECTED.

All these components are easily computed by visiting the *direct acyclic graph of strongly connected components* (SCC DAG): it is a graph having one node for each strongly connected component with an arc from x to y if some node in the component associated with x is connected with a node in the component associated with y . Such a graph can be easily generated using WebGraph's facilities. Figure 7 shows the size of bow-tie component.

Table 3 compares the sizes of the different components of the bow-tie structure between the web graph discussed in this paper (column two and three) and the web graph analysed by Broder *et al.* in 2000 (column four and five).¹²

¹²Broder *et al.* did not report the number of nodes belonging to the TUBE component separately, as they define as TUBE as a TENDRIL from the IN component hooked into the TENDRIL of a node from the OUT component.

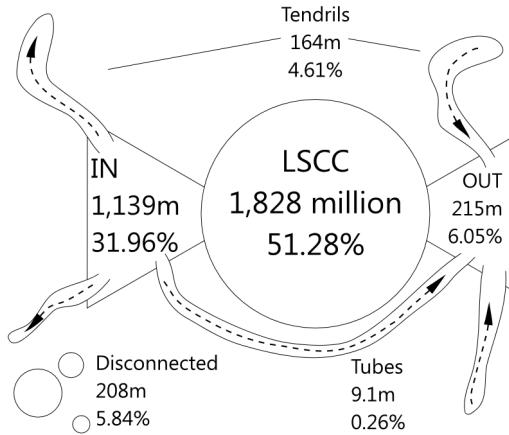


Figure 7: Bow-tie structure of the web graph

The main constant is the existence of a LSCC, which in our graph has almost doubled in relative size. We also witness a much smaller OUT component and a larger IN component. The different proportions are most likely to be attributed to different crawling strategies (in particular, to our large number of nodes with indegree zero, which cannot belong to the LSCC or OUT component). Unfortunately, basic data such as the seed size, the type of visit strategy, etc. are not available for the Broder *et al.* crawl. Certainly, however, the web has become significantly more dense and connected in the last 13 years.

Component	Common Crawl 2012		Broder <i>et al.</i>	
	# nodes (in thousands)	% nodes (in %)	# nodes (in thousands)	% nodes (in %)
LSCC	1 827 543	51.28	56 464	27.74
IN	1 138 869	31.96	43 343	21.29
OUT	215 409	6.05	43 166	21.21
TENDRILS	164 465	4.61	43 798	21.52
TUBES	9 099	0.26	-	-
DISC.	208 217	5.84	16 778	8.24

Table 3: Comparison of sizes of bow-tie components

3.5 Diameter and Distances

In this paper we report, for the first time, accurate measurements of distance-related features of a large web crawl. Previous work has tentatively used a small number of breadth-visit samples, but convergence guarantees are extremely weak (in fact, almost non-existent) for graphs that are not strongly connected. The data we report have been computed using HyperBall [8], a diffusion-based algorithm that computes an approximation of the distance distribution (technically, we computed four runs with relative standard deviation 9.25%). We report, for each datum, the empirical standard error computed by the jackknife resampling method.

In our web graph, $48.15 \pm 2.14\%$ of the pairs of pages have a connecting directed path. Moreover, the average distance is 12.84 ± 0.09 and the *harmonic diameter* (the harmonic mean of all distances, see [15] and [7] for motivation) is 24.43 ± 0.97 . These figures should be compared with the 25% of connected pairs and the average distance 16.12 reported by Broder *et al.* (which however have been computed averaging the result of few hundred breadth-first samples): even if our crawl is more than 15 times larger, it is significantly more connected, in contrast to commonly accepted predictions of logarithmic growth of the diameter in terms of the

number of nodes. This is a quite general phenomenon: the average distance between Facebook users, for instance, has been steadily going down as the network became *larger* [2].

We can also estimate that the graph has a diameter of at least 5 282 (the maximum number of iteration of a HyperBall run). Figure 8 shows the distance distribution, sharply concentrated around the average.

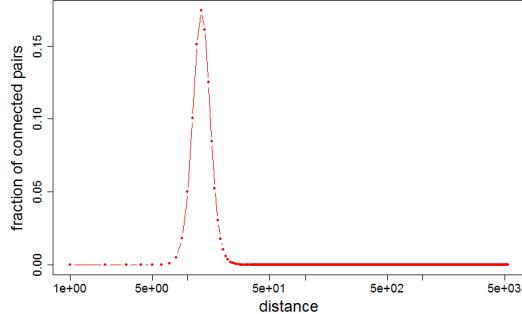


Figure 8: Distance distribution

4. CONCLUSION

We have reported a number of graph measurements on the largest web graph that is available to the public outside companies such as Google, Yahoo, Yandex, and Microsoft. Comparing our results with previous measurements performed in the last 13 years, and with previous literature on significantly smaller crawls, we reach the following conclusions:

- The average degree has significantly increased, almost by a factor of 5.
- At the same time, the connectivity of the graph (the percentage of connected pairs) has increased (almost twice) and the average distance between pages has decreased, in spite of a predicted growth that should have been logarithmic in the number of pages.
- While we can confirm the existence of a large strongly connected component of growing size, witnessing again the increase in connectivity, the structure of the rest of the web appears to be very dependent on the specific web crawl. While it is always possible to compute the components of the bow tie of Broder *et al.*, the proportion of the components is not intrinsic.
- The distribution of indegrees and outdegrees is extremely different. Previous work on a smaller scale did not detect or underplayed this fact, in part because of the little size of the concave (on a log-log plot) part of the distribution in smaller crawls. In our dataset, the two distributions have very little in common.
- The frequency plots of degree and component-size distributions are visually identical to previous work. However, using proper statistical tools, neither degree nor component-size distributions fit a power law. Moreover, visual inspection of the size-rank plots suggests that their tails are not fat (i.e., they decrease faster than a polynomial), in contrast with assumptions taken for granted in the current literature. Our data, nonetheless, leaves open the possibility of a heavy tail (e.g., lognormal).

5. ACKNOWLEDGEMENTS

The extraction of the web graph from the Common Crawl was supported by the FP7-ICT project PlanetData (GA 257641) and by an Amazon Web Services in Education Grant award. Sebastiano Vigna has been supported by the EU-FET grant NADINE (GA 288956), which provided part of the high-end hardware on which the analysis was performed.

6. REFERENCES

- [1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: Or, power-law degree distributions in regular graphs. *Journal ACM*, 56(4):21:1–21:28, 2009.
- [2] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *ACM Web Science 2012: Conference Proceedings*, pages 45–54. ACM Press, 2012.
- [3] R. Baeza-Yates and B. Poblete. Evolution of the Chilean web structure composition. In *Proc. of Latin American Web Conference 2003*, pages 11–13, 2003.
- [4] C. Bizer, K. Eckert, R. Meusel, H. Mühlleisen, M. Schuhmacher, and J. Völker. Deployment of RDFa, microdata, and microformats on the web - a quantitative analysis. In *Proc. of the In-Use Track International Semantic Web Conference 2013*, Oct 2013.
- [5] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural properties of the African web. In *Proc. WWW'02*, 2002.
- [6] P. Boldi and S. Vigna. The WebGraph framework I: Compression techniques. In *Proc. WWW'04*, pages 595–602. ACM, 2004.
- [7] P. Boldi and S. Vigna. Four degrees of separation, really. In *ASONAM 2012*, pages 1222–1227. IEEE Computer Society, 2012.
- [8] P. Boldi and S. Vigna. In-core computation of geometric centralities with HyperBall: A hundred billion nodes and beyond. In *ICDMW 2013*. IEEE, 2013.
- [9] P. Boldi and S. Vigna. Axioms for centrality. *Internet Math.*, 2014. To appear.
- [10] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web: experiments and models. *Computer Networks*, 33(1–6):309–320, 2000.
- [11] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, Nov. 2009.
- [12] D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas. Mining the inner structure of the web graph. In *WebDB*, pages 145–150, 2005.
- [13] W. Hall and T. Tiropanis. Web evolution and web science. *Computer Networks*, 56(18):3859 – 3865, 2012.
- [14] L. Li, D. L. Alderson, J. Doyle, and W. Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Math.*, 2(4), 2005.
- [15] M. Marchiori and V. Latora. Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications*, 285(3-4):539 – 546, 2000.
- [16] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library Technologies Project, Stanford University, 1998.
- [17] M. Serrano, A. Maguitman, M. Boguñá, S. Fortunato, and A. Vespignani. Decoding the structure of the WWW: A comparative analysis of web crawls. *TWEB*, 1(2):10, 2007.
- [18] S. Spiegler. Statistics of the Common Crawl Corpus 2012. Technical report, SwiftKey, June 2013.
- [19] S. Vigna. Fibonacci binning. *CoRR*, abs/1312.3749, 2013.
- [20] W. Willinger, D. Alderson, and J. C. Doyle. Mathematics and the Internet: A source of enormous confusion and great potential. *Notices of the AMS*, 56(5):586–599, 2009.
- [21] J. J. H. Zhu, T. Meng, Z. Xie, G. Li, and X. Li. A teapot graph and its hierarchical structure of the Chinese web. *Proc. WWW'08*, pages 1133–1134, 2008.

Cache-Oblivious Peeling of Random Hypergraphs*

Djamal Belazzougui¹, Paolo Boldi², Giuseppe Ottaviano³,
 Rossano Venturini⁴, and Sebastiano Vigna²

¹Department of Computer Science, University of Helsinki,
 djamal.belazzougui@cs.helsinki.fi

²Dipartimento di Informatica, Università degli Studi di Milano,
 {boldi,vigna}@di.unimi.it

³ISTI-CNR, Pisa, giuseppe.ottaviano@isti.cnr.it

⁴Dipartimento di Informatica, Università di Pisa, rossano@di.unipi.it

Abstract

The computation of a peeling order in a randomly generated hypergraph is the most time-consuming step in a number of constructions, such as perfect hashing schemes, random r -SAT solvers, error-correcting codes, and approximate set encodings. While there exists a straightforward linear time algorithm, its poor I/O performance makes it impractical for hypergraphs whose size exceeds the available internal memory.

We show how to reduce the computation of a peeling order to a small number of sequential scans and sorts, and analyze its I/O complexity in the cache-oblivious model. The resulting algorithm requires $O(\text{sort}(n))$ I/Os and $O(n \log n)$ time to peel a random hypergraph with n edges.

We experimentally evaluate the performance of our implementation of this algorithm in a real-world scenario by using the construction of minimal perfect hash functions (MPHF) as our test case: our algorithm builds a MPHF of 7.6 billion keys in less than 21 hours on a single machine. The resulting data structure is both more space-efficient and faster than that obtained with the current state-of-the-art MPHF construction for large-scale key sets.

*Paolo Boldi and Sebastiano Vigna were supported by the EU-FET grant NADINE (GA 288956). Giuseppe Ottaviano was supported by Midas EU Project (318786), MaRea project (POR-FSE-2012), and Tiscali. Rossano Venturini was supported by the MIUR of Italy project PRIN ARS Technomedia 2012 and the eCloud EU Project (325091).

1 Introduction

Hypergraphs can be used to model sets of dependencies among variables of a system: vertices correspond to variables and edges to relations of dependency among variables, such as equations binding variables together. This correspondence can be used to transfer graph-theoretical properties to solvability conditions in the original system of dependencies.

Among these, one of the most useful is the concept of peeling order. Given an r -hypergraph, a *peeling order* is an order of its edges such that each edge has a vertex of degree 1 in the subgraph obtained by removing the previous edges in the order. Such an order exists if the hypergraph does not have a non-empty 2-core, i.e. a set of vertices that induces a subgraph whose vertices have all degree at least 2.

In the above interpretation, if the equations of a system are arranged in peeling order, then each equation has at least one variable that does not appear in any equation that comes later in the ordering, i.e., the system becomes *triangular*, so it can be easily solved by backward substitution. For this reason, peeling orders found application in a number of fundamental problems, such as hash constructions [3, 6, 9, 10, 11, 12, 21], solving random instances of r -SAT [12, 24, 25], and the construction of error-correcting codes [15, 20, 23]. These applications exploit the guarantee that if the edge sparsity γ of a random r -hypergraph is larger than a certain sparsity threshold c_r (e.g., $c_3 \approx 1.221$), then with high probability the hypergraph has an empty 2-core [25].

The construction of *perfect hash functions* (PHF) is probably the most important of the aforementioned applications. Given a set S of n keys, a PHF for S maps the n keys onto the set of the first m natural numbers bijectively. A perfect hash function is *minimal* (MPHF) if $m = n = |S|$. A lower bound by Mehlhorn [22] states that $n \log e \approx 1.44n$ bits are necessary to represent a MPHF; a matching (up to lower order terms) upper bound is provided in [16], but the construction is impractical. Most practical approaches, instead, are based on random 3-hypergraphs, resulting in MPHFs that use about $2c_3n \approx 2.5n$ bits [6, 10, 21]. These solutions, which we review in Section 3, build on the MWHC technique [21], whose most demanding task is in fact the computation of a peeling order.

There is a surprisingly simple greedy algorithm to find a peeling order when it exists, or a 2-core when it does not: find a vertex of degree 1, remove (*peel*) its only edge from the hypergraph, and iterate this process until either no edges are left (in which case the removal order is a peeling order), or all the non-isolated vertices left have degree at least 2 (thus forming a 2-core). This algorithm can be easily implemented to run in linear time and space.

MPHFs are the main ingredient in many space-efficient data structures, such as (compressed) full-text indexes [4], monotone MPHFs [1], Bloom filter-like data structures [5], and prefix-search data structures [2].

It should be clear that the applications that benefit the most from such data structures are those involving large-scale key sets, often orders of magnitude larger than the main memory. Unfortunately, the standard linear-time peeling algorithm requires several tens of bytes per key of working memory, even if the final data structure can be stored in just a handful of bits per key. It is hence common that, while the data structure fits in memory, such memory is not enough to actually *build* it. It is then necessary to resort to external memory, but the poor I/O performance of the algorithm makes such an approach impossible.

Application-specific workarounds have been devised; for example, Botelho et al. [6] proposed an algorithm (called HEM) to build MPHFs in external memory by splitting the key set into small buckets and computing independent MPHFs for each bucket. A first-level index is used to find the bucket of a given key. The main drawback of this solution is that the first-level index introduces a non-negligible overhead in both space and lookup time; moreover, this construction cannot be extended to applications other than hashing.

In this paper we provide the first efficient algorithm in the *cache-oblivious* model that, given a random r -hypergraph with n edges and γn vertices (with $r = O(1)$ and $\gamma > c_r$), computes a peeling order in time $O(n \log n)$ and with $O(\text{sort}(n))$ I/Os w.h.p., where $\text{sort}(n)$ is the I/O complexity of sorting n keys. By applying this result we can construct (monotone) MPHFs, static functions, and Bloom filter-like data structures in $O(\text{sort}(n))$ I/Os. In our experimental evaluation, we show that the algorithm makes it indeed possible to peel very large hypergraphs: an MHPF for a set of 7.6 billion keys is computed in less than 21 hours; on the same hardware, the standard algorithm would not be able to manage more than 2.1 billion keys. Although we use minimal perfect hash functions construction as our test case, results of these experiments remain valid for all the other applications due to the random nature of the underlying hypergraphs.

2 Notation and tools

Model and assumptions We analyze our algorithms in the cache-oblivious model [14]. In this model, the machine has a two-level memory hierarchy, where the fast level has an unknown size of M words and a slow level of unbounded size where our data reside. We assume that the fast level plays the role of a cache for the slow level with an optimal replacement strategy where the transfers (a.k.a. I/Os) between the two levels are done in blocks of an unknown size of $B \leq M$ words; the I/O cost of an algorithm is the total number of such block transfers. *Scanning* and *sorting* are two fundamental building blocks in the design of cache-oblivious algorithms [14]: under the tall-cache assumption [8], given an array of N contiguous items the I/Os required for scanning and sorting are

$$\text{scan}(N) = O\left(1 + \frac{N}{B}\right) \text{ I/Os} \quad \text{and} \quad \text{sort}(N) = O\left(\frac{N}{B} \log_{M/B} \frac{N}{B}\right).$$

Hypergraphs An r -hypergraph on a vertex set V is a subset E of $\binom{V}{r}$, the set of subsets of V of cardinality r . An element of E is called an *edge*. We call an ordered r -tuple from V an *oriented edge*; if e is an edge, an oriented edge whose vertices are those in e is called an *orientation* of e . From now on we will focus on 3-hypergraphs; generalization to arbitrary r is straightforward. We define *valid* orientations those oriented edges (v_0, v_1, v_2) where $v_1 < v_2$ (for arbitrary r , $v_1 < \dots < v_{r-1}$). Then for each edge there are 6 orientations, but only 3 valid orientations ($r!$ orientations of which r are valid).

We say that a valid oriented edge (v_0, v_1, v_2) is the i -th orientation if v_0 is the i -th smallest among the three; in particular, the 0-th orientation is the *canonical* orientation. Edges correspond bijectively with their canonical orientations. Furthermore, valid orientations can be mapped bijectively to pairs (e, v) where e is an edge and v a vertex contained in e , simply by the correspondence $(v_0, v_1, v_2) \mapsto (\{v_0, v_1, v_2\}, v_0)$. In the following all the orientations are assumed to be valid, so we will use the term *orientation* to mean *valid orientation*.

3 The Majewski–Wormald–Havas–Czech technique

Majewski et al. [21] proposed a technique (MWHC) to compute an *order-preserving minimal perfect hash function*, that is, a function mapping a set of keys S in some specified way into $[\lceil |S| \rceil]$. The technique actually makes it possible to store succinctly any function $f : S \rightarrow [\sigma]$, for arbitrary σ . In this section we briefly describe their construction.

First, we choose three random¹ hash functions $h_0, h_1, h_2 : S \rightarrow [\gamma n]$ and generate a 3-hypergraph² with γn vertices, where γ is a constant above the *critical threshold* c_3 [25], by mapping each key x to the edge $\{h_0(x), h_1(x), h_2(x)\}$. The goal is to find an array u of γn integers in $[\sigma]$ such that for each key x one has $f(x) = u_{h_0(x)} + u_{h_1(x)} + u_{h_2(x)} \bmod \sigma$. This yields a linear system with n equations and γn variables u_i ; if the associated hypergraph is peelable, it is easy to solve the system. Since γ is larger than the critical threshold, the algorithm succeeds with probability $1 - o(1)$ as $n \rightarrow \infty$ [25].

By storing such values u_i , each requiring $\lceil \log \sigma \rceil$ bits, plus the three hash functions, we will be able to recover $f(x)$. Overall, the space required will be $\lceil \log \sigma \rceil \gamma n$ bits, which can be reduced to $\lceil \log \sigma \rceil n + \gamma n + o(n)$ using a ranking structure [17]. This technique can be easily extended to construct MPHFs: we define the function $f : S \rightarrow [3]$ as $x \mapsto i$ where $h_i(x)$ is a degree-1 vertex when the edge corresponding to x is peeled; it is then easy to see that $h_{f(x)}(x) : S \rightarrow [\gamma n]$ is a PHF. The function can be again made minimal by adding a ranking structure on the vector u [6].

As noted in the introduction, the peeling procedure needed to solve the linear system can be performed in linear time using a greedy algorithm (referred to as *standard linear-time peeling*). However, this procedure requires random access to several integers per key, needed for bookkeeping; moreover, since the graph is random, the visit order is close to random. As a consequence, if the key set is so large that it is necessary to spill to the disk part of the working data structures, the I/O volume slows down the algorithm to unacceptable rates.

¹Like most MWHC implementations, in our experiments we use a Jenkins hash function with a 64-bit seed in place of a fully random hash function.

²Although the technique works for r -hypergraphs, $r = 3$ provides the lowest space usage [25].

Practical workarounds (HEM) Botelho et al. [6] proposed a practical external-memory solution: they replace each key with a *signature* of $\Theta(\log n)$ bits computed with a random hash function, so that no collision occurs. The signatures are then sorted and divided into small buckets based on their most significant bits, and a separate MPHF is computed for each bucket with the approach described above. The representations of the bucket functions are then concatenated into a single array and their offsets stored in a separate vector.

The construction algorithm only requires to sort the signatures (which can be done efficiently in external memory) and to scan the resulting array to compute the bucket functions; hence, it is extremely scalable. The extra indirection needed to address the blocks causes however the resulting data structure to be both *slower* and *larger* than one obtained by computing a single function on the whole key set. In their experiments with a practical version of the construction, named HEM, the authors report that the resulting data structure is 21% larger than the one built with plain MWHC, and lookups are 30–50% slower. A similar overhead was confirmed in our experiments, which are discussed in Section 5.

4 Cache-oblivious peeling

In this section we describe a cache-oblivious algorithm to peel an r -hypergraph. We describe the algorithm for 3-hypergraphs, but it is easy to generalize it to arbitrary r .

4.1 Maintaining incidence lists

In order to represent the hypergraph throughout the execution of the algorithm, we need a data structure to store the *incidence list* of every vertex v_0 , i.e., the list $L_{v_0} = \{(v_0, v_1^0, v_2^0), \dots, (v_0, v_1^{d-1}, v_2^{d-1})\}$ of *valid* oriented edges whose first vertex is v_0 . To realize the peeling algorithm, it is sufficient to implement the following operations on the lists.

- $\text{Degree}(L_{v_0})$ returns the number of edges d in the incidence list of v_0 ;
- $\text{AddEdge}(L_{v_0}, e)$ adds the edge e to the incidence list of v_0 ;
- $\text{DeleteEdge}(L_{v_0}, e)$ deletes the edge e from the incidence list of v_0 ;
- $\text{RetrieveEdge}(L_{v_0})$ returns the only edge in the list if $\text{Degree}(L_{v_0}) = 1$.

For all the operations above, it is assumed that the edge is given through a *valid* orientation. Under this set of operations, the data structure does not need to *store* the actual list of edges: it is sufficient to store a tuple $(v_0, d, \tilde{v}_1, \tilde{v}_2)$, where d is the number of edges, $\tilde{v}_1 = \bigoplus_{j < d} v_1^j$, and $\tilde{v}_2 = \bigoplus_{j < d} v_2^j$, that is, all the vertices of the list in the same position are XORed together.

The operations *AddEdge* and *DeleteEdge* on an edge (v_0, v_1', v_2') simply XOR v_1' into \tilde{v}_1 and v_2' into \tilde{v}_2 , and respectively increment or decrement d . Since all the edges are assumed valid (i.e., it holds that $v_1' < v_2'$) these operations maintain the invariant. When $d = 1$, clearly $\tilde{v}_1 = v_1$ and $\tilde{v}_2 = v_2$ where (v_0, v_1, v_2) is the only edge in L_{v_0} , so it can be returned by *RetrieveEdge*. If necessary, the data structure can be trivially extended to *labeled* edges (v_0, v_1, v_2, ℓ) by XORing together the labels ℓ into a new field $\tilde{\ell}$.

We call this data structure *packed incidence list*, and we refer to this technique as the *XOR trick*. The advantage with respect to maintaining an explicit list, besides the obvious space savings, is that it is sufficient to maintain a single fixed-size record per vertex, regardless of the number of incident edges. This will make the peeling algorithm in the next section substantially simpler and faster. The same trick can be applied to the standard linear-time algorithm, replacing the linked lists traditionally used. As we will see in Section 5, the improvements are significant in both working space and running time.

4.2 Layered peeling

The peeling procedure we present is an adaptation of the CORE procedure presented by Molloy [25]. The basic idea is to proceed in rounds: at each round, all the vertices of degree 1 are removed, and then the next round is performed on the induced subgraph, until either a 2-core is left, or the graph is empty. In the latter case, the algorithm partitions the edges into a sequence of *layers*, one per round, by defining

each layer as the set of edges removed in its round. It is easy to see that by concatenating the layers the resulting edge order is a peeling order, regardless of the order within each layer.

The layered peeling process terminates in a small number of rounds: Jiang et al. [18] proved that if the hypergraph is generated randomly with a sparsity above the peeling threshold, then with high probability the number of rounds is bounded by $O(\log \log n)$. Moreover, the fraction of vertices remaining in each round decreases double-exponentially. In the following we show how to implement the algorithm in an I/O-efficient way by putting special care in the hypergraph representation and the update step.

Hypergraph representation At each round i , the hypergraph is represented by a list E_i of tuples $(v_0, d, \tilde{v}_1, \tilde{v}_2)$ as described in Section 4.1; each tuple represents the incidence list of v_0 . Each list E_i is sorted by v_0 . Note that each edge $e = \{v_0, v_1, v_2\}$ needs to be in the incidence list of all its vertices; hence, all the three orientations of e are present in the list E_i .

Construction of E_0 To construct E_0 , the edge list for the first round, we put together in a list all the valid orientations of all the edges in the hypergraph. The list is then sorted by v_0 , and from the sorted list we can construct the sorted list of incidence lists E_0 : after grouping the oriented edges by v_0 , we start with the empty packed incidence list $(v_0, 0, 0, 0)$ and, after performing AddEdge with all the edges in the group, we append it to E_0 . The I/O complexity is $O(\text{sort}(n) + \text{scan}(n)) = O(\text{sort}(n))$.

Round update At the beginning of each round we are given the list E_i of edges that are alive at round i , and we produce E_{i+1} . We first scan E_i to find all the tuples L such that $\text{Degree}(L) = 1$; for each tuple, we perform RetrieveEdge and put the edge in a list D_i , which represents all the edges to be removed in the current round i . The same edge may occur multiple times in D_i under different orientations (if more than one of its vertices have degree 1 in the current round); to remove the duplicates, we sort the oriented edges by their canonical orientation, keep one orientation for each edge, and store them in a list P_i .

Now we need to remove the edges from the hypergraph. To do so, we generate a *degree update list* U_i that contains all the three orientations for each edge in P_i , and sort U_i by v_0 . Since both E_i and U_i are sorted by v_0 , we can scan them both simultaneously joining them by v_0 ; for each tuple L_{v_0} in E_i , if no oriented edge starting with v_0 is in U_i the tuple is copied to E_{i+1} , otherwise for each such oriented edge e , DeleteEdge(L_{v_0}, e) is called to obtain a new L'_{v_0} which is written to E_{i+1} if non-empty. Note that E_{i+1} remains sorted by v_0 .

For each round, we scan E_i twice and U_i once, and sort D_i and U_i . The number of I/Os is then $2 \cdot \text{scan}(|E_i|) + \text{scan}(|U_i|) + \text{sort}(|D_i|) + \text{sort}(|U_i|)$. Summing over all rounds, we have $\sum_i (2 \cdot \text{scan}(|E_i|) + \text{sort}(|D_i|) + \text{sort}(|U_i|) + \text{scan}(|U_i|)) = O(\text{sort}(n))$ because each edge belongs to at most three lists D_i and three lists U_i . Since the fraction of vertices remaining at each round decreases doubly exponentially and, thanks to the XOR trick, E_i has exactly a tuple for each vertex alive in the i -th round, the cost of scanning the lists E_i sums up to $O(\text{scan}(n))$ I/Os. Hence, overall the algorithm takes $O(n \log n)$ time and $O(\text{sort}(n))$ I/Os.

We summarize the result in the following theorem.

THEOREM 4.1 A peeling order of a random r -hypergraph with n edges and γn vertices with constant r and $\gamma > c_r$, can be computed in the cache-oblivious model in time $O(n \log n)$ and with $O(\text{sort}(n))$ I/Os with high probability.

4.3 Implementation details

We report here the most important optimizations we used in our implementation. The source code used in the experiments is available at <https://github.com/ot/empfh> for the reader interested in further implementation details and in replicating the measurements.

File I/O Instead of managing file I/Os directly, we use a memory-mapped file by employing a C++ allocator that creates a file-backed area of memory. This way we can use the standard STL containers such as `std::vector` as if they resided in internal memory. We use `madvise` to instruct the kernel to optimize the mapped region for sequential access. We use the `madvise` system call with the parameter `MADV_SEQUENTIAL` on the memory-mapped region to instruct the kernel to optimize for sequential access.

Sorting Our sorting implementation performs two steps: in the first step we divide the domain of the values into k evenly spaced buckets, scanning the array to find the number of values that belong in each bucket, and then moving each value to its own bucket. In the second step, each bucket is sorted using `sort` of the C++ standard library. The number of buckets is chosen so that with very high probability

each bucket fits in internal memory; since the graph is random, its edges are uniformly distributed, which makes uniform bucketing balanced with high probability. To distribute the values into the k buckets, we use a buffer of size T for each bucket; when the buffer is full, it is flushed to disk. Note that this algorithm is technically not cache-oblivious, since it works as long as the available memory M is at least kT ; choosing k to be $\Theta(S/M)$, where S is the size of the data to be sorted, requires that M be $\Omega(\sqrt{TS})$. In our implementation we use $T \approx 1\text{MiB}$, thus for example $M = 1\text{GiB}$ is sufficient to sort $\approx 1\text{TiB}$ of data. When this condition holds, the algorithm performs just three scans of the array and it is extremely efficient in practice. Furthermore, contrary to existing cache-oblivious sorting implementations, it is *in-place*, using no extra disk space.

Reusing memory The algorithm as described in Section 4.2 uses a different list E_i for each round. Since tuples are appended to E_{i+1} at a slower pace than they are read from E_i , we can reuse the same array. A similar trick can be applied to D_i and U_i . Overall, we need to allocate just one array of γn packed incidence lists, and one for the $3n$ oriented edges.

Lists compression Reducing the size of the on-disk data structures can significantly improve I/O efficiency, and hence the running time of the algorithm. The two data structures that take nearly all the space are the lists of packed incidence lists E_i and the lists of edges P_i . Since the lists are read and written sequentially, we can (*de*)compress them on the fly.

Recall that the elements of E_i are tuples of the form $(v_0, d, \tilde{v}_1, \tilde{v}_2)$ sorted by v_0 . The first components v_0 of these tuples are gap-encoded with Elias γ codes. The overall size of the encoding is $\sum_{k=1}^{|E_i|} (2\lfloor \log g_k \rfloor + 1)$ bits, where g_k is the k -th gap.³ Since the gaps sum up to γn , by Jensen's inequality the sum is maximized when the gaps g_k are all equal to $\frac{\gamma n}{|E_i|}$ giving a space bound of $2|E_i|(\log \frac{\gamma n}{|E_i|} + 1)$ bits. Furthermore, this space bound is always at most $2\gamma n$ bits because it is maximized when E_i has size $\gamma n/2$. The degrees d are encoded instead with unary codes; since the sum of the degrees is $3n_i$, where n_i is the number of edges alive in round i , the overall size of their encoding is always upper bounded by $3n$ bits. The other two components, as well as the nodes in P_i , are represented with a fixed-length encoding using $\lceil \log \gamma n \rceil$ bits each. With $\gamma = 1.23$, and thanks to the memory reusing technique described above, the overall disk usage is approximately $(5.46 + 11.46\lceil \log \gamma n \rceil)n$ bits. On our largest inputs, using compression instead of plain 64-bit words makes the overall algorithm run about 2.5 times faster.

Exploiting the tripartition Many MWHC-based implementations, when generating the r -hypergraph edges $\{h_0(x), \dots, h_{r-1}(x)\}$, use random hash functions h_i with codomain $[i|V|/r, (i+1)|V|/r)$ instead of $[0, |V|)$, thus yielding a r -partite r -hypergraph. The main advantage is that by construction $h_i(x) \neq h_j(x)$ for $i \neq j$, so the process cannot generate hypergraphs with degenerate edges; this reduces considerably the number of trials needed to find a peelable hypergraph (in practice, just one trial is sufficient). Botelho et al. [7] proved that hypergraphs obtained with this process have the same peeling threshold as uniformly random hypergraphs. Jiang et al. [18] proved that the bound on the number of rounds of the layered peeling process also holds for random r -partite r -hypergraphs, so we can adopt this approach as well.

An additional advantage of the r -partition is that the first vertex of any 0-orientation is smaller than the first vertex of any 1-orientation, and so on; in general, if (u_0, \dots, u_{r-1}) is an i -orientation, (v_0, \dots, v_{r-1}) is a j -orientation, and $i < j$, then $u_0 < v_0$. We exploit this in our algorithm in the construction of E_0 : since our graph is 3-partite, instead of creating a list with every valid orientation of each edge and then sorting it by v_0 , we create a list with just the 0-orientations, sort it by v_0 , and append the obtained packed incidence lists to E_0 . Then we go through the sorted list, switch all the oriented edges to 1-orientation, and repeat the process. The same is done for the 2-orientations.

Thanks to this optimization the amount of memory required in the first step of the algorithm, which is the most I/O intensive, is reduced to one third.

Avoiding backward scans For MWHC-based functions construction, the final phase that assigns the u_i 's needs to scan the edges in *reverse* peeling order. Unfortunately, operating systems and disks are highly optimized for *forward* reading, by performing an aggressive lookahead. However, as we noted in Section 4.2, the ordering of the edges *within* the layers is irrelevant; thus it is sufficient to scan the layers in reverse order, but each layer may be safely scanned forward. The number of forward scans is then bounded by the number of rounds, which is negligible. The performance improvement of the assignment phase with respect to reading the array backwards is almost ten-fold.

³Elias Gamma code [13] uses $2\lceil \log j \rceil + 1$ bits to encode any integer $j \geq 1$.

5 Experimental analysis

Although our code can be easily extended to construct *any* static function, to evaluate experimentally the performance of the peeling algorithm we tested it on the task of constructing a minimal perfect hash function, as discussed in Section 3. In this task, the peeling process largely dominates the running time.

Testing details The tests of MPHF construction were performed on an Intel Xeon i7 E5520 (Nehalem) at 2.27GHz with 32GiB of RAM, running Linux 3.5.0 x86-64. The storage device is a 3TB Western Digital WD30EFRX hard drive. Before running each test, the kernel page cache was cleared to ensure that all the data were read from disk. The experiments were written in C++11 and compiled with g++ 4.8.1 at -O3.

We tested the following algorithms.

- **Cache-Oblivious:** The cache-oblivious algorithm described in Section 4.
- **Standard+XOR:** The standard linear-time peeling implemented using the packed incidence list, with the purpose of evaluating the impact of the XOR-trick by itself.
- **cmph:** A publicly available, widely used and optimized library for minimal perfect hashing⁴, implementing the same MWHC-based MPHF construction with the standard in-memory peeling algorithm.

Datasets We tested the above algorithms on the following datasets.

- **URLs:** a set of ≈ 4.8 billion URLs from the ClueWeb09 dataset⁵ (average string length ≈ 67 bytes, summing up to ≈ 304 GiB);
- **ngrams:** a set of ≈ 7.6 billion $\{1, 2, 3\}$ -grams obtained from the Google Books Ngrams English dataset⁶ (average string length ≈ 23 bytes, summing up to ≈ 168 GiB).

Since the strings are hashed in the first place, the nature of the data is fairly irrelevant: the only aspect that may be relevant is the average string length (that affects the time to load the input from disk). In fact tests on randomly generated data produced the same results.

Experimental results The running time of the algorithms as the number of keys increases is plotted in Figure 1; to evaluate the performance in the regime where the working space fits in main memory, the figure also shows an enlarged version of the first part of the plot.

The first interesting observation is that the cache-oblivious algorithm performs almost as well as cmph, with Cache-Oblivious being slightly slower because it has to perform file I/O even when the working space would fit in memory.

We can also see that the XOR trick pays off, as shown by the performance of Standard+XOR, which is up to 3 times faster than cmph, and the smaller space usage enables to process up to almost twice the number of keys for the given memory budget. Both non-external algorithms, though, cease to be useful as soon as the available memory gets exhausted: the machine, then, starts to thrash because of the random patterns of access to the swap. In fact, we had to kill the processes after 48 hours. Actually, one can make a quite precise estimate of when this is going to happen: cmph occupies 34.62 bytes/key, as estimated by the authors, whereas Standard+XOR occupies about 26.76 bytes/key, and these figures almost exactly justify the two points where the construction times slow down and then explode. On the other hand, Cache-Oblivious scales well with the input size, exhibiting eventually almost linear performance in our larger input ngrams, while remaining competitive even on small key sets.

Comparison with HEM Finally, we compare our algorithm with HEM [6]. Recall that their technique consists in splitting the set of keys into several buckets and building a separate MPHF for every bucket; at query time, a first-level index is used to drive the query to the correct bucket. Choosing a sufficiently small size for the buckets allows the use of a standard internal memory algorithm to construct the bucket MPHFs. Although technically not a peeling algorithm, this external-memory solution is simple and elegant.

To make a fair comparison, we re-implemented the HEM algorithm using our sort implementation for the initial bucketing, and the Standard+XOR algorithm to build the bucket MPHFs. The signature

⁴We used cmph 2.0, available at <http://cmph.sourceforge.net/>.

⁵Downloaded from <http://lemurproject.org/clueweb09/>.

⁶Downloaded from <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>.

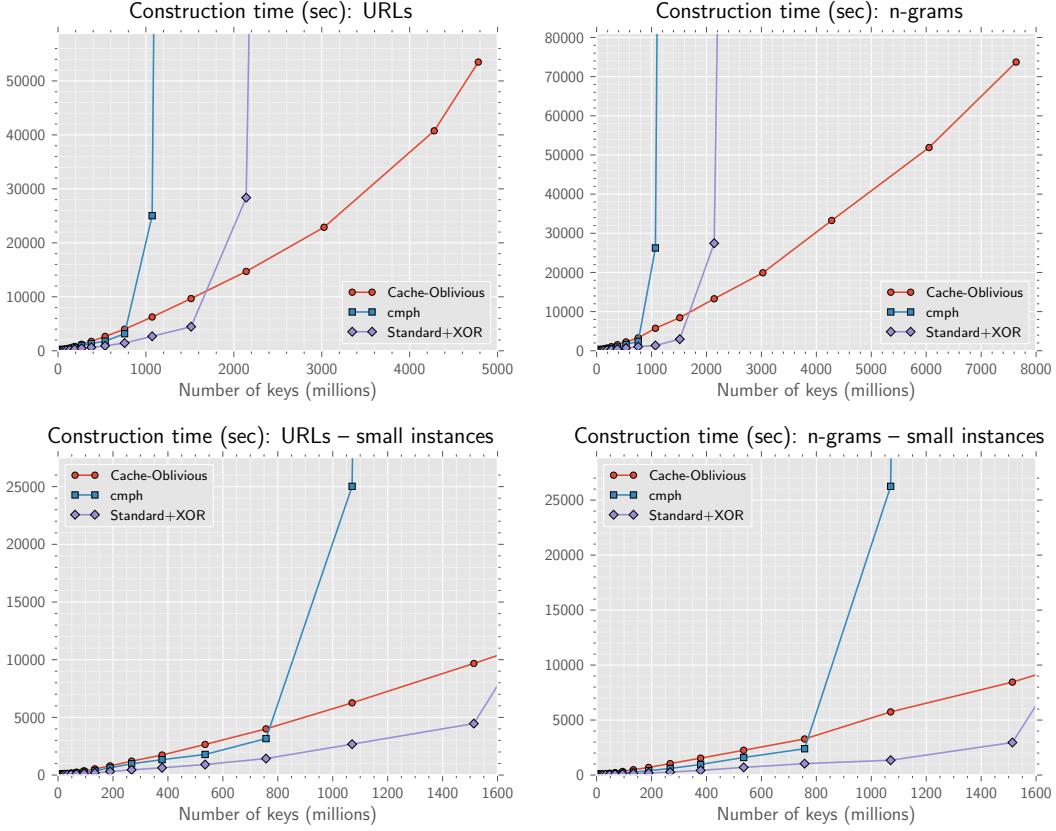


Figure 1: Above: construction times on the two datasets. Below: close-up for n up to $1.6 \cdot 10^9$ keys.

function is the same 96-bit hash function used in [6] (which suffice for sets of up to 2^{48} keys), but we employed 64-bit bucket offsets in place of 32-bit, since our key sets are larger than $2^{32}/\gamma$.

The result, as shown in Figure 2, is a construction time between 2 and 6 times smaller than Cache-Oblivious. However, this efficiency has a cost in term of lookup time (because of the double indirection) and size (because of the extra space needed for the first-level index). Since, in most applications, MPHFs are built rarely and queried frequently, the shorter construction time may not be worth the increase in space and query time.

Indeed, as shown in Table 1, the space loss is 17% to 27%. The variability in space overhead is due to the fact that in HEM the number of buckets must be a power of 2, hence the actual average bucket size can vary by a factor of 2 depending on the number of keys. We also include the space taken by cmph on the largest inputs we were able to construct in-memory. Despite using the same data structure as our implementation of MWHC, its space occupancy is slightly larger because it uses denser ranking tables.

	URLs		ngrams	
	$0.76 \cdot 10^9$ keys	$4.8 \cdot 10^9$ keys	$0.76 \cdot 10^9$ keys	$7.6 \cdot 10^9$ keys
MWHC	2.61 b/key	2.61 b/key	2.61 b/key	2.61 b/key
HEM	3.16 b/key	3.31 b/key	3.16 b/key	3.05 b/key
cmph	2.77 b/key	-	2.77 b/key	-

Table 1: Space comparison of MWCH, HEM, and cmph.

The evaluation of lookup efficiency is much subtler, as it depends on a number of factors, some of which are subject to hardware architecture. For this reason, we decided to perform the experiments on three different machines: an Intel Intel i7-4770 (Haswell) at 3.40GHz, the same Intel i7 (Nehalem) machine used for the construction experiments (see above), and an AMD Opteron 6276 at 2.3GHz.

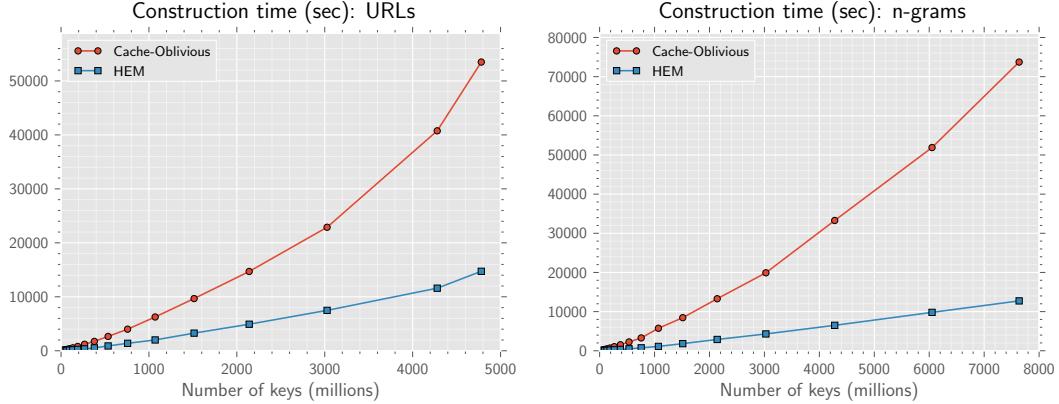


Figure 2: Construction time with the Cache-Oblivious algorithm and HEM [6].

For both machines and both datasets we performed lookups of 10 M distinct keys, repeated 10 times. Since lookup times are in the order of less than a microsecond, it is impossible to measure individual lookups accurately; for this reason, we divided the lookups into 1,525 batches of 2^{16} keys each, and measured the average lookup time for each batch. Out of these average times, we computed the global average and the standard deviation. The results in Table 2 show that HEM is slower than MWHC in all cases. On AMD Opteron the slowdown is the smallest, ranging from 17% to 20%; on the Intel i7 (Nehalem) the range goes up to 19%–26%; on the Intel i7 (Haswell), the most recent and fastest CPU, the slowdown goes up to 30%–35%, suggesting that as the speed of the CPU increases, the cost of the causal cache miss caused by the double indirection of HEM becomes more substantial. In all cases, the standard deviation is negligibly small, making the comparison statistically significant.

We also remark that our implementation of the MWHC lookup (which is used also in HEM) is roughly twice as fast than `cmph` despite using a sparser ranking table; this is because to perform the ranking we adopt a *broadword* [19] algorithm that counts the number of non-zero pairs in a 64-bit words in just a few non-branching instructions, rather than a linear bit scan with a loop; the smaller ranking table also imposes a lower cache pressure. Finally, we use a 64-bit implementation of the Jenkins hash function, which is faster on long strings than the 32-bit one used in `cmph`.

	URLs		ngrams	
	$0.76 \cdot 10^9$ keys	$4.8 \cdot 10^9$ keys	$0.76 \cdot 10^9$ keys	$7.6 \cdot 10^9$ keys
Intel i7 (Haswell)				
MWHC	219 ns \pm 0.3%	253 ns \pm 1.3%	199 ns \pm 0.2%	251 ns \pm 1.8%
HEM	284 ns \pm 0.3%	335 ns \pm 1.1%	262 ns \pm 0.3%	338 ns \pm 0.9%
cmph	466 ns \pm 0.3%	-	303 ns \pm 0.3%	-
Intel i7 (Nehalem)				
MWHC	365 ns \pm 0.1%	433 ns \pm 0.1%	334 ns \pm 0.1%	422 ns \pm 0.2%
HEM	450 ns \pm 0.1%	523 ns \pm 0.1%	420 ns \pm 0.1%	502 ns \pm 0.7%
cmph	799 ns \pm 0.1%	-	532 ns \pm 0.1%	-
AMD Opteron				
MWHC	415 ns \pm 0.1%	419 ns \pm 0.1%	373 ns \pm 0.1%	386 ns \pm 0.1%
HEM	484 ns \pm 0.1%	493 ns \pm 0.1%	442 ns \pm 0.2%	463 ns \pm 0.1%
cmph	908 ns \pm 0.2%	-	578 ns \pm 0.3%	-

Table 2: Lookup-time comparison (with relative standard deviation) of MWHC, HEM, and `cmph`.

References

- [1] Djamal Belazzougui, Paolo Boldi, Rasmus Pagh, and Sebastiano Vigna. Monotone minimal perfect hashing: Searching a sorted table with $O(1)$ accesses. In *SODA*, pages 785–794, 2009.
- [2] Djamal Belazzougui, Paolo Boldi, Rasmus Pagh, and Sebastiano Vigna. Fast prefix search in little space, with applications. In *ESA* (1), pages 427–438, 2010.
- [3] Djamal Belazzougui, Paolo Boldi, Rasmus Pagh, and Sebastiano Vigna. Theory and practice of monotone minimal perfect hashing. *ACM Journal of Experimental Algorithms*, 16, 2011.
- [4] Djamal Belazzougui and Gonzalo Navarro. Alphabet-independent compressed text indexing. In *ESA* (1), pages 748–759, 2011.
- [5] Djamal Belazzougui and Rossano Venturini. Compressed static functions with applications. In *SODA*, pages 229–240, 2013.
- [6] Fabiano C. Botelho, Rasmus Pagh, and Nivio Ziviani. Practical perfect hashing in nearly optimal space. *Inf. Syst.*, 38(1):108–131, 2013.
- [7] Fabiano C. Botelho, Nicholas Wormald, and Nivio Ziviani. Cores of random r -partite hypergraphs. *Information Processing Letters*, 112(8–9):314 – 319, 2012.
- [8] Gerth Stølting Brodal and Rolf Fagerberg. On the limits of cache-obliviousness. In *STOC*, pages 307–315, 2003.
- [9] Denis Charles and Kumar Chellapilla. Bloomier filters: A second look. In *ESA*, pages 259–270, 2008.
- [10] Bernard Chazelle, Joe Kilian, Ronitt Rubinfeld, and Ayelet Tal. The Bloomier filter: an efficient data structure for static support lookup tables. In *SODA*, pages 30–39, 2004.
- [11] Zbigniew J. Czech, George Havas, and Bohdan S. Majewski. Perfect hashing. *Theor. Comput. Sci.*, 182(1-2):1–143, 1997.
- [12] Martin Dietzfelbinger, Andreas Goerdt, Michael Mitzenmacher, Andrea Montanari, Rasmus Pagh, and Michael Rink. Tight thresholds for cuckoo hashing via xorsat. In *ICALP* (1), pages 213–225, 2010.
- [13] Peter Elias. Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory*, 21(2):194–203, 1975.
- [14] Matteo Frigo, Charles E. Leiserson, Harald Prokop, and Sridhar Ramachandran. Cache-oblivious algorithms. *ACM Transactions on Algorithms*, 8(1):4, 2012.
- [15] Michael T. Goodrich and Michael Mitzenmacher. Invertible Bloom lookup tables. In *CCC*, pages 792–799, 2011.
- [16] Torben Hagerup and Torsten Tholey. Efficient minimal perfect hashing in nearly minimal space. In *STACS*, pages 317–326. 2001.
- [17] G. Jacobson. Space-efficient static trees and graphs. In *FOCS*, pages 549–554, 1989.
- [18] Jiayang Jiang, Michael Mitzenmacher, and Justin Thaler. Parallel peeling algorithms. *CoRR*, abs/1302.7014, 2013.
- [19] Donald E. Knuth. The Art of Computer Programming. Pre-Fascicle 1A. Draft of Section 7.1.3: Bitwise Tricks and Techniques, 2007.
- [20] Michael Luby, Michael Mitzenmacher, Mohammad Amin Shokrollahi, and Daniel A. Spielman. Efficient erasure correcting codes. *IEEE Transactions on Information Theory*, 47(2):569–584, 2001.
- [21] Bohdan S. Majewski, Nicholas C. Wormald, George Havas, and Zbigniew J. Czech. A family of perfect hashing methods. *Comput. J.*, 39(6):547–554, 1996.

- [22] Kurt Mehlhorn. On the program size of perfect and universal hash functions. In *FOCS*, pages 170–175, 1982.
- [23] Michael Mitzenmacher and George Varghese. Biff (Bloom filter) codes: Fast error correction for large data sets. In *ISIT*, pages 483–487, 2012.
- [24] Michael Molloy. The pure literal rule threshold and cores in random hypergraphs. In *SODA*, pages 672–681, 2004.
- [25] Michael Molloy. Cores in random hypergraphs and Boolean formulas. *Random Structures and Algorithms*, 27(1):124–135, 2005.

A Network Model characterized by a Latent Attribute Structure with Competition

Paolo Boldi*, Irene Crimaldi†, Corrado Monti‡

July 30, 2014

Abstract

The quest for a model that is able to explain, describe, analyze and simulate real-world complex networks is of uttermost practical as well as theoretical interest. In this paper we introduce and study a network model that is based on a latent attribute structure: each node is characterized by a number of features and the probability of the existence of an edge between two nodes depends on the features they share. Features are chosen according to a process of Indian-Buffet type but with an additional random “fitness” parameter attached to each node, that determines its ability to transmit its own features to other nodes. As a consequence, a node’s connectivity does not depend on its age alone, so also “young” nodes are able to compete and succeed in acquiring links. One of the advantages of our model for the latent bipartite “node-attribute” network is that it depends on few parameters with a straightforward interpretation. We provide some theoretical, as well experimental, results regarding the power-law behavior of the model and the estimation of the parameters. By experimental data, we also show how the proposed model for the attribute structure naturally captures most local and global properties (e.g., degree distributions, connectivity and distance distributions) real networks exhibit.

keyword: Complex network, social network, attribute matrix, Indian Buffet process

1 Introduction

Complex networks are a unifying theme that emerged in the last decades as one of the most important topics in many areas of science; the starting point is the

*Dipartimento di Informatica, Università degli Studi di Milano, Via Comelico 39/41 - 20135
Milano, Italy

†IMT Institute for Advanced Studies Lucca, Piazza San Ponziano 6, I-55100 Lucca, Italy

‡Dipartimento di Informatica, Università degli Studi di Milano, Via Comelico 39/41 - 20135
Milano, Italy

observation that many networks arising from different types of interactions (e.g., in biology, physics, chemistry, economics, technology, on-line social activity) exhibit surprising similarities that are partly still unexplained. The quest for a model that is able to explain, describe, analyze and simulate those real-world complex networks is of utmost practical as well as theoretical interest.

The classical probabilistic model of graphs by Erdős and Rényi [11] soon revealed itself unfit to describe complex networks because, for example, it fails to produce a power-law degree distribution. One of the first attempts to try to obtain more realistic models was [3], where the idea of *preferential attachment* was first introduced: nodes tend to attach themselves more easily to other nodes that are already very popular, i.e. with a high number of links. Similar models were proposed by [1] and [22]. The general approach of these and other attempts is to produce probabilistic frameworks (typically with one or more parameters) giving rise to networks with statistical properties that are compatible with the ones that are observed in real-world graphs: degree distribution is just one example; other properties are degree-degree correlation, clustering coefficients, distance distribution etc. [19].

The task of modeling the network is often undertaken directly [3, 23], but recently some authors proposed to split it into two steps (see, e.g., [24, 25]). This proposal stems from the observation that many complex networks contain two types of entities: actors on one hand, and groups (or features) on the other; every actor belongs to one, or more, groups (or can exhibit one, or more, features), and the common membership to groups (or the sharing of features) determines a relation between actors. The idea of a *bipartite network*, where interpersonal connections follow from intergroup connections, derives from sociology; a seminal paper presented by Breiger [7] in 1974 described this dualism between “persons and groups”. This idea has been proved precious in social networks and their mathematical modelization [25].

In particular, many authors [26] distinguish between two kinds of models: class-based models – such as [31] – assume that every node belongs to a single class, while feature-based models use many features to describe each node. A well-known shortcoming of the first is the proliferation of classes, since dividing a class according to a new feature leads to two different classes. To overcome this limitation, classical class-based models have been extended to allow mixed membership, like in [2]. Feature-based models naturally assume this possibility. Within them, some authors (such as [18]) propose real-valued vectors to associate features to nodes; others instead assume only binary features, in which a node either exhibits a feature or it does not (see e.g. [26]). This assumption is simple and natural, and it significantly simplifies the analysis of the model.

A natural model for the evolution of such binary bipartite graphs comes from Bayesian statistics and it is known as the Indian Buffet process, introduced by Griffiths *et al.* [14, 15, 16] and, subsequently extended and studied by many authors [8, 33, 34]. The process defines a plausible way for features to evolve, always according to a *rich-get-richer* principle: because of this, it represents a promising model for affiliation networks. Since the Indian Buffet process provides *a prior distribution* in

Bayesian statistics, these models have been used to reconstruct affiliation networks with an unknown number of features from data where only friendship relations between actors are available. An important work in this direction is [26]. However, the standard Indian Buffet process has a drawback as a model for real networks: the exchangeability assumption is often untenable in applications.

In this paper we propose and analyze a model that combines two features characterizing the evolution of a network:

1. Behind the adjacency matrix of a network there is a *latent attribute structure* of the nodes, in the sense that each node is characterized by a number of features and the probability of the existence of an edge between two nodes depends on the features they share. In other words, the adjacency matrix of a network hides a bipartite network describing the attributes of the nodes.
2. Not all nodes are equally successful in transmitting their own attributes to the new nodes. Each node n is characterized by a *random fitness parameter* R_n describing its ability to transmit the node's attributes: the greater the value of the random variable R_n , the greater the probability that a feature of n will also be a feature of a new node, and so the greater the probability of the creation of an edge between n and the new node. Consequently, a node's connectivity does not depend on its age alone (so also “young” nodes are able to compete and succeed in acquiring links). We refer to this aspect as *competition*.

We shape the first aspect by the definition of a model which connects the pair of attribute-vectors of two nodes, say i and j , to the probability of the existence of an edge between i and j . Other examples can be found in [26, 27, 28, 30], which are related to the Bayesian framework based on the standard Indian Buffet model.

We model the second aspect by the definition of a stochastic dynamics for a bipartite “node-attribute” network, where the probability that a new node exhibits a certain attribute depends on the ability, represented by some random fitness parameter, of the previous nodes possessing that attribute in transmitting it. It is worthwhile to underline that, as in the standard Indian Buffet process, the collection of attributes is potentially unbounded. Thus, we do not need to specify a maximum number of latent attributes *a priori*.

We were inspired by the recent generalization of the Indian buffet process presented in [5]. However, the model presented here is in some sense simpler since the parameters (that will be introduced and analyzed in the next sections) play a role that is clearer and more intuitive. Specifically, we have two parameters (α and β) that control the number of new attributes each new node exhibits (in particular $\beta > 0$ tunes the power-law behaviour of the overall number of different observed attributes), whereas the random fitness parameters R_i impact on the probability of the new nodes to inherit the attributes of the previous nodes. With respect to the model in [5], we lose some mathematical properties, but we will show that some important results still hold true and they allow us to estimate the parameters and, in particular, the exponent of the power-law behavior.

Regarding the use of fitness parameters, we recall the work by Bianconi and Barabási [6] that introduced some fitness parameters describing the ability of the nodes to compete for links. The difference between their model and ours consists in the fact that in [6] the fitness parameters appear explicitly in the edge-probabilities; while in our model they affect the evolution of the attribute matrix and then play an implicit role in the evolution of the connections.

Summing up, the present work have different aims: firstly, we propose a simple model for the latent bipartite “node-attribute” network, where the role played by the single parameters is straightforward and easy to be interpreted; secondly, differently from other network models based on the standard Indian Buffet process, we take into account the aspect of competition and, like in [5], we introduce random fitness parameters so that nodes have a different relevance in transmitting their features to the next nodes; finally, we provide some theoretical, as well experimental, results regarding the power-law behavior of the model and the estimation of the parameters. By experimental data, we will also show how the proposed attribute structure naturally leads to a complex network model.

The paper is structured as follows. In Section 2, we introduce a model for the evolution of the attribute matrix and we provide theoretical results and tools regarding the estimation and the analysis of the quantities characterizing the model. These methods are then tested by simulations in Section 3. In order to produce a graph out of the attribute structure, in Section 4 we illustrate different models for the edge-probabilities that are based on the attribute matrix. The properties of the generated graphs are studied by simulation in Section 5. Finally, in Section 6 we analyze a real dataset and, then, in Section 7 we sum up the main novelties and merits of our work and we illustrate some possible future lines of research.

2 A model for the evolution of the attribute matrix

We assume that the nodes enter the network sequentially so that node i represents the node that comes into the network at time i . Let \mathcal{X} be an unbounded collection of possible attributes that a node can exhibit. (This means that we do not specify the total number of possible attributes *a priori*.) Each node is assumed to have only a finite number of attributes and different nodes can share one or more attributes.

Let Z be a binary bipartite network where each row Z_n represents the attributes of the node n : $Z_{n,k} = 1$ if node n has attribute k , $Z_{n,k} = 0$ otherwise. We assume that each Z_n remains unchanged in time, in the sense that every node decides its own features/attributes when it arrives and then it will never change them thereafter. This assumption is quite natural in many contexts, e.g., in genetics.

In all the sequel we postulate that Z is left-ordered. This means that in the first row the columns for which $Z_{1,k} = 1$ are grouped on the left and so, if the first node has N_1 features, then the columns of Z with index $k \in \{1, \dots, N_1\}$ represent these

features. The second node could have some features in common with the first node (those corresponding to indices k such that $k = 1, \dots, N_1$ and $Z_{2,k} = 1$) and some, say N_2 , new features. The latter are grouped on the right of the sets for which $Z_{1,k} = 1$, i.e., the columns of Z with index $k \in \{N_1 + 1, \dots, N_2\}$ represent the new features brought by the second node. This grouping structure persists throughout the matrix Z .

Here is an example of a Z matrix with $n = 4$ nodes; in gray we show the new features adopted by each node ($N_1 = 3$, $N_2 = 2$, $N_3 = 3$, $N_4 = 2$ in this example); observe that, for every node i , the i -th row contains 1's for all the columns with indices $k \in \{N_1 + \dots + N_{i-1} + 1, \dots, N_1 + \dots + N_i\}$ (they represent the new features brought by i); moreover some of the columns with indices $k \in \{1, \dots, N_1 + \dots + N_{i-1}\}$ are also 1's (features that were chosen by previous nodes and that also node i decided to adopt):

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

We will describe the dynamics using a culinary metaphor (similarly to what some authors do for other models, see Chinese Restaurant [29], Indian Buffet process [15, 16, 33] and their generalizations [4, 5]). We identify the nodes with the customers of a restaurant and the attributes with the dishes, so that the dishes tried by a customer represent the attributes that a node exhibits.

Fix $\alpha > 0$ and $\beta \in (-\infty, 1]$. Also, let $\text{Poi}(\lambda)$ denote the Poisson distribution with mean $\lambda \geq 0$ (where $\text{Poi}(0) = \delta_0$). Customer (node) n is attached a random weight (that we call, in accordance with the usage in Network Theory, *fitness parameter*) R_n . We assume that each R_n is independent of R_1, \dots, R_{n-1} and of the dishes experimented by customers $1, \dots, n$. The fitness parameter R_n affects the choices of the future customers (those after n), while the choices of customer n are affected by the fitness parameters and the choices of the previous ones. Indeed, it may be the case that different customers have different relevance, due to some random cause, that does not affect their choices but is relevant to the choices of future customers (i.e., their capacity of being followed).

The dynamics is as follows. Customer (node) 1 tries N_1 dishes, where N_1 is $\text{Poi}(\alpha)$ -distributed. For each $n \geq 1$, let S_n be the collection of dishes experimented by the first n customers (nodes). For the customers (nodes) following the first one, we have that:

- Customer $n + 1$ selects a subset S_n^* of S_n . Each $k \in S_n$ is included or not into S_n^* independently of the other members of S_n . The inclusion probability is

$$P_n(k) = \frac{\sum_{i=1}^n R_i Z_{i,k}}{\sum_{i=1}^n R_i}. \quad (2.1)$$

where $Z_{i,k} = 1$ if $\{ \text{customer } i \text{ has selected dish } k \}$ and $Z_{i,k} = 0$ otherwise. It is a preferential attachment rule: the larger the weight of a dish k at time n (given by the numerator of (2.1), i.e., the total value of the random variables R_i associated to the customers that have chosen it until time n), the greater the probability that it will be chosen by the future customer $n+1$.

- In addition to S_n^* , customer $n+1$ also tries N_{n+1} new dishes, where N_{n+1} is $\text{Poi}(\Lambda_n)$ -distributed with

$$\Lambda_n = \frac{\alpha}{(\sum_{i=1}^n R_i)^{1-\beta}}. \quad (2.2)$$

For each k in S_{n+1} , the matrix element $Z_{n+1,k}$ is set equal to 1 if customer $n+1$ has selected dish k , equal to zero otherwise.

Besides the assumption of independence, we also assume that the random parameters R_n are identically distributed with $R_n \geq v$ for each n and a certain number $v > 0$, and $E[R_n^2] < +\infty$.

We set $E[R_n] = m_R$ and $L_n = \text{card}(S_n) = \sum_{i=1}^n N_i$, i.e.

L_n = overall number of different dishes experimented by the first n customers
= overall number of different observed attributes for the first n nodes.

In the previous example, we have $L_1 = 3, L_2 = 5, L_3 = 8, L_4 = 10$.

The meaning of the parameters is the following. The random fitness parameters R_n controls the probability of transmitting the attributes to the new nodes. The main effect of β is that it regulates the asymptotic behavior of the random variable L_n (see Theorem 2.1). In particular, $\beta > 0$ is the power-law exponent of L_n . The main effect of α is the following: the larger α , the larger the total number of new tried dishes by a customer (and so the larger the total number of 1's in a row of the binary matrix Z). It is worth to note that β fits the asymptotic behaviour of L_n (in particular, the power-law exponent of L_n) and, separately, α fits the number of observed features.

The mathematical formalization of the above model can be performed by means of random measures [21] with atoms corresponding to the tried dishes (observed attributes), similarly to [5, 8, 34]. More precisely, besides the sequence of positive real random variables (R_n) , we can define a sequence of random measures (M_n) , such that each M_{n+1} is, conditionally on the past $(M_i, R_i : i \leq n)$, a Bernoulli random measure with a hazard measure ν_n , having a discrete part related to the points k in S_n and their weights $P_n(k)$ and a diffuse part with total mass equal to Λ_n .

2.1 Theoretical results regarding the estimation of the parameters α and β

In this section we prove some properties regarding the asymptotic behavior of L_n . In particular, the first result shows a logarithmic behavior for $\beta = 0$ and a power-law

behavior for $\beta \in (0, 1]$. These results allow us to define suitable estimators for β and α .

Theorem 2.1. *Using the previous notation and setting $\Lambda_0 = \alpha$, the following statements hold true:*

- a) $\sup_n L_n = L < +\infty$ a.s. for $\beta < 0$;
- b) $L_n/\ln(n) \xrightarrow{a.s.} \alpha/m_R$ for $\beta = 0$;
- c) $L_n/n^\beta \xrightarrow{a.s.} \alpha/(\beta m_R^{1-\beta})$ for $\beta \in (0, 1]$.

Proof. Let us prove assertion a), first. Let \mathcal{F}_i be the natural σ -field associated to the model until time i . Since, conditionally on \mathcal{F}_i , the distribution of N_{i+1} is $\text{Poi}(\Lambda_i)$, we have

$$P(N_{i+1} \geq 1) = E[P(N_{i+1} \geq 1 \mid \mathcal{F}_i)] \leq E[\Lambda_i].$$

Since $R_i \geq v > 0$, we obtain

$$\sum_i P(N_{i+1} \geq 1) \leq \alpha \sum_i \frac{1}{(vi)^{1-\beta}} < +\infty$$

(where the convergence of the series is due to the assumption $\beta < 0$). By the Borel-Cantelli lemma, we conclude that

$$P(N_i > 0 \text{ infinitely often}) = P(N_i \geq 1 \text{ infinitely often}) = 0.$$

Hence, if $\beta < 0$, there is a random index N such that $L_n = L_N$ a.s. for all $n \geq N$, which concludes the proof of a).

The assertion c) is trivial for $\beta = 1$ since, in this case, L_n is the sum of n independent random variables with distribution $\mathcal{P}(\alpha)$ and so, by the classical strong law of large numbers, $L_n/n \xrightarrow{a.s.} \alpha$.

Now, let us prove assertions b) and c) for $\beta \in [0, 1)$. Define

$$\begin{aligned} \lambda(\beta) &= \frac{\alpha}{m_R} \text{ if } \beta = 0 \quad \text{and} \quad \lambda(\beta) = \frac{\alpha}{\beta m_R^{1-\beta}} \text{ if } \beta \in (0, 1), \\ a_n(\beta) &= \log n \text{ if } \beta = 0 \quad \text{and} \quad a_n(\beta) = n^\beta \text{ if } \beta \in (0, 1). \end{aligned}$$

We need to prove that

$$\frac{L_n}{a_n(\beta)} \xrightarrow{a.s.} \lambda(\beta).$$

First, we observe that we can write

$$\frac{\sum_{i=1}^{n-1} \Lambda_i}{a_n(\beta)} = \alpha \frac{\sum_{i=1}^{n-1} i^{\beta-1} (\bar{R}_i)^{\beta-1}}{a_n(\beta)},$$

where, by the strong law of the large numbers,

$$\bar{R}_i = \frac{\sum_{j=1}^i R_j}{i} \xrightarrow{a.s.} m_R.$$

Therefore, since $\sum_{i=1}^{n-1} i^{\beta-1}/a_n(\beta)$ converges to 1 when $\beta = 0$ and to $1/\beta$ when $\beta \in (0, 1)$, we get

$$\frac{\sum_{i=1}^{n-1} \Lambda_i}{a_n(\beta)} \xrightarrow{a.s.} \lambda(\beta). \quad (2.3)$$

Next, let \mathcal{F}_i be the natural σ -field associated to the model until time i and define

$$T_0 = 0 \quad \text{and} \quad T_n = \sum_{i=1}^n \frac{N_i - E[N_i \mid \mathcal{F}_{i-1}]}{a_i(\beta)} = \sum_{i=1}^n \frac{N_i - \Lambda_{i-1}}{a_i(\beta)}.$$

Then, (T_n) is a martingale with respect to (\mathcal{F}_n) and

$$E[T_n^2] = \sum_{i=1}^n \frac{E[(N_i - \Lambda_{i-1})^2]}{a_i(\beta)^2} = \sum_{i=1}^n \frac{E\{E[(N_i - \Lambda_{i-1})^2 \mid \mathcal{F}_{i-1}]\}}{a_i(\beta)^2} = \sum_{i=1}^n \frac{E[\Lambda_{i-1}]}{a_i(\beta)^2}.$$

Since $R_i \geq v > 0$, it is easy to verify that $E[\Lambda_i] = O(i^{-(1-\beta)})$ and so $\sup_n E[T_n^2] = \sum_{i=1}^{\infty} \frac{E[\Lambda_{i-1}]}{a_i(\beta)^2} < \infty$. Thus, (T_n) converges a.s., and the Kronecker's lemma implies

$$\frac{1}{a_n(\beta)} \sum_{i=1}^n a_i(\beta) \frac{(N_i - \Lambda_{i-1})}{a_i(\beta)} \xrightarrow{a.s.} 0,$$

so finally

$$\lim_n \frac{L_n}{a_n(\beta)} = \lim_n \frac{\sum_{i=1}^n N_i}{a_n(\beta)} = \lim_n \frac{\sum_{i=1}^n \Lambda_{i-1}}{a_n(\beta)} = \lim_n \frac{\Lambda_0 + \sum_{i=1}^{n-1} \Lambda_i}{a_n(\beta)} = \lambda(\beta) \quad \text{a.s.} \quad (2.4)$$

■

The above result entails that $\ln(L_n)/\ln(n)$ is a strongly consistent estimator of $\beta \in [0, 1]$. In fact:

- if $\beta = 0$ then $L_n \xrightarrow{a.s.} \frac{\alpha}{m_R} \ln(n)$ as $n \rightarrow +\infty$; hence $\ln(L_n) \xrightarrow{a.s.} \ln(\alpha/m_R) + \ln(\ln(n))$, therefore $\ln(L_n)/\ln(n) \xrightarrow{a.s.} \ln(\alpha/m_R)/\ln(n) + \ln(\ln(n))/\ln(n) \xrightarrow{a.s.} 0 = \beta$;
- if $\beta > 0$, we have $L_n \xrightarrow{a.s.} \lambda(\beta)n^{\beta}$ as $n \rightarrow +\infty$ so $\ln(L_n) \xrightarrow{a.s.} \ln(\lambda(\beta)) + \beta \ln(n)$, hence $\ln(L_n)/\ln(n) \xrightarrow{a.s.} \ln(\lambda(\beta))/\ln(n) + \beta \xrightarrow{a.s.} \beta$.

Remark 2.2. In practice, the value of $\ln(L_n)/\ln(n)$ may be quite far from the limit value β when n is small. Hence, it may be worth trying to fit the power-law dependence of L_n as a function of n with standard techniques [10] and use the slope of the regression line $\hat{\beta}_n$ as an effective estimator for β .

Finally, assuming that $\beta \in [0, 1]$ and m_R are known, we can get a strongly consistent estimator of α , as:

$$m_R \frac{L_n}{\ln(n)} \quad \text{for } \beta = 0 \quad \text{and} \quad m_R^{1-\beta} \beta \frac{L_n}{n^\beta} \quad \text{for } 0 < \beta \leq 1.$$

In practice, we assume β equal to the estimated value $\widehat{\beta}_n$ (as defined before) and we take m_R equal to the estimated value $\overline{R}_n = \sum_{i=1}^n R_i/n$, if the random parameters R_i are known. In Section 3.2, we will discuss the case when the random variables R_i are unknown.

Remark 2.3. Once more, it may be better in practice to estimate α as

$$\begin{aligned} \widehat{\alpha}_n &= m_R \widehat{\gamma}_n && \text{when } \beta = 0 \\ \widehat{\alpha}_n &= \beta m_R^{1-\beta} \widehat{\gamma}_n && \text{when } 0 < \beta \leq 1, \end{aligned} \tag{2.5}$$

where $\widehat{\gamma}_n$ is the slope of the regression line in the plot $(\ln(n), L_n)$ or in the plot (n^β, L_n) according to whether $\beta = 0$ or $\beta \in (0, 1]$.

We complete this section with a central-limit theorem that gives the rate of convergence of $L_n/a_n(\beta)$ to $\lambda(\beta)$ when $\beta \in [0, 1]$.

Theorem 2.4. *If $\beta \in [0, 1]$, then we have the following convergence in distribution¹:*

$$\sqrt{a_n(\beta)} \left\{ \frac{L_n}{a_n(\beta)} - \lambda(\beta) \right\} \xrightarrow{d} \mathcal{N}(0, \lambda(\beta)).$$

Proof. The result for $\beta = 1$ follows from the classical central limit theorem, since, in this case, L_n is the sum of n independent random variables with distribution $\mathcal{P}(\alpha)$. Assume now $\beta \in [0, 1)$. We first prove that

$$\sqrt{a_n(\beta)} \left\{ \frac{\sum_{i=1}^n \Lambda_{i-1}}{a_n(\beta)} - \lambda(\beta) \right\} \xrightarrow{P} 0. \tag{2.6}$$

By some calculations, condition (2.6) is equivalent to

$$\frac{\sum_{i=1}^{n-1} \left\{ \left(\sum_{j=1}^i R_j \right)^{\beta-1} - (m_R i)^{\beta-1} \right\}}{\sqrt{a_n(\beta)}} \xrightarrow{P} 0. \tag{2.7}$$

¹Actually, the convergence is in the sense of the *stable* convergence, which is stronger than the convergence in distribution. Indeed, stable convergence is a form of convergence intermediate between convergence in distribution and convergence in probability.

Since $R_j \geq v > 0$, we have $m_R \geq v > 0$ and we obtain

$$\begin{aligned}
E \left[\left| (m_R i)^{\beta-1} - \left(\sum_{j=1}^i R_j \right)^{\beta-1} \right| \right] &\leq \frac{E \left[\left| \left(\sum_{j=1}^i R_j \right)^{1-\beta} - (m_R i)^{1-\beta} \right| \right]}{(v i)^{2(1-\beta)}} \\
&\leq \frac{1}{(v i)^{2(1-\beta)}} \frac{1-\beta}{(v i)^\beta} E \left[\left| \sum_{j=1}^i R_j - m_R i \right| \right] \\
&= \frac{1-\beta}{v^{2-\beta}} \frac{1}{i^{1-\beta}} E \left[|\bar{R}_i - m_R| \right] \\
&\leq \frac{1-\beta}{v^{2-\beta}} \frac{1}{i^{1-\beta}} \sqrt{Var[\bar{R}_i]} = \frac{(1-\beta)\sqrt{Var[R_1]}}{v^{2-\beta}} \frac{i^{\beta-1}}{\sqrt{i}}.
\end{aligned}$$

This proves condition (2.7) (and so (2.6)). Indeed, we have

$$\begin{aligned}
\frac{1}{\sqrt{a_n(\beta)}} E \left[\left| \sum_{i=1}^{n-1} \left\{ \left(\sum_{j=1}^i R_j \right)^{\beta-1} - (m_R i)^{\beta-1} \right\} \right| \right] &\leq \\
\frac{1}{\sqrt{a_n(\beta)}} \sum_{i=1}^{n-1} E \left[\left| (m_R i)^{\beta-1} - \left(\sum_{j=1}^i R_j \right)^{\beta-1} \right| \right] &\leq \\
\frac{(1-\beta)\sqrt{Var[R_1]}}{v^{2-\beta}} \frac{1}{\sqrt{a_n(\beta)}} \sum_{i=1}^{n-1} \frac{1}{i^{1-(\beta-1/2)}} &\rightarrow 0.
\end{aligned}$$

Next, define

$$T_n = \sqrt{a_n(\beta)} \left\{ \frac{L_n}{a_n(\beta)} - \frac{\sum_{i=1}^n \Lambda_{i-1}}{a_n(\beta)} \right\} = \frac{\sum_{i=1}^n (N_i - \Lambda_{i-1})}{\sqrt{a_n(\beta)}}.$$

In view of (2.6), it suffices to show that $T_n \xrightarrow{d} \mathcal{N}(0, \lambda(\beta))$.

To this end, for $n \geq 1$ and $i = 1, \dots, n$, define

$$T_{n,i} = \frac{N_i - \Lambda_{i-1}}{\sqrt{a_n(\beta)}}, \quad \mathcal{G}_{n,0} = \mathcal{F}_0 \quad \text{and} \quad \mathcal{G}_{n,i} = \mathcal{F}_i,$$

where \mathcal{F}_i is the natural σ -field associated to the model until time i . Then, we have $E[T_{n,i} | \mathcal{G}_{n,i-1}] = 0$, $\mathcal{G}_{n,i} \subseteq \mathcal{G}_{n+1,i}$ and $T_n = \sum_{i=1}^n T_{n,i}$. Thus, by a martingale central limit theorem (see [17]), $T_n \xrightarrow{d} \mathcal{N}(0, \lambda(\beta))$ provided

$$(i) \quad \sum_{i=1}^n T_{n,i}^2 \xrightarrow{P} \lambda(\beta), \quad (ii) \quad \max_{1 \leq i \leq n} |T_{n,i}| \xrightarrow{P} 0, \quad (iii) \quad \sup_n E \left[\max_{1 \leq i \leq n} T_{n,i}^2 \right] < \infty;$$

Let

$$D_i = (N_i - \Lambda_{i-1})^2 \quad \text{and} \quad U_n = \frac{\sum_{i=1}^n \{D_i - E[D_i \mid \mathcal{F}_{i-1}]\}}{a_n(\beta)} = \frac{\sum_{i=1}^n (D_i - \Lambda_{i-1})}{a_n(\beta)}.$$

By the same martingale argument used in the proof of the previous theorem and by Kronecker's lemma, $U_n \xrightarrow{a.s.} 0$. Then, by (2.3),

$$\sum_{i=1}^n T_{n,i}^2 = \frac{\sum_{i=1}^n D_i}{a_n(\beta)} = U_n + \frac{\sum_{i=1}^n \Lambda_{i-1}}{a_n(\beta)} \xrightarrow{a.s.} \lambda(\beta).$$

This proves condition (i). As to (ii), fix $k \geq 1$ and note that

$$\max_{1 \leq i \leq n} T_{n,i}^2 \leq \frac{\max_{1 \leq i \leq k} D_i}{a_n(\beta)} + \max_{k < i \leq n} \frac{D_i}{a_i(\beta)} \leq \frac{\max_{1 \leq i \leq k} D_i}{a_n(\beta)} + \sup_{i > k} \frac{D_i}{a_i(\beta)} \quad \text{for } n > k.$$

Hence, $\limsup_n \max_{1 \leq i \leq n} T_{n,i}^2 \leq \limsup_n \frac{D_n}{a_n(\beta)}$ and condition (ii) follows since

$$\frac{D_n}{a_n(\beta)} = \frac{\sum_{i=1}^n D_i}{a_n(\beta)} - \frac{\sum_{i=1}^{n-1} D_i}{a_n(\beta)} \xrightarrow{a.s.} 0.$$

Finally, condition (iii) is a consequence of

$$\begin{aligned} E \left[\max_{1 \leq i \leq n} T_{n,i}^2 \right] &\leq \frac{\sum_{i=1}^n E[D_i]}{a_n(\beta)} = \frac{\sum_{i=1}^n E[\Lambda_{i-1}]}{a_n(\beta)} = \\ &= \frac{\Lambda_0 + \sum_{i=1}^{n-1} E[\Lambda_i]}{a_n(\beta)} \leq \frac{\alpha \left(1 + \sum_{i=1}^{n-1} (vi)^{\beta-1} \right)}{a_n(\beta)}. \end{aligned}$$

■

2.2 Analysis of the random fitness parameters R_i

Now our purpose is to find, under the assumption of our model, a procedure to get information on the random variables R_i from the data, that typically are the values of Z_1, \dots, Z_n , i.e., n rows of the matrix Z , where n is the number of the observed nodes.

Unfortunately, this goal is not easily tractable as we will point out in the sequel. The method we empirically tested extracts from the data, with a maximum log-likelihood procedure (see Section 3.2), a plausible realization $\hat{r}_1, \dots, \hat{r}_{k_n}$ of R_1, \dots, R_{k_n} , for a suitable k_n ; this information could be useful, for instance, to reconstruct the ranking induced by R_i . Note that we ideally would like to find a probable realization for all the fitness parameters of the observed nodes (not only for the first k_n nodes), but we do not possess the same amount of information about all R_i : in particular, while R_1 influences all the observed rows of the matrix Z ,

R_{n-1} has only influence over Z_n . So we cannot expect to find good values for all the random variables.

With the above purpose in mind, we now give a general expression for the conditional probability of observing $Z_1 = z_1, \dots, Z_n = z_n$ given R_1, \dots, R_{n-1} . We refer to Section 2 for the notation.

The first row Z_1 is simply identified by $L_1 = N_1$ and so

$$\begin{aligned} P(Z_1 = z_1) &= P(N_1 = n_1 = \text{card}\{k : z_{1,k} = 1\}) \\ &= \text{Poi}(\alpha)\{n_1\} = e^{-\alpha} \frac{\alpha^{n_1}}{n_1!}. \end{aligned}$$

Then the second row is identified by the values $Z_{2,k}$ with $k = 1, \dots, L_1 = N_1$ and by N_2 and so

$$\begin{aligned} P(Z_2 = z_2 | Z_1, R_1) &= \\ P(Z_{2,k} = z_{2,k} \text{ for } k = 1, \dots, L_1, N_2 = n_2 = \text{card}\{k > L_1 : z_{2,k} = 1\} | Z_1, R_1) &= \\ \prod_{k=1}^{L_1} P_1(k)^{z_{2,k}} (1 - P_1(k))^{1-z_{2,k}} \times \text{Poi}(\Lambda_1)\{n_2\}, \end{aligned}$$

where $P_1(k)$ is defined in (2.1) and Λ_1 is defined in (2.2).

The general formula is

$$\begin{aligned} P(Z_{j+1} = z_{j+1} | Z_1, R_1, \dots, Z_j, R_j) &= \\ P(Z_{j+1,k} = z_{j+1,k} \text{ for } k = 1, \dots, L_j, \\ N_{j+1} = n_{j+1} = \text{card}\{k > L_j : z_{j+1,k} = 1\} | Z_1, R_1, \dots, Z_j, R_j) &= \\ \prod_{k=1}^{L_j} P_j(k)^{z_{j+1,k}} (1 - P_j(k))^{1-z_{j+1,k}} \times \text{Poi}(\Lambda_j)\{n_{j+1}\}, \end{aligned}$$

where $P_j(k)$ is defined in (2.1) and Λ_j is defined in (2.2).

Hence, for n nodes, we can write a formula for the conditional probability of observing $Z_1 = z_1, \dots, Z_n = z_n$ given R_1, \dots, R_{n-1} :

$$\begin{aligned} P(Z_1 = z_1, \dots, Z_n = z_n | R_1, \dots, R_{n-1}) &= \\ P(Z_1 = z_1) \prod_{j=1}^{n-1} P(Z_{j+1} = z_{j+1} | Z_1, R_1, \dots, Z_j, R_j). \end{aligned} \tag{2.8}$$

2.2.1 A Monte Carlo method

The algorithm we applied is essentially a MCMC (Markov Chain Monte Carlo) method [13], which uses the basic principle of Gibbs sampling [9]: fix all components of a vector except one and compare the different values of the likelihood obtained for various values of the non-fixed component.

The method employs the aforementioned formula (2.8) for the conditional probability of observing $Z_1 = z_1, \dots, Z_n = z_n$ given the values of R_1, \dots, R_{n-1} . Precisely, using the symbol \bar{z} in order to denote the matrix with rows z_1, \dots, z_n and the symbol \bar{r} in order to denote a vector of component r_1, \dots, r_n , set

$$P(Z = \bar{z} | R = \bar{r}) = P(Z_1 = z_1, \dots, Z_n = z_n | R_1 = r_1, \dots, R_n = r_n). \quad (2.9)$$

We want to find a vector $\hat{\bar{r}}$ that is a point maximizing the likelihood function (2.9) corresponding to the observed \bar{z} . ²

The basic algorithm is described in Alg. 1. It is regulated by these parameters:

- $\bar{r}^0 \in \mathbb{R}^n$ is the initial guess for $\hat{\bar{r}}$;
- $J \in \mathbb{N}^+$ is the number of “jumps to a new value”, i.e., the number of the new values analyzed for a certain component at each step;
- $\sigma \in \mathbb{R}^+$ is the standard deviation of each “jump”.

Algorithm 1 Basic Monte Carlo algorithm to find $\hat{\bar{r}}$.

INPUT: z_1, \dots, z_n , the observed features of each of the n observed nodes, i.e., the first n rows of the attribute matrix Z

OUTPUT: $\hat{\bar{r}}$, a maximum point for the likelihood function associated to the input data

DESCRIPTION:

1. $\hat{\bar{r}} \leftarrow \bar{r}^0$
2. Repeat the following loop until convergence:
 - (a) Choose a random node $i \in \{1, \dots, n\}$
 - (b) Extract J values h_1, \dots, h_J from the normal distribution $\mathcal{N}(r_i, \sigma^2)$; re-sample each h_j until $h_j > 0$.
 - (c) For each value h_j , compute

$$\mathcal{L}(h_j) = P(Z = \bar{z} | R_1 = r_1, \dots, R_i = h_j, \dots, R_n = r_n)$$

$$(d) \hat{r}_i \leftarrow \arg \max_{h \in \{r_i, h_1, \dots, h_J\}} \mathcal{L}(h)$$

²We point out that our algorithm can not be considered a proper statistical estimation procedure for the fitness parameters. In particular, although it resembles the Bayesian *Maximum a posteriori estimation* (MAP) when the a priori distribution is (improperly) uniform, we do not have a vector of parameters with a fixed dimension: the number of parameters in our case increases with the number of observations.

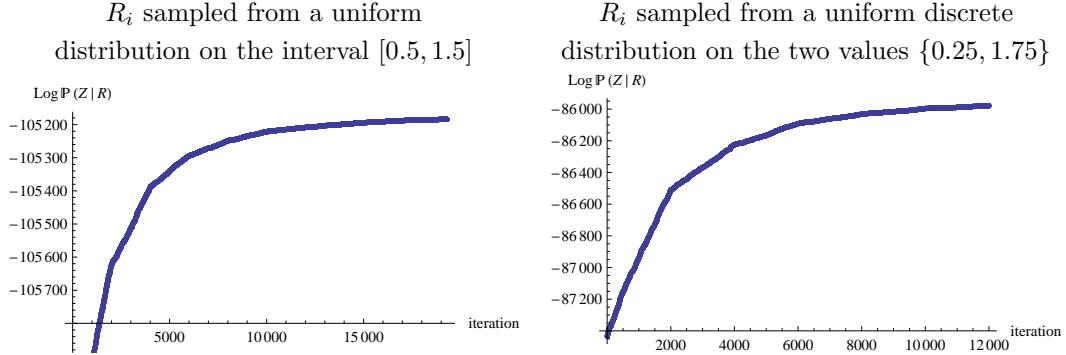


Figure 1: Value of the log-likelihood during the execution of the algorithm, for different distributions of R_i . The chosen algorithm parameters are $\sigma^2 = 1$, $J = 4$ and $\bar{r}^0 = \mathbf{1}$ (the vector with all 1's). The matrix Z has 2000 nodes and it was generated with $\alpha = 3$ and $\beta = 0.9$.

It is worth to note that, given $\mathcal{L}(r_i)$, it is possible to find $\mathcal{L}(h)$ without re-doing the whole computation. In fact, let us consider the product in eq. (2.8): a change from r_i to h must be taken into account only from the i -th factor onward – that is, for the factors that come after $P(Z_i = z_i | Z_1, R_1, \dots, Z_{i-1}, R_{i-1})$. In particular, let $\delta = h - r_i$; then, for each j -th factor, with $j \geq i$, we have to:

- add δ to the term $\sum_{i=1}^j R_i$, inside Λ_j and $P_j(k)$ (defined in eq. (2.1) and (2.2));
- add δ to the numerator of $P_j(k)$ when k is s.t. $z_{i,k} = 1$; that is to say, change the global weight of a feature only if the node we changed has that feature.

Every other term in the equation remains unchanged and does not need to be computed again. This remark allows us to speed up the implementation considerably.

Figure 1 confirms that the algorithm moves toward a vector \hat{r} maximizing $P(Z = \bar{z} | R = \bar{r})$ and shows that the algorithm effectively converges. As a stopping criterion, we can use the maximum increase in the log-likelihood in the last iterations: when this is under a certain threshold t , we stop the algorithm. The obtained outputs will be discussed in details in Section 3.2.

As already said, one point that we need to keep in mind is that we do not possess the same amount of information about all the random variables R_i : in particular, while R_1 influences all the rows of the matrix Z , R_{n-1} has only influence over the last one. So we cannot expect the output values to be very accurate for the last segment. For this reason, we also implemented a variant of the algorithm that considers only the first k_n nodes. Thus, we have another algorithm parameter k_n so that the choice of the jumping node at step 2(a) is restricted to $i \in \{1, \dots, k_n\}$.

and, finally, the output will be the corresponding segment of $\widehat{\bar{r}}$, i.e., $\widehat{r}_1, \dots, \widehat{r}_{k_n}$. This variant converges faster and moreover it allows to use larger values of the algorithm parameter J .

Another relevant point is that the parameters α and β enter the expression (2.8). Therefore, in practice, before applying the algorithm, we need to estimate them. As shown in Remark 2.2, we are able to estimate β starting from the observed values of the matrix Z . On the other hand, as shown in Remark 2.3, the estimation of α presupposes the knowledge of the mean value m_R of the fitness parameters R_i (except for the special case $\beta = 1$). Hence, we are in the situation in which, in order to get information on the fitness parameters by the proposed algorithm, we need to estimate α and β , but, in order to estimate α , we need to know the mean value m_R . This problem can be partially solved as follows. Since the term $P(Z_1 = z_1)$ does not contain the R_i 's, the research of a vector $\widehat{\bar{r}}$ that maximizes (2.9) is equivalent to the research of a vector $\widehat{\bar{r}}$ maximizing the product

$$\prod_{j=1}^{n-1} P(Z_{j+1} = z_{j+1} | Z_1, R_1, \dots, Z_j, R_j)$$

in formula (2.8). On the other hand, each term of the above product contains the inclusion probabilities $P_j(k)$, that are invariant with respect to the normalization of the R_i 's by their mean value m_R , and the Λ_j 's that have the property

$$\Lambda_j = f(\alpha, \beta, \bar{r}) = f(\alpha/(m_R)^{1-\beta}, \beta, \bar{r}/m_R)$$

(where \bar{r}/m_R denotes the vector with components r_i/m_R). Consequently, starting from the observed values of the matrix Z , we can

- first, estimate β by Remark 2.2;
- then estimate $\alpha' = \alpha/(m_R)^{1-\beta}$ by Remark 2.3 (i.e., $\widehat{\alpha}'_n$ equal to $\widehat{\gamma}_n$ or $\beta \widehat{\gamma}_n$ according to the estimated value of β);
- finally, extract a plausible realization $\widehat{\bar{r}}' = \widehat{\bar{r}}/m_R$ (of the random variables $R'_i = R_i/m_R$) as a maximum point of the corresponding expression of the likelihood with the estimated value of β and α' .

Therefore, the output of the algorithm will be α' , β and a plausible realization $\widehat{\bar{r}}'$ of the random variables $R'_i = R_i/m_R$.

Finally, we highlight that it is possible to experiment other variants of the algorithm, for example, by using a distribution different from the normal for the jumps, or changing σ during the execution (e.g., reducing it according to some “cooling schedule”, as it happens in simulated annealing [12]). Additionally, instead of looking for the values on the whole positive real line, we could restrict the research on a suitable interval (guessed for the particular real case).

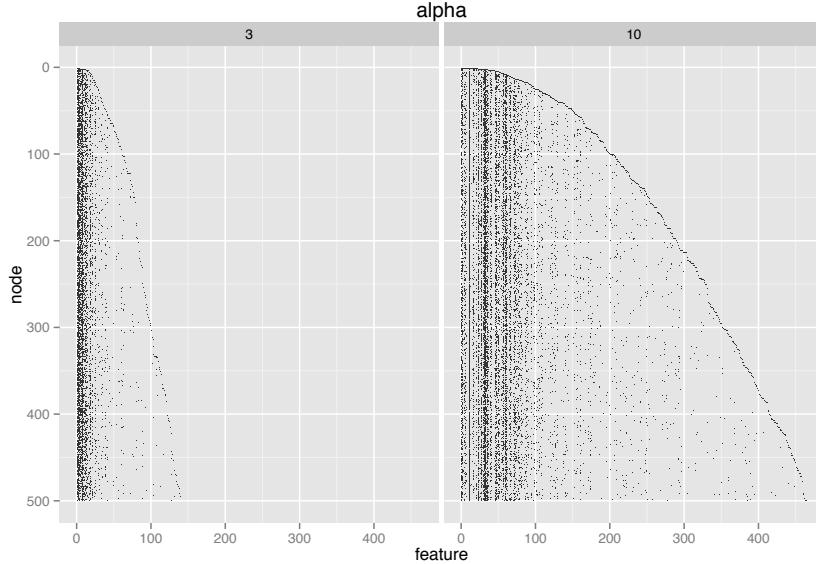


Figure 2: The Z matrix for $n = 500$, two different values of α ($\alpha = 3$ and $\alpha = 10$) and a fixed $\beta = 0.5$. The random variables R_i are uniformly distributed on the interval $[0.25, 1.75]$.

3 Simulations for the attribute matrix

In this section, we shall present a number of simulations we performed in order to illustrate the role of the parameters of the model and also to see how good the proposed tools turn out to be.

3.1 Estimating α and β

Firstly, we aim at pointing out the role played by the model parameters α and β . Therefore, we fix a distribution for the random fitness parameters with $m_R = 1$ and we simulate the matrix Z for different values of α and β (fixing one and making the other one change). More precisely, we assume that the random variables R_n are uniformly distributed on the interval $[0.25, 1.75]$.

In Figure 2, we visualize the effect of α : a larger α yields a larger number of new attributes per node.

In Figure 3, instead, we visualize how different positive values of β yield a different power-law (asymptotic) behavior of L_n . Indeed, in this figure, we have the log-log plot of L_n as a function of n . In the first two panels, we present two different positive values of β (0.75 and 0.5), showing the correspondence with the power-law exponent of L_n , estimated by the slope of the regression line. Moreover, in the third panel, we point out that the parameter α do not affect the power-law exponent of L_n .

Figure 4 underlines that the estimator proposed in remark 2.2 works better (i.e. with a more precision) for large values of β since L_n reaches the power-law behavior more quickly for larger values of β .

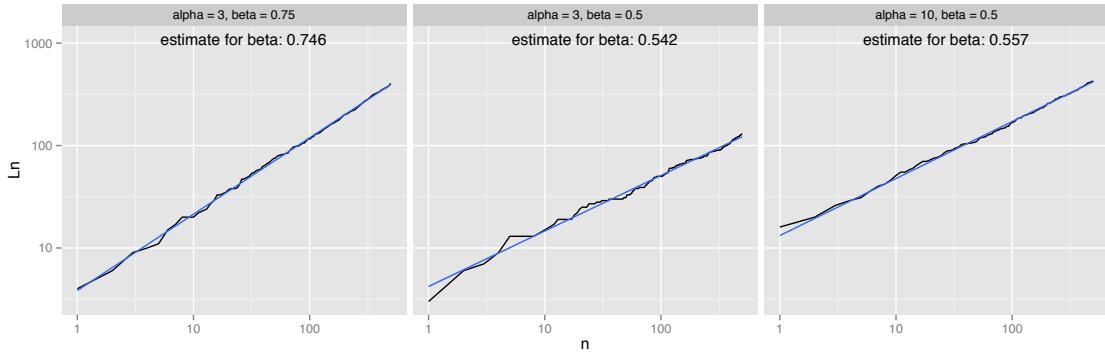


Figure 3: Correspondence between the parameter β and the power-law exponent of L_n as a function of n . The estimate of β is the slope of the regression line. Here, we have 500 nodes, the random variables R_i are uniformly distributed, on the interval $[0.25, 1.75]$. Values for α and β are indicated above; we can see how different values for α do not affect the power-law behaviour.

Similarly, we evaluated the estimator $\hat{\alpha}_n$ of α , obtained by using the slope of the regression line in the plot of L_n as a function of n^β , as said in Remark 2.3 (note that we have $m_R = 1$ and so α coincides with α'). Results are illustrated in Figure 5 and show how this estimator yields good results.

We also checked how the shape of the matrix Z is influenced by the distribution of the random parameters R_n . More precisely, we analyzed the effect of ε on the shape of Z when the random variables R_i are uniformly distributed on the interval $[\varepsilon, 2 - \varepsilon]$, with $0 < \varepsilon \leq 1$, so that $E[R_i] = m_R = 1$ and the variance of R_i is $Var[R_i] = (1 - \varepsilon)^2/3$, which goes to zero as $\varepsilon \rightarrow 1$. Hence, when ε is smaller, the variance of the R_n 's is larger, so that also a “young” nodes i have some chance of transmitting their attributes to the other nodes (recall that a larger R_i makes i more successful in transmitting its own attributes). This is witnessed (see Figure 6) by the number of “blackish” vertical lines, that are more or less widespread in the whole spectrum of nodes; whereas for larger ε they are more concentrated on the left-hand side (i.e., only the first nodes successfully transmit their attributes).

3.2 Analysis of the random fitness parameters R_i

We proceeded to test empirically how the Monte Carlo method performs in recovering the information on the fitness parameters R_i . We tested its behavior against various distributions of R_i ; specifically, a uniform distribution on an interval, a two-class uniform distribution, and finally a discrete power-law distribution with 10

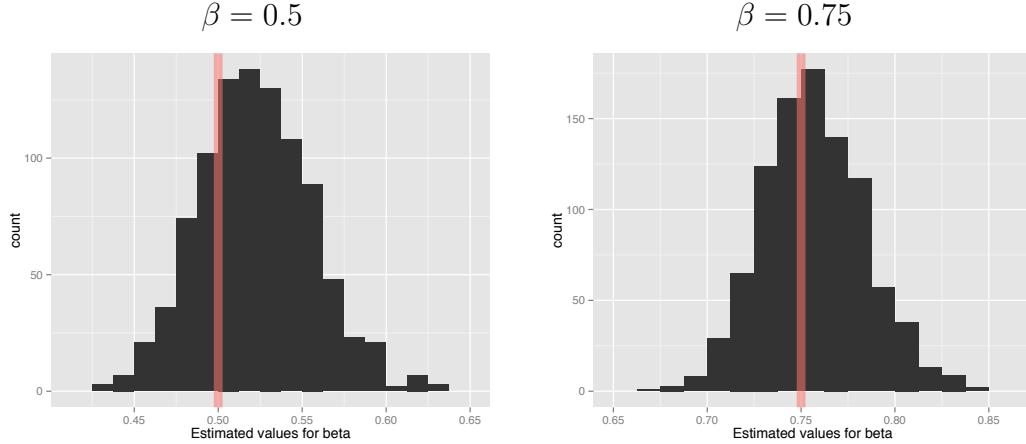


Figure 4: Distribution of the estimator $\hat{\beta}_n$ of β over 1000 experiments, each with $n = 2000$ and $\alpha = 3$. The random variables R_i are uniformly distributed on the interval $[0.25, 1.75]$. The red line indicates the true value of β .

classes. In the following of this section we illustrate the details of such experiments, while, in the next section, we will try to measure the performance of the proposed technique.

In every experiment, the matrix Z has $n = 2000$ nodes and it was generated with $\alpha = 3$ and $\beta = 0.9$. The Monte Carlo algorithm parameters were set as follows: $\sigma^2 = 1$, $J = 4$ and $\bar{r}^0 = \mathbf{1}$ (the vector with all 1's).

For the first experiment, each R_i is sampled from the uniform distribution on the interval $[0.5, 1.5]$. We used the previously discussed techniques to find the estimates of α and β : the estimated values are $\hat{\alpha} = 3.095$ and $\hat{\beta} = 0.893$ (note that we have $m_R = 1$ and so $\alpha = \alpha'$ and $\hat{r} = \hat{r}'$). Then, we tried the proposed Monte Carlo algorithm with the stopping threshold $t = 1/4$. Results are visualized in Figure 7, according to two different orderings of the nodes:

- i) in the natural order, so that we confirm that our predictions are better for the first (i.e., the oldest) nodes than for the last (i.e., the youngest) ones;
- ii) ordered by their true fitness values, so that we can show that we are, more or less, able to reconstruct the relative order of the fitness parameters (this fact will be made clearer in Section 3.3).

In the second experiment, we applied our algorithm to a discrete case: we sampled the fitness parameters R_i from a set of only two values, $\{0.25, 1.75\}$, each with probability $\frac{1}{2}$. We left the parameters of the model and the ones of the algorithm unaltered (except for moving the stopping threshold t from $1/4$ to 1). The estimated

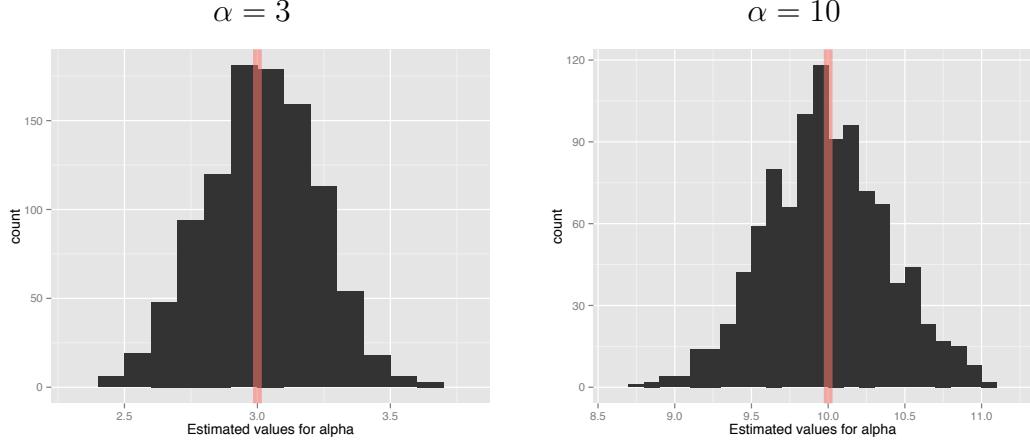


Figure 5: Distribution of the estimator $\hat{\alpha}_n$ of α over 1000 experiments, each with $n = 2000$ and $\beta = 0.5$. The random variables R_i are uniformly distributed on the interval $[0.25, 1.75]$ (so $m_R = 1$ and $\alpha = \alpha'$). The red line indicates the true value of α .

values for $\alpha = 3$ and $\beta = 0.9$ are, respectively, $\hat{\alpha} = 2.922$ (again $m_R = 1$ and so $\alpha = \alpha'$ and $\hat{r} = \hat{r}'$) and $\hat{\beta} = 0.903$. The results of this second experiment are more encouraging (we will see precise measurements in Section 3.3). In this case, the output values of the algorithm are closer to the true ones (see Figure 8). Moreover, we can still observe the same phenomena, i) and ii), described above.

Finally, we applied the algorithm to a third case: we sampled R_i from a normalized power-law discrete distribution, with 10 possible values – specifically, a normalized discrete Zipf's law with exponent 2 and number of values 10. We left both algorithm and model parameters unaltered (and we used 1 as the stopping threshold t).

The estimated values for $\alpha = 3$ and $\beta = 0.9$ are, respectively, $\hat{\alpha} = 3.595$ (again $m_R = 1$ and so $\alpha = \alpha'$ and $\hat{r} = \hat{r}'$) and $\hat{\beta} = 0.868$.

Results for this case show that – despite the fact that we have now a discrete distribution with more than two values – our approach can recover information (especially for larger values of fitness), as can be seen in Figure 9 and in Section 3.3.

We conclude this section noting that, for each of the experiments, the Monte Carlo algorithm looks for the values of the fitness parameters on the whole positive real line. We would obtain better outputs if we could restrict the research on a suitable interval for each case, assuming a partial knowledge of the shape of the distribution.

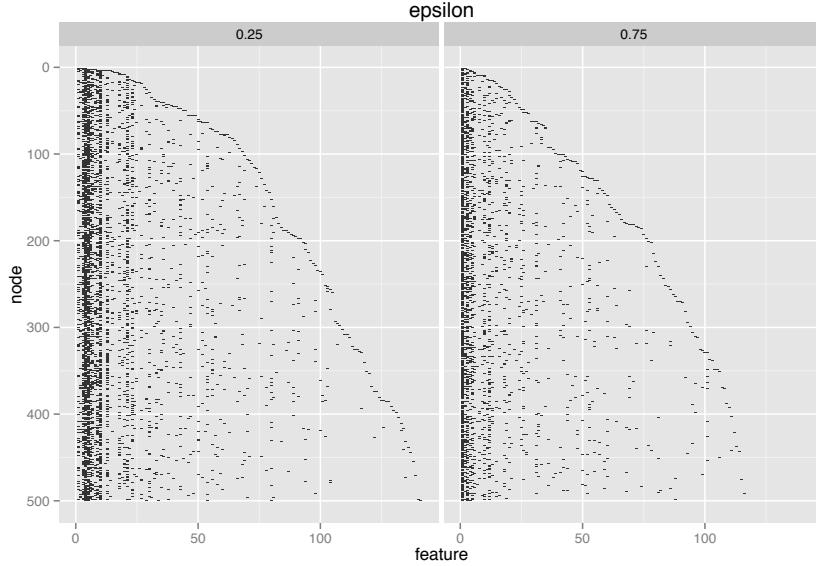


Figure 6: Here $n = 500$, $\alpha = 3$, $\beta = 0.5$ and the random variables R_i are uniformly distributed on the interval $[\varepsilon, 2 - \varepsilon]$ (so that $m_R = 1$ and $Var[R_i] = (1 - \varepsilon)^2/3$), for different values of ε ($\varepsilon = 0.25$ and $\varepsilon = 0.75$). The figure shows how ε affects the shape of Z .

3.3 Analysis of the ordering of the nodes

In a real application, we may content ourselves in finding not the realized fitness parameters themselves but rather their ordering, that is, the ordering of the nodes from larger to smaller values of the fitness parameter. To evaluate if we can at least extract values \hat{r}_i that respect this ordering, we decided to compare the drawn vector $\hat{\bar{r}}$ with the true realization \bar{r} by the use of Kendall's τ and some variants of it.

To keep track of the fact that, as said before, the first nodes contain more information than the last ones, we evaluated Kendall's τ not only on the whole vector but also on a short initial segment of size $k_n = n/2$ or $k_n = \sqrt{n}$. Besides this, we tried to use a variant of Kendall's τ (proposed in [35]), that we apply in two separate and different ways:

1. inducing a hyperbolic decay based on the position of the nodes – that is, weighting more the first (the oldest) nodes, and less the last (the youngest) ones;
2. inducing a hyperbolic decay based on the true realized values r_i – that is, assigning a higher weight to the nodes with a greater fitness parameter r_i .

The results of these measures are summarized in Table 1 for the experiment with the uniform distribution on an interval, in Table 2 for the experiment with

Considered nodes	Kendall's τ	τ weighted by position	τ weighted by value
$k_n = \lfloor \sqrt{n} \rfloor = 44$.281	.206	.463
$k_n = \frac{n}{2} = 1000$.229	.188	.337
$k_n = n = 2000$.150	.139	.155

Table 1: Comparing orderings induced by the true realization \bar{r} versus the extracted one $\hat{\bar{r}}$ in the case of the uniform distribution on the interval $[0.5, 1.5]$.

Considered nodes	Kendall's τ	τ weighted by position	τ weighted by value
$k_n = \lfloor \sqrt{n} \rfloor = 44$.676	.593	.713
$k_n = \frac{n}{2} = 1000$.586	.585	.625
$k_n = n = 2000$.438	.477	.434

Table 2: Comparing orderings induced by the true realization \bar{r} versus the extracted one $\hat{\bar{r}}$ in the case of the uniform discrete distribution on the two values $\{0.25, 1.75\}$.

Considered nodes	Kendall's τ	τ weighted by position	τ weighted by value
$k_n = \lfloor \sqrt{n} \rfloor = 44$.735	.762	.772
$k_n = \frac{n}{2} = 1000$.453	.516	.803
$k_n = n = 2000$.313	.402	.543

Table 3: Comparing orderings induced by the true realization \bar{r} versus the extracted one $\hat{\bar{r}}$ in the case of the normalized discrete Zipf's distribution with exponent 2 and 10 values.

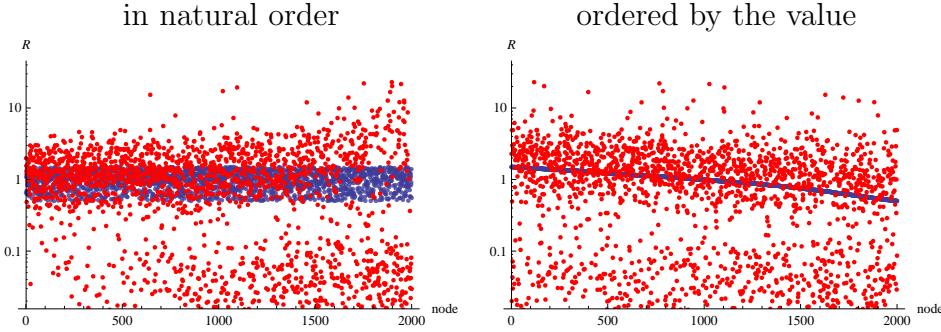


Figure 7: The extracted realization \hat{r} (in red) versus the true realization \bar{r} (in blue), with two different orderings, in the case of uniform distribution on the interval $[0.5, 1.5]$. The empirical mean of the the first $\frac{n}{2}$ extracted values is 1.18.

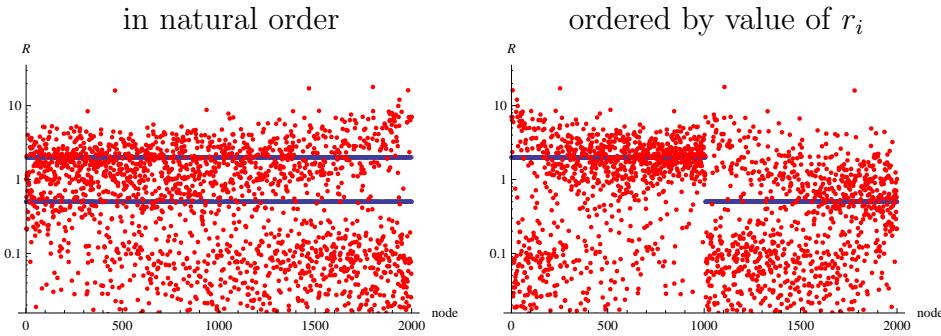


Figure 8: The extracted realization \hat{r} (in red) versus the true realization \bar{r} (in blue), with two different orderings, in the case of the uniform discrete distribution on the two values $\{0.25, 1.75\}$. The empirical mean of the the first $\frac{n}{2}$ extracted values is 1.33.

the uniform discrete distribution on the two values $\{0.25, 1.75\}$, and in Table 3 for the discrete Zipf's distribution with 10 values and exponent 2. The tables show that, although we are unable to reconstruct the actual realized values of the fitness parameters, our approach actually recovers some information about node ranking. As already seen before, the output of the Monte Carlo algorithm is better for the discrete cases.

4 From the attribute structure to the graph

We now extend the model to produce a graph out of the attribute structure (that may itself be latent and unknown). In general, we may assume that the presence of an edge between two nodes depends on the features that those nodes exhibit, but

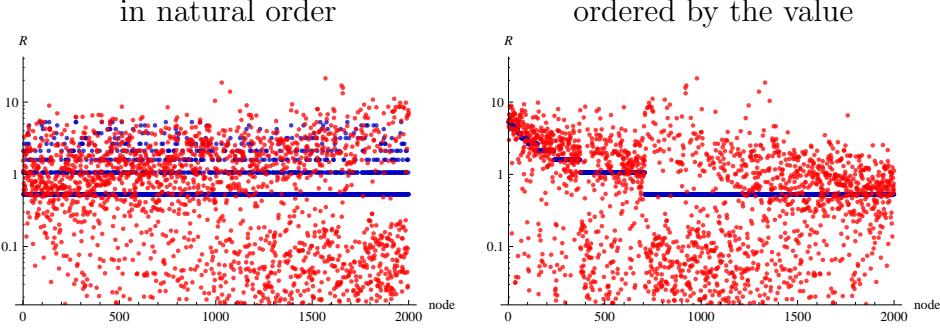


Figure 9: The extracted realization \widehat{r} (in red) versus the true realization \bar{r} (in blue), with two different orderings, for the normalized discrete Zipf's distribution with exponent 2 and 10 values. The empirical mean of the the first $\frac{n}{2}$ extracted values is 1.25.

there are many nuances to this idea and possible approaches.

In general, we postulate that the connections are undirected (we omit self-loops, i.e., edges of type (i, i)) and we assume that, conditioned on Z (and some other variables), the probability of having at time n a certain adjacency matrix (symmetric by assumption) $a = (a_{i,j})_{1 \leq i,j \leq n}$ (with $a_{i,j} \in \{0, 1\}$) is

$$\begin{aligned} P(A = a | Z, \text{other variables}) &= P\left(\bigcap_{1 \leq j < i \leq n} \{A_{i,j} = a_{i,j}\} | Z, \text{other variables}\right) \\ &= \prod_{1 \leq j < i \leq n} P(A_{i,j} = a_{i,j} | Z, \text{other variables}). \end{aligned}$$

4.1 Feature/Feature probability model (FF)

In the first, basic model, we assume that the probability of having an edge (i, j) depends *solely* on the features that i and j possess; each pair of feature that node i and node j exhibit contributes in tuning the edge probability. In other words, letting L_n be the total number of different features (i.e., columns of Z), we assume that there is a symmetric feature/feature influence matrix $\Xi = (\xi_{h,k})_{1 \leq h,k \leq L_n}$ that determines a node-node weight matrix W given by

$$W = Z \cdot \Xi \cdot Z^T$$

or, more explicitly,

$$w_{i,j} = \sum_{h,k} Z_{i,h} \xi_{h,k} Z_{j,k}.$$

The probability of the presence of an edge (i, j) , then, depends monotonically on $w_{i,j}$. The choice of Ξ determines different relations between features and edge probabilities. If $\xi_{h,k} > 0$ (resp., $\xi_{h,k} < 0$), then the simultaneous presence of attributes h

and k increases (resp., decreases) the edge-probability; if $\xi_{h,k} = 0$, the simultaneous presence of attributes h and k does not affect the edge-probability. In particular, if $\xi_{h,k} = 0$ for $h \neq k$, then the edge-probability is affected only by the presence of the same attribute in both nodes (positively or negatively affected depending on the sign of $\xi_{h,h}$).

The actual probabilities are computed as some function applied to the corresponding weight; i.e., some monotone function $\Phi : \mathbf{R} \rightarrow [0, 1]$ is fixed and

$$P(A_{i,j} = 1|Z) = \Phi(w_{i,j}) = \Phi\left(\sum_{h,k} Z_{i,h} \xi_{h,k} Z_{j,k}\right). \quad (4.1)$$

4.2 Feature/Feature+BA probability model (FFBA)

A variant of the feature/feature (FF) probability model takes into account the fact that some edges exist independently of the features that the involved nodes exhibit, but they are there simply because of the popularity of a node, as in the traditional “preferential attachment” model by Barabási and Albert [3]. To take this into consideration, instead of using (4.1), we rather define for $1 \leq j < i \leq n$

$$P(A_{i,j} = 1|Z, D_j(i-1), m(i-1)) = \delta \Phi\left(\sum_{h,k} Z_{i,h} \xi_{h,k} Z_{j,k}\right) + (1-\delta) \frac{D_j(i-1)}{2m(i-1)}, \quad (4.2)$$

where $D_j(k)$ and $m(k)$ are, respectively, the degree of node j and the overall number of edges just after node k was added. The parameter δ controls the mixture between the pure feature/feature model and the preferential-attachment model (degenerating to the first when $\delta = 1$, and to the second when $\delta = 0$).

4.3 Feature/feature+JR probability model (FFJR)

Jackson and Rogers [20] observed that preferential-attachment can be induced also injecting a “friend-of-friend” approach in the creation of edges. Their behavior can be mimicked in our model as follows: we first generate a graph with adjacency matrix A' using the pure FF model, i.e., letting

$$P(A'_{i,j} = 1|Z) = \Phi(w_{i,j}) = \Phi\left(\sum_{h,k} Z_{i,h} \xi_{h,k} Z_{j,k}\right).$$

After this, every node i looks at the set of the neighbors of its neighbors, according to A' . If this set is not empty, it then selects one node from the set uniformly at random; the resulting node is chosen as an “extra” friend of i with some probability $1 - \delta$ (for suitably chosen $\delta \in [0, 1]$). The adjacency matrix obtained in this way is A . Once more, if $\delta = 0$ we have $A = A'$ so we get back to the pure FF model.

5 Simulations for the graph structure

The purpose of this collection of experiments is to determine the topology of the graph generated with the models described above. We fix *a priori* the number of nodes n and the (approximate) number of edges m (i.e., density) we aim at; then, every experiment consists essentially in two phases:

- generating an attribute matrix Z for n nodes (with certain values for the parameters α and β and with R_i uniformly distributed on the interval $[\varepsilon, 2 - \varepsilon]$ for a certain ε);
- building the graph according to one of the models described in Section 4.

The second phase needs to fix some further parameters: Ξ (the feature/feature influence matrix), the function Φ and, for the mixed models (FFBA and FFJR), the parameter δ .

For the sake of simplicity, throughout this section, we assume that $\Xi = I$ and we take Φ as a sigmoid function given by

$$\Phi(x) = \frac{1}{e^{K(\vartheta-x)} + 1}.$$

In other words, the existence of an edge (i, j) depends simply on the number of features that i and j share (this is an effect of choosing $\Xi = I$). More features induce larger probability: the sigmoid function smoothly increases (from 0 to 1) around a threshold ϑ , and $K > 0$ controls its smoothness; when $K \rightarrow \infty$ we obtain a step function and edges are chosen deterministically based on whether the two involved nodes share more than ϑ features or not.

In the experiments, we fix K and determine ϑ on the basis of the desired density of the graph (or, equivalently, the desired number of edges m); in practice³, this is obtained by solving numerically the equation

$$E \left[\sum_{1 \leq j < i \leq n} A_{i,j} \right] = \sum_{1 \leq j < i \leq n} \Phi \left(\sum_{h,k} Z_{i,h} \xi_{h,k} Z_{j,k} \right) = m$$

for the indeterminate ϑ (using, for example, Newton's method). Since $\Xi = I$ the equation in fact simplifies into

$$\sum_{1 \leq j < i \leq n} \frac{1}{e^{K(\vartheta - \sum_h Z_{i,h} Z_{j,h})} + 1} = m.$$

With these assumptions, every experiment depends on the parameters used for generating Z (i.e., α , β and ε), on K (that controls the smoothness of the sigmoid

³The described method needs some (obvious) adjustments when applied to the mixed models, to take into account the edges generated by preferential attachment.

function) and on δ (for the mixed models). In the graphs produced by each simulation, we took into consideration the degree distribution, the percentage of reachable pairs (i.e., the fraction of pairs of nodes that are reachable) and the distribution of distances (lengths of shortest paths); the latter data are computed using a probabilistic algorithm [32].

Some of the results obtained (for $n = 2000$ and⁴ $m = 4000$) for the FF model are shown in Figure 10. For those experiments, the underlying attribute matrix is generated with $\beta = 0.75$ and R_i uniformly distributed on the interval $[0.75, 1.25]$; we compare $\alpha = 3$ (resulting in ≈ 1200 features) with $\alpha = 10$ (≈ 4000 features). Results regarding mixed models are reported in Figure 11.

The properties of the obtained graphs can be summarized as follows:

- the pure FF model exhibits a behavior that strongly depends on the smoothness parameter K (see Fig. 10):
 - for $K = 1$, the degree distribution is power-law only when α is large (e.g., $\alpha = 10$), whereas the distribution is often non-monotonic for smaller α 's, especially on large graphs; the fraction of reachable pairs is quite large (between 40% and 90%);
 - for $K = 4$, degrees are always distributed as a power-law (with exponents around 3), but the graph becomes largely disconnected (the reachable pairs are never more than 20%): this is because nodes with the same degree tend to stick together (assortativity), forming a highly connected component and leaving the remaining nodes isolated;
 - for $K \rightarrow \infty$, the power-law distribution of degrees is even more clear-cut, but the number of reachable pairs becomes smaller (no more than 10%); the exponent of the power-law distribution depends on α , with larger α 's yielding larger absolute values of the (negative) exponents;
- the FFBA model (see Fig. 11) increases slightly the number of reachable pairs in all cases; the shape of the power-law distribution is essentially unchanged with respect to the pure FF model;
- finally, for the FFJR model (see Fig. 11) we observe a reduced connectivity; this is due to holding the expected number of edges as a constant, while devoting some of them to closing triangles – an operation that cannot increase connectivity. The degree distribution seems closer to a power-law with respect to the pure FF model.

⁴We observed absolutely analogous phenomena also for larger and denser networks; we hereby report only the smaller case for the sake of readability of the pictures.

6 A real dataset

We considered a dataset of scientific papers⁵ (originally released as part of the 2003 KDD Cup) consisting of 27 770 papers from the “High energy physics (theory) arXiv” database. For each paper (node), we considered as features the words appearing in its title and abstract, excluding those that are dictionary words⁶. The papers are organized in order of publication date.

In Figure 12a the reader can see a fragment of the attribute matrix (for the first 500 nodes and the features they exhibit).

The overall number of features is 21 933, with a matrix density of $0.35 \cdot 10^{-3}$ (there are 214 510 ones in the matrix). The estimated values of α' and β are 15.038 and 0.671, respectively. In particular, we recall that β is the power-law exponent of the asymptotic behavior of L_n , i.e. the overall number of distinct attributes. We show the estimate for this real case in Figure 12b. A reconstruction of the ordering of the nodes according to their fitness parameter values is possible, but we lack any ground truth to compare it to.

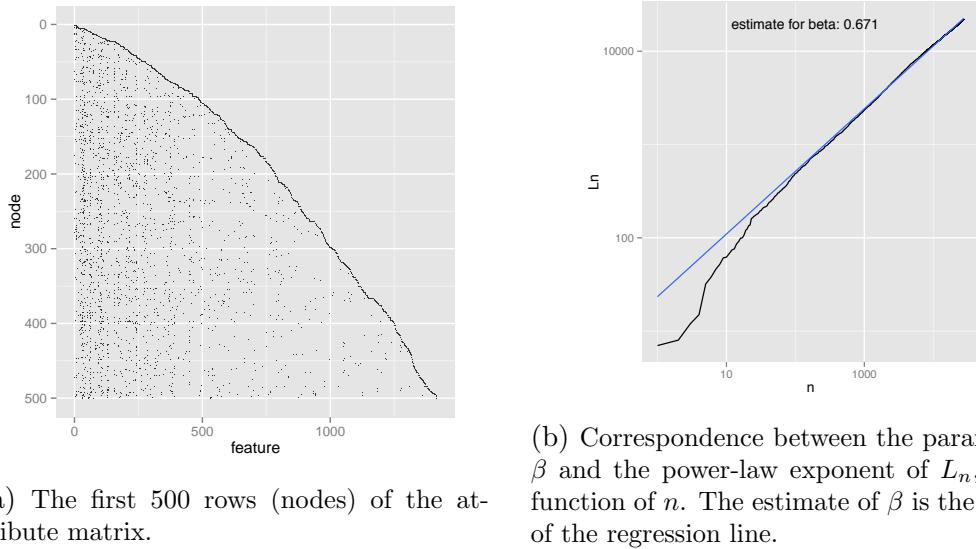


Figure 12: Analysis of the `cit-HepTh` dataset.

We conclude this section with a comparison between the graph produced by the FF model using as the underlying matrix the attribute matrix of the `cit-HepTh` dataset and the corresponding (symmetrized) citation graph. After some experiments, we observed that we can obtain a good fit with $K = 2.5$, that produces a quite similar degree and distance distribution (see Figure 13). It is striking to observe that the two graphs have such a strong similarity in their topology, albeit

⁵The dataset is available within the SNAP (Stanford Large Network Dataset Collection) at <http://snap.stanford.edu/data/cit-HepTh.html>.

⁶According to the Unix `words` dictionary.

having positively no direct relation with each other (in one case the edges represent citations, in the other they were obtained by the model basing on the textual similarity of their abstracts!).

7 Conclusions

In this paper we introduce and study a network model that combines two features:

1. Behind the adjacency matrix of a network there is a *latent attribute structure* of the nodes, in the sense that each node is characterized by a number of features and the probability of the existence of an edge between two nodes depends on the features they share.
2. Not all nodes are equally successful in transmitting their own attributes to the new nodes (*competition*). Each node n is characterized by a random fitness parameter R_n describing its ability to transmit the node's attributes: the greater the value of the random variable R_n , the greater the probability that a feature of n will also be a feature of a new node, and so the greater the probability of the creation of an edge between n and the new node. Consequently, a node's connectivity does not depend on its age alone (so that also “young” nodes are able to compete and succeed in acquiring links).

Our work has different merits: firstly, we propose a simple model for the latent bipartite “node-attribute” network, where the role played by each single parameter is straightforward and easy to interpret: specifically, we have the two parameters, α and β , that control the number of new attributes each new node exhibits (in particular, $\beta > 0$ tunes the power-law behavior of the total number of distinct observed features); whereas the fitness parameters R_i ’s impact on the probability of the new nodes to inherit the attributes of the previous nodes. Secondly, unlike other network models based on the standard Indian Buffet Process, we take into account the aspect of competition and, like in [5], we introduce random fitness parameters so that nodes have a different relevance in transmitting their features to the next nodes; finally, we provide some theoretical, as well experimental, results regarding the power-law behavior of the model and the estimation of the parameters. By experimental data, we also show how the proposed model for the attribute structure naturally leads to a complex network model.

The comparison with real datasets is promising: our model seems to produce quite realistic attribute matrices while at the same time capturing most local and global properties (e.g., degree distributions, connectivity and distance distributions) real networks exhibit.

Some possible future developments are the following. First, we could introduce another parameter $c \geq 0$ in the model of the node-attribute bipartite network so

that the inclusion probabilities are

$$P_n(k) = \frac{\sum_{i=1}^n R_i Z_{i,k}}{c + \sum_{i=1}^n R_i}$$

(we now have $c = 0$): the bigger c , the smaller the inclusion probabilities and so the sparser the attributes. This can allow to obtain attribute matrices that are sparser on the left side. To this purpose, we note that the proofs of the theoretical results change only slightly and so, from a theoretical point of view, we have no problem. The problem is, instead, in the fact that we have an additional parameter to estimate.

Second, a possible variant of the feature/feature (FF) model is to consider, for each incoming new node i , a feature/feature influence matrix $\Xi(i)$ which depends on i : for instance, a diagonal matrix with

$$\xi_{k,k}(i) = \frac{1}{\sum_{\ell=1}^{i-1} Z_{\ell,k}}$$

so that the edge-probability is smaller as the number of nodes with k as a feature is larger.

Acknowledgments

Paolo Boldi and Corrado Monti acknowledge the EU-FET grant NADINE (GA 288956). They also would like to thank Andrea Marino for useful discussions.

Irene Crimaldi acknowledges support from CNR PNR Project “CRISIS Lab”. Moreover, she is a member of the Italian group “Gruppo Nazionale per l’Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA)” of the Italian Institute “Istituto Nazionale di Alta Matematica (INdAM)”.

References

- [1] William Aiello and Fan Chung. Random evolution in massive graphs. In *Proc. of the 42Nd IEEE Symposium on Foundations of Computer Science*, FOCS ’01, pages 510–, Washington, DC, USA, 2001. IEEE Computer Society.
- [2] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, June 2008.
- [3] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.
- [4] Federico Bassetti, Irene Crimaldi, and Fabrizio Leisen. Conditionally identically distributed species sampling sequences. *Advances in Applied Probability*, 42(2):433–459, 06 2010.

- [5] Patrizia Berti, Irene Crimaldi, Luca Pratelli, and Pietro Rigo. Central limit theorems for an indian buffet model with random weights. *The Annals of Applied Probability*, 2014. (forthcoming). Currently available on http://imstat.org/aap/future_papers.html and on arXiv (1304.3626, 2013).
- [6] Ginestra Bianconi and Albert-László Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54(4):436, 2001.
- [7] Ronald L. Breiger. The Duality of Persons and Groups. *Social Forces*, 53(2):181–190, 1974.
- [8] Tamara Broderick, Michael I. Jordan, and Jim Pitman. Beta processes, stick-breaking and power laws. *Bayesian Analysis*, 7(2):439–476, 06 2012.
- [9] George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.
- [10] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, November 2009.
- [11] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.
- [12] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, 1984.
- [13] Charles J. Geyer and Minnesota Univ. (Minneapolis School Of Statistics). *Markov Chain Monte Carlo Maximum Likelihood*. Defense Technical Information Center, 1992.
- [14] Zoubin Ghahramani. Bayesian nonparametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A*, 371(20110553), 2012.
- [15] Thomas L. Griffiths and Zoubin Ghahramani. Infinite latent feature models and the indian buffet process. In *Advances in Neural Information Processing Systems*, pages 475–482. MIT Press, 2005.
- [16] Thomas L. Griffiths and Zoubin Ghahramani. The indian buffet process: An introduction and review. *J. Mach. Learn. Res.*, 12:1185–1224, 2011.
- [17] Peter Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, New York, NY, 1980.
- [18] Peter D. Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*, 15(4):261–272, 2009.

- [19] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, Princeton, NJ, USA, 2008.
- [20] Matthew O. Jackson and Brian W. Rogers. Meeting strangers and friends of friends: How random are social networks? *American Economic Review*, 97(3):890–915, 2007.
- [21] John F. Charles Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.
- [22] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’06, pages 611–617, New York, NY, USA, 2006. ACM.
- [23] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, and Eli Upfal. Stochastic models for the web graph. In *Proc. of the 41st Annual Symposium on Foundations of Computer Science*, FOCS ’00, pages 57–, Washington, DC, USA, 2000. IEEE Computer Society.
- [24] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31 – 48, 2008.
- [25] Silvio Lattanzi and D. Sivakumar. Affiliation networks. In *Proc. of the Forty-first Annual ACM Symposium on Theory of Computing*, STOC ’09, pages 427–434, New York, NY, USA, 2009. ACM.
- [26] Kurt T. Miller, Thomas L. Griffiths, and Michael I. Jordan. Nonparametric latent feature models for link prediction. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *In NIPS*, pages 1276–1284. Curran Associates, Inc., 2009.
- [27] Morten Mørup, Mikkel N. Schmidt, and Lars Kai Hansen. Infinite multiple membership relational modeling for complex networks. *CoRR*, abs/1101.5097, 2011.
- [28] Konstantina Palla, David A. Knowles, and Zoubin Ghahramani. An Infinite Latent Attribute Model for Network Data. In *Proc. of the 29th International Conference on Machine Learning*, pages 1–8, 2012.
- [29] Jim Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006.
- [30] Purnamrita Sarkar, Deepayan Chakrabarti, and Michael I. Jordan. Nonparametric link prediction in dynamic networks. In *Proc. of the 29th International Conference on Machine Learning*, pages 1–8, 2012.

- [31] Tom A.B. Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.
- [32] Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M.P. Ravindra, Elisa Bertino, and Ravi Kumar, editors. *HyperANF: approximating the neighbourhood function of very large graphs on a budget*. ACM, 2011.
- [33] Yee W. Teh and Dilan Gorur. Indian buffet processes with power-law behavior. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1838–1846. Curran Associates, Inc., 2009.
- [34] Romain Thibaux and Michael I. Jordan. Hierarchical beta processes and the indian buffet process. In *Proc. 11th Conference on Artificial Intelligence and Statistics (AISTAT)*, pages 1–8, Puerto Rico, 2007.
- [35] Sebastiano Vigna. A weighted correlation index for rankings with ties. *CoRR*, abs/1404.3325, 2014.

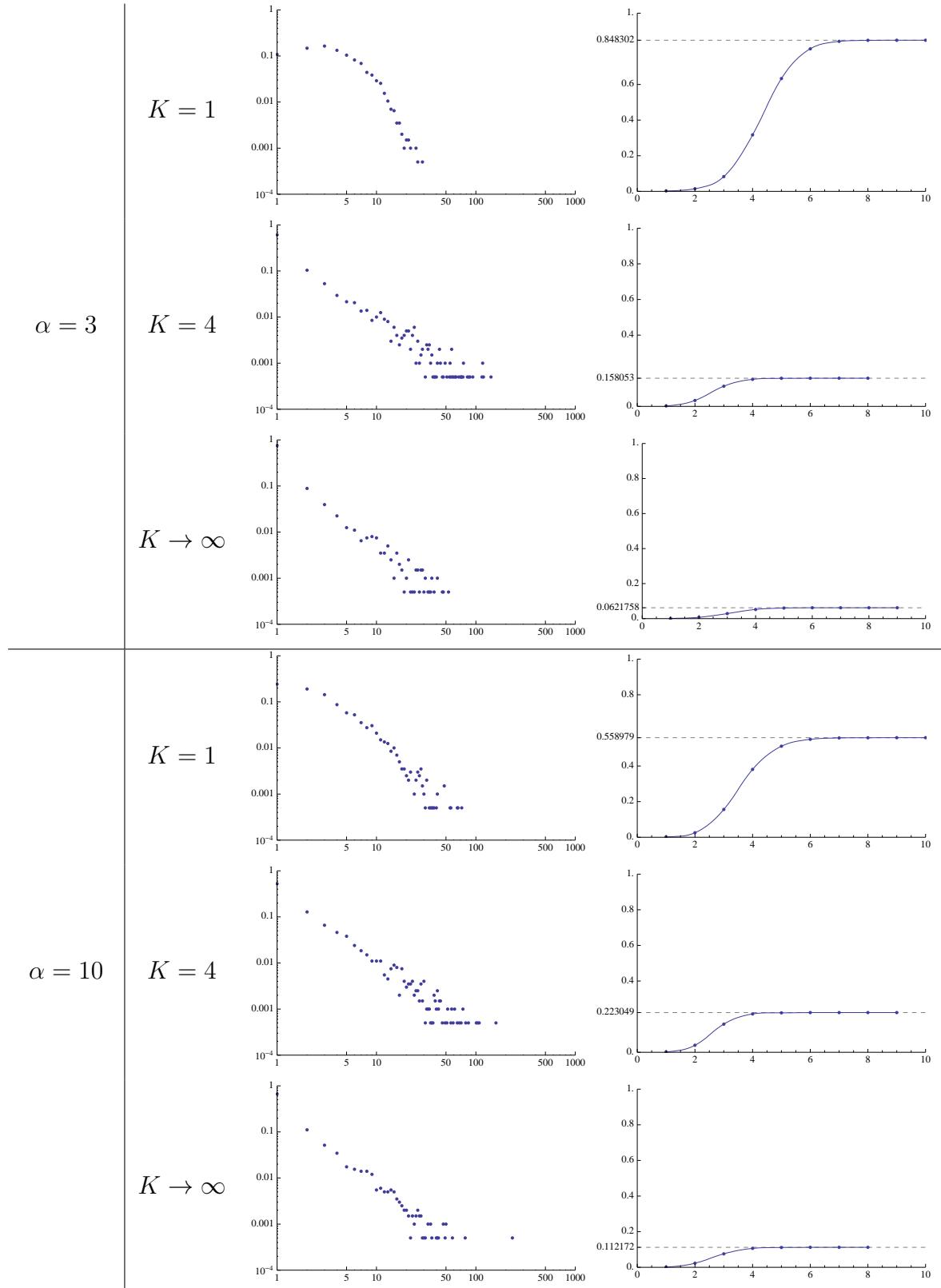


Figure 10: Properties of graphs generated by the FF model. We show the degree distribution in a log-log plot and the fraction of pairs at distance at most k ; in the latter, we highlight the peak value (fraction of mutually reachable pairs). 33

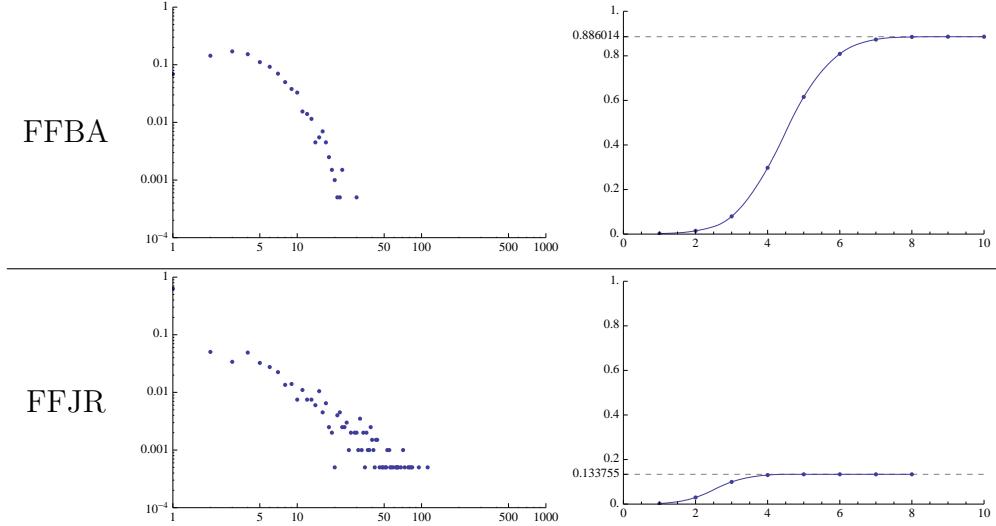


Figure 11: Properties of graphs generated by mixed models with $K = 1$ and $\delta = 0.75$. We show the degree distribution in a log-log plot and the fraction of pairs at distance at most k ; in the latter, we highlight the peak value, indicating how many pairs of nodes are mutually reachable. The parameters of the underlying attribute-matrix model are $\alpha = 3$ and $\beta = 0.75$.

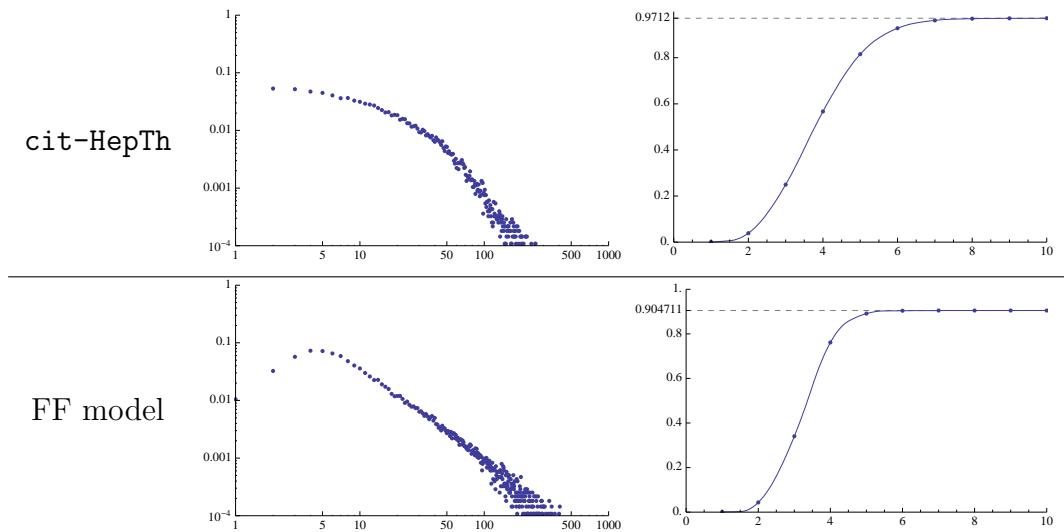


Figure 13: Comparison of the cit-HepTh dataset *versus* a graph generated by the FF model applied to the feature matrix. We show the degree distribution in a log-log plot, and the fraction of pairs at distance at most k ; in the latter, we highlight the peak value, indicating how many pairs of nodes are mutually reachable.

Entity-Linking via Graph-Distance Minimization

Roi Blanco

Yahoo! Research
Barcelona, Spain
roi@yahoo-inc.com

Paolo Boldi

Dipartimento di informatica
Università degli Studi di Milano
paolo.boldi@unimi.it

Andrea Marino*

Dipartimento di informatica
Università degli Studi di Milano
marino@di.unimi.it

Entity-linking is a natural-language-processing task that consists in identifying the entities mentioned in a piece of text, linking each to an appropriate item in some knowledge base; when the knowledge base is Wikipedia, the problem comes to be known as *wikification* (in this case, items are Wikipedia articles). One instance of entity-linking can be formalized as an optimization problem on the underlying concept graph, where the quantity to be optimized is the average distance between chosen items. Inspired by this application, we define a new graph problem which is a natural variant of the Maximum Capacity Representative Set. We prove that our problem is NP-hard for general graphs; nonetheless, under some restrictive assumptions, it turns out to be solvable in linear time. For the general case, we propose two heuristics: one tries to enforce the above assumptions and another one is based on the notion of hitting distance; we show experimentally how these approaches perform with respect to some baselines on a real-world dataset.

1 Introduction

Wikipedia¹ is a free, collaborative, hypertextual encyclopedia that aims at collecting articles on different (virtually, all) branches of knowledge. The usage of Wikipedia for automatically tagging documents is a well-known methodology, that includes in particular a task called *wikification* [13]. Wikification is a special instance of *entity-linking*: a textual document is given and within the document various fragments are identified (either manually or automatically) as being (*named*) *entities* (e.g., names of people, brands, places...); the purpose of entity-linking is assigning a specific reference (a Wikipedia article, in the case of wikification) as a tag to each entity in the document.

Entity-linking happens typically in two stages: in a first phase, every entity is assigned to a set of items, e.g., Wikipedia articles (the *candidate nodes* for that entity); then a second phase consists in selecting a single node for each entity, from within the set of candidates. The latter task, called *candidate selection*, is the topic on which this paper focuses.

To provide a concrete example, suppose that the target document contains the entity “jaguar” and the entity “jungle”. Entity “jaguar” is assigned to a set of candidates that contains (among others) both the Wikipedia article about the feline living in America and the one about the Jaguar car producer. On the other hand, “jungle” is assigned to the article about tropical forests and to the one about the electronic music genre. Actually, there are more than 30 candidates for “jaguar”, and more about 20 for “jungle”.

In this paper, we study an instance of the candidate selection problem in which the selection takes place based on some cost function that depends on the average distance between the selected candidates, where the distance is measured on the Wikipedia graph²: the rationale should be clear enough—concepts

*The second and third authors were supported by the EU-FET grant NADINE (GA 288956).

¹<http://en.wikipedia.org/>

²The undirected graph whose vertices are the Wikipedia articles and whose edges represent hyperlinks between them.

appearing in the same text are related, and so we should choose, among the possible candidates for each entity, those that are more closely related to one another.

Getting back to the example above, there is an edge connecting “jaguar” the feline with “jungle” the tropical forest, whereas the distance between, say, the feline and the music genre is much larger.

The approach we assume here highlights the *collective* nature of the entity-linking problem, as mentioned already in [10]: accuracy of the selection can be improved by a global (rather than local) optimization of the choices. As [10] observes, however, trying to optimize all-pair compatibility is a computationally difficult problem.

In this paper, we prove that the problem itself, even in the simple instance we take into consideration, is NP-hard; however, it becomes efficiently solvable under some special assumptions. We prove that, although these assumptions fail to hold in real-world scenarios, we can still provide heuristics to solve real instances.

We test our proposals on a real-world dataset showing that one of our heuristics is very effective, actually more effective than other methods previously proposed in the literature, and more than a simple greedy approach using the same cost function adopted here.

2 Related Work

Named-entity linking (NEL)- also referred to as *named entity disambiguation* grounds mentions of entities in text (*surface forms*) into some knowledge base (e.g. Wikipedia, Freebase). Early approaches to NEL [13] make use of measures derived from the frequency of the keywords to be linked in the text and in different Wikipedia pages. These include *tf-idf*, χ^2 and *keyphraseness*, which stands for a measure of how much a certain word is used in Wikipedia links in relation to its frequency in general text. Cucerzan [7] employed the context in which words appears and Wikipedia page categories in order to create a richer representation of the input text and candidate entities. These approaches were extended by Milne and Witten [14] who combined commonness (i.e., prior probability) of an entity with its relatedness to the surrounding context using machine learning. Further, Bunescu [4] employed a *disambiguation* kernel which uses the hierarchy of classes in Wikipedia along with its word contents to derive a finer-grained similarity measure between the candidate text and its context with the potential named entities to link to. In this paper we will make use of Kulkarni et al.’s dataset [11]. They propose a general collective disambiguation approach, under the premise that coherent documents refer to entities from one or a few related topics. They introduce formulations that account for the trade-off between local spot-to-entity compatibility and measures of global coherence between entities. More recently, Han et al. [10] propose a graph-based representation which exploits the global interdependence of different linking decisions. The algorithm infers jointly the disambiguated named mentions by exploiting the graph.

It is worth to remark that NEL is a task somehow similar to Word Sense Disambiguation (determining the right sense of a word given its context) in which the role of the knowledge base is played by Wordnet [8]. WSD is a problem that has been extensively studied and its explicitly connection with NEL was made by Hachey et al [9]. WSD has been an area of intense research in the past, so we will review here the approaches that are directly relevant to our work. Graph-based approaches to word sense disambiguation are pervasive and yield state of the art performance [15]; however, its use for NEL has been restricted to ranking candidate named entities with different flavors of centrality measures, such as in-degree or PageRank [9].

Mihalcea [12] introduced an unsupervised method for disambiguating the senses of words using random walks on graphs that encode the dependencies between word senses.

Navigli and Lapata [18, 16, 17] present subsequent approaches to WSD using graph connectivity metrics, in which nodes are ranked with respect to their local *importance*, which is regarded using centrality measures like in-degree, centrality, PageRank or HITS, among others.

Importantly, even if the experimental section of this paper deals with a NEL dataset exclusively, the theoretical findings could be equally applied to WSD-style problems. Our *greedy* algorithm is an adaptation of Navigli and Velardi’s Structural Semantic Interconnections algorithms for WSD [18, 6]. The original algorithm receives an ordered list of words to disambiguate. The procedure first selects the *unambiguous* words from the set (the ones with only one synset), and then for every ambiguous word, it iteratively selects the sense that is *closer* to the sense of disambiguated words, and adds the word to the unambiguous set. This works in the case that a sufficiently connected amount of words is unambiguous; this is not the case in NEL and in our experimental set-up, where there could potentially exist hundreds of candidates for a particular piece of text.

3 Problem statement and NP-completeness

In this section we will introduce the general formal definition of the problem, in the formulation we decided to take into consideration. We will make use of the classical graph notation: in particular, given an undirected graph $G = (V, E)$, we will denote with $G[W]$ the graph induced by the vertices in W , and with $d(u, v)$ the distance between the nodes u and v , that is, the number of edges in the shortest path from u to v (or the sum of the weights of the lightest path, if G is weighted).

If G is a graph and e is an edge of G , $G - e$ is the graph obtained by removing e from G ; we say that e is a *bridge* if the number of connected components of $G - e$ is larger than that of G . A connected bridgeless graph is called *biconnected*; a maximal set of vertices of G inducing a biconnected subgraph is called a *biconnected component* of G .

We call our main problem the *Minimum Distance Representative*, in short MINDR, and we define it as follows. Given an undirected graph $G = (V, E)$ (possibly weighted) and k subsets of its set of vertices, $X_1, \dots, X_k \subseteq V$, a feasible solution for MINDR is a sequence of vertices of G , x_1, \dots, x_k , such that for any i , with $1 \leq i \leq k$, $x_i \in X_i$ (i.e., the solution contains exactly one element from every set, possibly with repetitions).

Given the instance $G, \{X_1, \dots, X_k\}$, the measure (the *distance cost*) of a solution S , x_1, \dots, x_k , is $f(S) = \sum_{i=1}^k \sum_{j=1}^k d(x_i, x_j)$. The goal is finding the solution of minimum distance cost, i.e., a feasible solution S such that $f(S)$ is minimum.

We call the restriction of this problem, in which the sets of vertices in input $\{X_1, \dots, X_k\}$ are disjoint, MINDIR (Minimum Independent Distance Representative). In this case, for the sake of simplicity, we will refer to a solution as the multiset composed by its elements.³

3.1 NP-completeness of MINDR

The MINDIR problem seems to be similar and related to the so-called Maximum Capacity Representatives [5], in short MAXCRS. The Maximum Capacity Representatives problem is defined as follows: given some disjoint sets X_1, \dots, X_m and for any $i \neq j$, $x \in X_i$, and $y \in X_j$, a nonnegative capacity $c(x, y)$, a

³We shall make free use of multiset membership, intersection and union with their standard meaning: in particular, if A and B are multisets with multiplicity function a and b , respectively, the multiplicity functions of $A \cup B$ and $A \cap B$ are $x \mapsto \max(a(x), b(x))$ and $x \mapsto \min(a(x), b(x))$, respectively.

solution is a set $S = \{x_1, \dots, x_m\}$, such that, for any i , $x_i \in X_i$; such a solution is called *system of representatives*. The measure of a solution is the capacity of the system of representatives, that is $\sum_{x \in S} \sum_{y \in S} c(x, y)$, and the MAXCRS problem aims at *maximizing* it. The MAXCRS problem was introduced by [1], who showed that it is NP-complete and gave some non-approximability results. Successively, in [19], tight inapproximability results for the problem were presented.

The MINDIR problem differs from MAXCRS just for in the sense that we are dealing with distances instead of capacities, and therefore we ask for a minimum instead of a maximum. Nonetheless the following Lemma, whose proof is given in Appendix A, shows that also MINDIR problem is NP-complete.

Lemma 1. *The MINDIR (hence, MINDR) problem is NP-complete.*

4 The decomposable case

In this section we study the MINDR problem under some restrictive hypothesis and we will show that in this case a linear exact algorithm exists.

Even if it may seem that these hypothesis are too strong to make the algorithm useful in practice, in the next section we will use our algorithm to design an effective heuristic for the general problem. In particular, we assume that the graph G (possibly weighted) is such that:

- any set X_i induces a connected subgraph on G , i.e., $G[X_i]$ is connected,
- for any $i \neq j$, for any $x \in X_i$ and $y \in X_j$, x and y do not belong to the same biconnected component.

The problem, under these further restrictions, will be called *decomposable* MINDR. Note that the second condition implies that a decomposable MINDR is in fact an instance of MINDIR, because it implies that no two sets can have nonempty intersection.

Let us consider an instance $(G, \{X_1, \dots, X_k\})$ of decomposable MINDR problem on a graph $G = (V, E)$.

An edge $e = (x, y) \in E$ is called *useful* if it is a bridge, x and y do not belong to the same set X_i , and there are at least two indices i and j such that X_i and X_j are in different components of $G - e$ (since e is a bridge, the graph obtained removing the edge e from G is no more connected).

4.1 Decomposing the problem

The main trick that allows to obtain a linear-time solution for the decomposable case is that we can actually decompose the problem (hence the name) through useful edges. First observe that, trivially:

Remark 1. *Let $e = (x, y)$ be a useful edge and let Z_x and Z_y be the two connected components of $G - e$ containing x and y , respectively. In G , all paths from any $x' \in Z_x$ to any $y' \in Z_y$ must contain e .*

Moreover:

Remark 2. *Let $e = (x, y)$ be a useful edge. There cannot be an index i such that X_i has a nonempty intersection with both components of $G - e$.*

In fact, assume by contradiction that one such X_i exists, and let $u, w \in X_i$ be two vertices living in the two different components of $G - e$: since $G[X_i]$ is connected, there must be a path connecting u and w and made only of elements of X_i ; because of Remark 1, this path passes through e , but this would imply that $x, y \in X_i$, in contrast with the definition of useful edge.

Armed with the previous observations, we can give the following further definitions. Let Y_x (respectively, Y_y) be the set of sets X_i such that $X_i \subseteq Z_x$ (respectively, $X_i \subseteq Z_y$); we denote the sets of nodes in Y_x and Y_y by $V(Y_x) \subseteq Z_x$ and $V(Y_y) \subseteq Z_y$, respectively.

By virtue of Remark 1, all the paths in G from any $x' \in V(Y_x)$ to any $y' \in V(Y_y)$ pass through e . This implies also that there is no simple cycle in the graph including both $x' \in V(Y_x)$ and $y' \in V(Y_y)$.

Given a solution S for $\text{MINDIR}(G, \{X_1, \dots, X_k\})$, and a useful edge (x, y) , we have:

$$\begin{aligned} \sum_{x_i, x_j \in S} d(x_i, x_j) &= \sum_{x_i, x_j \in S \cap V(Y_x)} d(x_i, x_j) + \sum_{x_i, x_j \in S \cap V(Y_y)} d(x_i, x_j) + \\ &\quad 2 \sum_{x_i \in S \cap V(Y_x), x_j \in S \cap V(Y_y)} (d(x_i, x) + d(x, y) + d(y, x_j)). \end{aligned}$$

Indeed all the shortest paths from any $x_i \in S \cap V(Y_x)$ to any $x_j \in S \cap V(Y_y)$ pass through the useful edge (x, y) by Remark 1. Moreover, since the sets X_1, \dots, X_k are disjoint, we have that $|S \cap V(Y_x)| = |Y_x|$ and $|S \cap V(Y_y)| = |Y_y|$, that is, a solution has exactly one element for each set in Y_x (respectively, Y_y). Hence we can rewrite the last summand of the above equation as follows:

$$\begin{aligned} \sum_{x_i \in S \cap V(Y_x), x_j \in S \cap V(Y_y)} (d(x_i, x) + d(y, x_j) + d(x, y)) &= |Y_y| \cdot \sum_{x_i \in S \cap V(Y_x)} d(x_i, x) + \\ &\quad |Y_x| \cdot \sum_{x_j \in S \cap V(Y_y)} d(y, x_j) + \\ &\quad |Y_x| \cdot |Y_y| \cdot d(x, y). \end{aligned}$$

By combining the two equations, we can conclude that finding a solution for $\text{MINDIR}(G, \{X_1, \dots, X_k\})$ can be decomposed into the following two subproblems:

1. finding S_x minimizing $\sum_{x_i, x_j \in S \cap V(Y_x)} d(x_i, x_j) + 2 \sum_{x_i \in S \cap V(Y_x)} |Y_y| d(x_i, x)$ in the instance $(G[Z_x], Y_x)$;
2. finding S_y minimizing $\sum_{x_i, x_j \in S \cap V(Y_y)} d(x_i, x_j) + 2 \sum_{x_j \in S \cap V(Y_y)} |Y_x| d(y, x_j)$ in the instance $(G[Z_y], Y_y)$.

Note that both instances are smaller than the original one because of the definition of a useful edge. The idea of our algorithm generalizes this principle; note that the new objective function we must take into consideration is slightly more complex than the original one: in fact, besides the usual all-pair-distance cost there is a further summand that is a weighted sum of distances from some fixed nodes (such as x for the instance $G[Z_x], Y_x$ and y for the instance $G[Z_y], Y_y$).

We hence define an extension of the MINDR problem, that we call **EXTMINDR** (for *Extended Minimum Distance Representatives*). In this problem, we are given:

- an undirected graph $G = (V, E)$ (possibly weighted)
- k subsets of its set of vertices, $X_1, \dots, X_k \subseteq V$
- a multiset B of vertices, each $x \in B$ endowed with a weight $b(x)$.

A feasible solution for the EXTMINDR is a multiset $S = \{x_1, \dots, x_k\}$ of vertices of G , such that for any i , with $1 \leq i \leq k$, $S \cap X_i \neq \emptyset$ (i.e., the set contains at least one element from every set). Its cost is

$$f(S) = \sum_{i=1}^h \sum_{j=1}^k d(x_i, x_j) + \sum_{i=1}^k \sum_{z \in B} b(z) d(x_i, z).$$

The goal is finding the solution of minimum cost, i.e., a feasible solution S such that $f(S)$ is minimum. The original version of the problem is obtained by letting $B = \emptyset$.

We are now ready to formalize our decomposition through the following Theorem, whose proof is given in Appendix B.

Theorem 1. Let us be given a decomposable EXTMINDR instance $(G, \{X_1, \dots, X_k\}, B, b)$ and a useful edge $e = (t_0, t_1)$. For every $s \in \{0, 1\}$, let Z_s be the connected component of $G - e$ containing t_s , Y_s be the set of sets X_i such that $X_i \subseteq Z_s$ and $V(Y_s)$ be the union of those X_i 's. Let also B_s be the intersection of B with Z_s . Define a new instance $I_s = (T[Z_s], \{X_i, i \in Y_s\}, B_s \cup \{t_s\}, b_s)$ where

$$b_s(t_s) = 2|Y_{1-s}| + \sum_{z \in B_{1-s}} b(z) \text{ and } b_s(z) = b(z), \text{ for any } z \in B.$$

Then the cost $f(S)$ of an optimal solution S of the original problem is equal to

$$f(S_0) + f(S_1) + 2|Y_0||Y_1|d(t_0, t_1) + \sum_{s \in \{0, 1\}} \left(|S \cap V(Y_s)| \cdot \sum_{z \in B \cap Z_{1-s}} b(z)d(t_s, z) \right)$$

where S_s is an optimal solution for the instance I_s .

For completeness, we need to consider the base case of an instance with just one set $G, \{X_1\}, B, b$: the solution in this case is just one node $x \in X_1$ and the objective function to be minimized is simply $\sum_{z \in B} d(x, z)b(z)$. The optimal solution can be found by performing a BFS from every $z_j \in B$ (in increasing order of j), maintaining for each node $y \in X_1$, $g(y) = \sum_{z_t \in B, t < j} d(x, z_t)b(z_t)$, and picking the node having maximum final $g(y)$. This process takes $O(|B| \cdot |E(G[X_1])|)$. It is worth observing that in our case the size of the multiset B is always bounded by k . Moreover since $\sum_{i=1}^k |E(G[X_i])| \leq |E(G)| = m$, the overall complexity for all these base cases is bounded by $O(k \cdot m)$.

4.2 Finding useful edges

For every instance with more than one set, given an useful edge e the creation of the subproblems as described above is linear, so we are left with the issue of finding useful edges. This task can be seen as a variant of the standard depth-first search of bridges, as shown in Algorithm 2 and 3, in Appendix C. Recall that bridges can be found by performing a standard DFS that numbers the nodes as they are found (using the global counter visited, and keeping the DFS numbers in the array dfs); every visit returns the index of the least ancestor reachable through a back edge while visiting the DFS-subtree rooted at the node where the visit starts from. Every time a DFS returns a value that is larger than the number of the node currently being visited, we have found a bridge.

The variant consists in returning not just the index of the least ancestor reachable, but also the set of indices i that are found while visiting the subtree. If the set of indices and its complement are both different from \emptyset then the bridge is useful: at this point, a “rapid ascent” is performed to get out of the recursive procedure.

4.3 The final algorithm

Combining the observations above, we can conclude that the overall complexity of the algorithm is $O(k \cdot m)$. The algorithm is presented in Algorithm 1.

5 The general case

As we observed at the beginning, the MINDR problem is NP-complete in general, although the decomposable version turns out to be linear. We want to discuss how we can deal with a general instance of the problem. To start with, let us consider a general connected MINDR instance, that is:

Algorithm 1: DECOMPOSABLEMINDR

Input: A graph $G = (V, E)$, $X_1, \dots, X_k \subseteq V$, a weighted multiset B of nodes in V , where each element in B has a weight b . $G[X_i]$ is connected for every i and moreover for all $i \neq j$ and $x \in X_i$, $y \in X_j$, the two vertices x and y do not belong to the same biconnected component of G .

Output: A solution $S = \{x_1, \dots, x_k\}$ such that for any i , with $1 \leq i \leq k$, $x_i \in X_i$, minimizing $\sum_{i=1}^k \sum_{j=1}^k d(x_i, x_j) + \sum_{i=1}^k \sum_{z \in B} b(z)d(x_i, z)$

Find a useful edge $e = (x, y)$, if it exists, using Algorithm 2

if the useful edge does not exist **then**

- if** $k \neq 1$ **then**
- I** Fail!
- end**
- Output the element $x_1 \in X_1$ minimizing $\sum_{z \in B} b(z)d(x_1, z)$

else

- Let Z_x (respectively Z_y) be the connected component of $T - e$ containing x (respectively y).
- Let Y_x (respectively Y_y) be the indices i such that $X_i \subseteq Y_x$ ($X_i \subseteq Y_y$, respectively)
- $B' \leftarrow B \cup \{x\}$ (multiset union) with $b(x) = 2|Y_y| + \sum_{z \in B \cap Z_y} b(z)$
- $B' \leftarrow B' \cap Z_x$ (multiset intersection)
- $S' \leftarrow \text{DECOMPOSABLEMINDR}(T[Z_x], Y_x, B')$
- $B'' \leftarrow B \cup \{y\}$ (multiset union) with $b(y) = 2|Y_x| + \sum_{z \in B \cap Z_x} b(z)$
- $B'' \leftarrow B'' \cap Z_y$ (multiset intersection)
- $S'' \leftarrow \text{DECOMPOSABLEMINDR}(T[Z_y], Y_y, B'')$
- return** $S' \cup S''$

end

- a connected undirected (possibly weighted) graph $G = (V, E)$,
- k subsets of its set of vertices, $X_1, \dots, X_k \subseteq V$,

with the additional assumption that $G[X_i]$ is connected for every i . Recall that a feasible solution is a sequence S of vertices of G , x_1, \dots, x_k , such that for any i , with $1 \leq i \leq k$, we have $x_i \in X_i$; its (distance) cost is $f(S) = \sum_{i=1}^k \sum_{j=1}^k d(x_i, x_j)$.

We shall discuss two heuristics to approach this problem: the first is related to Algorithm 1 in that it tries to modify the problem to make it into a decomposable one, whereas the second is based on the notion of hitting distance.

Before describing the two heuristics, let us briefly explain the rationale behind the additional assumption (i.e., that every $G[X_i]$ be connected). In our main application (entity-linking) the structure of the graph within each X_i is not very important, and can actually be misleading: a very central node in a large candidate set may seem very promising (and may actually minimize the distance to the other sets) but can be blatantly wrong. It is pretty much like the distinction between nepotistic and non-nepotistic links in PageRank computation: the links *within* each host are not very useful in determining the importance of a page—on the contrary, they may be confusing, and are thus often disregarded.

Based on this observation, we can (and probably want to) modify the structure of the graph within each set X_i to avoid this kind of trap. This is done by preserving the *external* links (those that connect vertices of X_i to the outside), but at the same time adding or deleting edges within each X_i in a suitable way. In our experiments, we considered two possible approaches:

- one consists in making $G[X_i]$ *maximally connected*, i.e., transforming it into a clique;
- the opposite approach makes $G[X_i]$ *minimally connected* by adding the minimum number of edges needed to that purpose; this can be done by computing the connected components of $G[X_i]$ and then adding enough edges to join them in a single connected component.

Both approaches guarantee that $G[X_i]$ is connected, so that the two heuristics described below can be applied.

5.1 The spanning-tree heuristic

The first heuristic aims at modifying the graph G in such a way that the resulting instance becomes decomposable. For the moment, let us assume that the sets X_i are pairwise disjoint. To guarantee that the problem be decomposable, we proceed as follows. Define an equivalence relation \sim on V by letting $x \sim y$ whenever x and y belong to the same X_i .⁴ The quotient graph $G/\sim = (V/\sim, E/\sim)$ has vertices V/\sim and an edge between $[x]$ and $[y]$ whenever there is some edge $(x', y') \in E$ with $x' \sim x$ and $y' \sim y$ (here, and in the following, $[x]$ denotes the \sim -equivalence class including x). Thus, there is a surjective (but not injective) map $\iota : E \rightarrow E/\sim$.

Since G is connected, so is G/\sim , and we perform a breadth-first traversal of G building a spanning tree T . Every tree edge is an edge of G/\sim , so its pre-image with respect to ι is a nonempty set of edges in G . Let us arbitrarily choose one edge of G from $\iota^{-1}(t)$ for every tree edge t , and let T' be the resulting set of edges of G .

Define the new graph $G' = (V, E')$ where $E' = T' \cup \bigcup_{i=1}^k E(G[X_i])$: this graph contains all the edges within each set X_i , plus the set T' of external edges.

It is easy to see that $G'[X_i]$ is connected (it is in fact equal to $G[X_i]$), and moreover all the elements of T' are bridges dividing all the X_i 's in distinct biconnected components. In other words, we have turned the instance into a *decomposable* one, where Algorithm 1 can be run.

The non-disjoint case If the sets X_i are not pairwise disjoint, we can proceed as follows. Let us define maximal mutually disjoint sets of indices $I_1, \dots, I_h \subseteq \{1, \dots, k\}$ such that for all $t \neq s$, $\cup_{i \in I_t} X_i \cap \cup_{i \in I_s} X_i = \emptyset$.

Now, take the new problem instance with the same graph and sets Y_1, \dots, Y_h where $Y_t = \cup_{i \in I_t} X_i$: this instance is disjoint, so the previous construction applies. The only difference is that, at the very last step of Algorithm 1, when we are left with a graph and a *single* Y_t , we will not select a single $y \in Y_t$ optimizing the cost function

$$\sum_{z \in B} b(z) d(y, z).$$

Rather, we will choose one element x_i for every $i \in I_t$ optimizing

$$\sum_{i \in I_t} \sum_{z \in B} b(z) d(x_i, z).$$

Discussion Both steps presented above introduce some level of imprecision, that make the algorithm only a heuristic in the general case. The first approximation is due to the fact that building a tree on G will produce distances (between vertices living in different X_i) much larger than they are in G ; the second approximation is that when we have non-disjoint sets, we only optimize with respect to bridges, disregarding the sum of distances of the nodes of different sets. Actually, we should optimize

$$\sum_{i \in I_t} \sum_{j \in I_t} d(x_i, x_j) + \sum_{i \in I_t} \sum_{z \in B} b(z) d(x_i, z).$$

but this would make the final optimization step NP-complete.

⁴Note that, since the sets X_i are pairwise disjoint, \sim is transitive.

5.2 The hitting-distance heuristic

The second heuristic we propose is based on the notion of *hitting distance*: given a vertex x and a set of vertices Y , define the hitting distance of x to Y as $d(x, Y) = \min_{y \in Y} d(x, y)$. The hitting distance can be easily found by a breadth-first traversal starting at x and stopping as soon as an element of Y is hit. Given a general connected instance of MINDR, as described above, we can consider, for every i and every $x \in X_i$, the average hitting distance of x to the other sets:

$$\frac{\sum_{j=1}^k d(x, X_j)}{k}.$$

The element $x_i^* \in X_i$ minimizing the average hitting distance (or any such an element, if there are many) is the candidate chosen for the set X_i in that solution.

The main problem with this heuristic is related to its locality (optimization is performed separately for each X_i); moreover the worst-case complexity is $O(m \sum_i |X_i|)$, that reduces to $O(k \cdot m)$ only under the restriction that the sets X_i have $O(1)$ size.

6 Experiments

All our experiments were performed on a snapshot of the English portion of Wikipedia as of late February 2013; the graph (represented in the BVGraph format [3]) was symmetrized and only the largest component was kept. The undirected graph has 3 685 351 vertices (87.2% of the vertices of the original graph) and 36 066 162 edges (99.9% of the edges of the original graph). Such a graph will be called the “Wikipedia graph” and referred to as G throughout this experimental section.

Our experiments use actual real-world entity-linking problems for which we have a human judgment, and tries the two heuristics proposed in Section 5, as well as a greedy baseline and other heuristics.

The greedy baseline works as follows: it first chooses an index i at random, and draws an element $x_i \in X_i$ also at random. Then, it selects a vertex of $x_{i+1} \in X_{i+1}, x_{i+2} \in X_{i+2}, \dots, x_k \in X_k, x_1 \in X_1, \dots, x_{i-1} \in X_{i-1}$ (in this order) minimizing each time the sum of the distances to the previously selected vertices; the greedy algorithm continues doing the same also for $x_i \in X_i$ to get rid of the only element (the first one) that was selected completely at random. Moreover we have considered also two other heuristics, that have been observed to be effective in practice [9]: these are *degree* and *PageRank based*. They respectively select the highest degree and the highest PageRank vertex for each set.

The real-world entity-linking dataset has been taken from [11] which contains a larger number of human-labelled annotations. For retrieving the candidates, we created an index over all Wikipedia pages with different fields (title, body, anchor text) and used a variant of BM25F [2] for ranking, returning the top 100 scoring candidate entities. Since the candidate selection method was the same for every graph-based method employed, there should be no bias in the experimental outcomes.

The problem instances contained in the dataset have 11.73 entities on average (with a maximum of 53), and the average number of candidates per entity is 95.90 (with a maximum of 200). Each of the 100 problem instances in the NEL dataset is annotated, and in particular, for every i there is a subset $X_i^* \subseteq X_i$ of *fair* vertices (that is, vertices that are good candidates for that set): typically $|X_i^*| = 1$. Note that, for every instance in the NEL dataset, we deleted the sets X_i such that X_i^* were not included in the largest connected component of the Wikipedia graph. The number of sets X_i deleted was at maximum 2 (for two instances). We have not considered instances in which, after these modifications, we have just one set X_i : this situation happened in 5 cases. So the problem set on which we actually ran our algorithm contains 95 instances.

HEURISTIC	DISTANCE-COST RATIO		VALUE	
	MAXIMAL CONNECTION	MINIMAL CONNECTION	MAXIMAL CONNECTION	MINIMAL CONNECTION
	Average (\pm Std Error)	Average (\pm Std Error)	Average (\pm Std Error)	Average (\pm Std Error)
Spanning-tree	122.747 (\pm 2.812)	130.998 (\pm 2.917)	0.369 (\pm 0.023)	0.360 (\pm 0.023)
Hitting-distance	103.945 (\pm 1.320)	105.797 (\pm 2.322)	0.454 (\pm 0.027)	0.459 (\pm 0.027)
Greedy	101.969 (\pm 0.429)	102.785 (\pm 0.426)	0.428 (\pm 0.025)	0.426 (\pm 0.026)
Degree based	114.182 (\pm 2.386)	113.285 (\pm 2.305)	0.411 (\pm 0.024)	0.394 (\pm 0.023)
PageRank based	114.894 (\pm 2.452)	112.392 (\pm 2.266)	0.407 (\pm 0.025)	0.398 (\pm 0.023)
GROUND TRUTH	115.117 (\pm 1.782)	119.243 (\pm 1.873)		

Table 1: Distance-cost ratio and value.

For every instance, we considered the maximal and minimal connection⁵ approach, and then ran both heuristics described in Section 5, comparing them with the greedy baseline, and also with the degree and PageRank heuristics.

For any instance, when comparing the distance cost f of the solutions S_j returned by some algorithm A_j , we have computed the *distance-cost ratio* of each algorithm A_j , defined as

$$\frac{f(S_j)}{\min_j f(S_j)} \cdot 100.$$

Intuitively this corresponds to the approximation ratio of each solution with respect to the best solution found by all the considered algorithms: hence the best algorithm has minimum distance-cost ratio and it equals 100.

Besides evaluating the distance cost of the solutions found by the various heuristics, we can compute how many of the elements found are fair: we normalize this quantity by k , so that 1.0 means that all the k candidates selected are fair. We call such a quantity the *value* of a solution.

In the last two columns of Table 1 we report, for each heuristic, the average value (across all the instances) along with the standard error. For both the connection approaches, we have that the hitting-distance heuristic outperforms all the other heuristics, and it selects more than 45% of fair candidates. The variability of the results seems not to differ too much for all the methods. The second best heuristic is the greedy baseline, that selects almost 42.8% and 42.6% fair candidates respectively in a maximal and minimal connected scenario.

It is worth observing that the greedy approach comes second (as far as the value is concerned), and outperforms the baseline techniques (degree and PageRank). The spanning tree heuristic, instead, perform worse than any other method.

The latter outcome is easily explained by the fact that it transforms completely the topology of the graph in order to make the instance decomposable, and the distances between vertices are mostly scrambled. This interpretation of the bad result obtained can also be seen looking at the distance cost (central columns of Table 1): the spanning-tree heuristic is the one that is less respectful of distances, selecting candidates that are far apart from one another.

In the central columns of Table 1, we report also the distance-cost ratio for all the other heuristics. For both the maximal and the minimal connection approaches, the greedy baseline seems to obtain more

⁵To obtain the minimal connection of each $G[X_i]$, we chose to connect the vertex of maximum degree of its largest component with an (arbitrary) vertex of each of its remaining components.

often a minimum distance cost solution. The second best option is the hitting distance heuristic, while the other methods seems to be more far away from an optimal result.

In the last row of Table 1, we report the distance-cost ratio for the ground-truth solution given by the fair candidates. It seems that for any instance, the ground truth has distance cost averagely 15%-20% higher than the best solution we achieve by using the heuristics. This observation suggests that probably our objective function (that simply aims at minimizing the graph distances) is too simplistic: the distance cost is an important factor to be taken into account but certainly not the unique one.

It is interesting to remark, though, that the average Jaccard coefficient between the solution found by the degree based and the hitting-distance heuristic is 0.3 (for both maximal and minimal connection approaches): this fact means that the degree and distance can be probably used as complementary features that hint at different good candidates, although we currently do not know how to combine these pieces of information.

Finally, we remark that we also tried to apply the degree and PageRank based heuristics by using the same problem set but *in the original directed graph*; in this case, we did not enforce any connectivity of the subgraphs $G[X_i]$: the resulting average values (\pm standard error) are respectively 0.327 (± 0.020) and 0.336 (± 0.022), and they are both worse than the values achieved by degree and PageRank heuristics in Table 1. This fact suggests that our experimental approach (of considering the undirected version and of enforcing some connectivity on the subgraphs) not only guarantees the applicability of our heuristics in a more suitable scenario, but also improves the effectiveness of the other existing techniques.

7 Conclusions and future work

Inspired by the entity-linking task in NLP, we defined and studied a new graph problem related to Maximum Capacity Representative Set and we proved that this problem is NP-hard in general (although it remains an open problem to determine its exact approximability). Moreover, we showed that the problem can be solved efficiently in some special case, and that we can anyway provide reasonable heuristics for the general scenario. We tested our proposals on a real-world dataset showing that one of our heuristics is very effective, actually more effective than other methods previously proposed in the literature, and more than a simple greedy approach using the same cost function adopted here.

The other heuristic proposed in this paper seem to work poorly (albeit it reduces to a case where we know how to produce the optimal solution), but we believe that this is just because of the very rough preprocessing phase it adopts; we plan to devise a more refined way to induce the conditions needed for Algorithm 1 to work, without having to resort to the usage of a spanning tree—the latter scrambles the distances too much, resulting in a bad selection of candidates.

Finally, we observed that a distance-based approach is complementary to other methods (e.g., the local techniques based solely on the vertex degree), hinting at the possibility of obtaining a new, better cost function that exploits both features at the same time.

References

- [1] Mihir Bellare (1993): *Interactive Proofs and Approximation: Reduction from Two Provers in One Round*. In: *ISTCS*, pp. 266–274, doi:10.1109/ISTCS.1993.253462.
- [2] Roi Blanco & Paolo Boldi (2012): *Extending BM25 with Multiple Query Operators*. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12*, ACM, New York, NY, USA, pp. 921–930, doi:10.1145/2348283.2348406.

- [3] Paolo Boldi & Sebastiano Vigna (2004): *The WebGraph Framework I: Compression Techniques*. In: Proc. of the Thirteenth International World Wide Web Conference (WWW 2004), ACM Press, Manhattan, USA, pp. 595–601, doi:10.1145/988672.988752.
- [4] Razvan C. Bunescu & Marius Pasca (2006): *Using Encyclopedic Knowledge for Named entity Disambiguation*. In: EACL, The Association for Computer Linguistics.
- [5] Pierluigi Crescenzi & Viggo Kann (1997): *Approximation on the Web: A Compendium of NP Optimization Problems*. In: RANDOM, pp. 111–118, doi:10.1007/3-540-63248-4_10.
- [6] Montse Cuadros & German Rigau (2008): *KnowNet: A proposal for building highly connected and dense knowledge bases from the web*. In: First Symposium on Semantics in Systems for Text Processing, pp. 71–84, doi:10.3115/1626481.1626488.
- [7] Silviu Cucerzan (2007): *Large-scale named entity disambiguation based on Wikipedia data*. In: In Proc. 2007 Joint Conference on EMNLP and CNLL, pp. 708–716.
- [8] Christiane Fellbaum, editor (1998): *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, doi:10.2307/417141.
- [9] Ben Hachey, Will Radford & James R. Curran (2011): *Graph-based Named Entity Linking with Wikipedia*. In: Proceedings of the 12th International Conference on Web Information System Engineering, WISE’11, Springer-Verlag, Berlin, Heidelberg, pp. 213–226, doi:10.1007/978-3-642-24434-6_16.
- [10] Xianpei Han, Le Sun & Jun Zhao (2011): *Collective Entity Linking in Web Text: A Graph-based Method*. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’11, ACM, pp. 765–774, doi:10.1145/2009916.2010019.
- [11] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan & Soumen Chakrabarti (2009): *Collective annotation of Wikipedia entities in web text*. In: Knowledge Discovery and Data Mining, pp. 457–466, doi:10.1145/1557019.1557073.
- [12] Rada Mihalcea (2005): *Unsupervised Large-vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling*. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT ’05, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 411–418, doi:10.3115/1220575.1220627.
- [13] Rada Mihalcea & Andras Csomai (2007): *Wikify!: Linking Documents to Encyclopedic Knowledge*. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM ’07, ACM, New York, NY, USA, pp. 233–242, doi:10.1145/1321440.1321475.
- [14] David Milne & Ian H. Witten (2008): *Learning to Link with Wikipedia*. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM ’08, ACM, New York, NY, USA, pp. 509–518, doi:10.1145/1458082.1458150.
- [15] Roberto Navigli (2009): *Word sense disambiguation: a survey*. ACM COMPUTING SURVEYS 41(2), pp. 1–69, doi:10.1145/1459352.1459355.
- [16] Roberto Navigli & Mirella Lapata (2007): *Graph Connectivity Measures for Unsupervised Word Sense Disambiguation*. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI’07, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1683–1688.
- [17] Roberto Navigli & Mirella Lapata (2010): *An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation*. IEEE Trans. Pattern Anal. Mach. Intell. 32(4), pp. 678–692, doi:10.1109/TPAMI.2009.36.
- [18] Roberto Navigli & Paola Velardi (2005): *Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation*. IEEE Trans. Pattern Anal. Mach. Intell. 27(7), pp. 1075–1086, doi:10.1109/TPAMI.2005.149.
- [19] Maria Serna, Luca Trevisan & Fatos Xhafa (2005): *The approximability of non-Boolean satisfiability problems and restricted integer programming*. Theoretical Computer Science 332(13), pp. 123 – 139, doi:10.1016/j.tcs.2004.10.014.

A Proof of Lemma 1

Proof. We reduce MAXCRS to MINDIR. Given an instance of MAXCRS, $\{X_1, \dots, X_k\}$ and for any $i \neq j$, $x \in X_i$, and $y \in X_j$, a nonnegative capacity $c(x, y)$, we construct the instance of MINDIR $G, \{X_1, \dots, X_k\}$; the vertices of G are $X_1 \cup \dots \cup X_k$, and for any pair $x \in X_i$, $y \in X_j$, with $i \neq j$, we add a weighted edge between x and y , i.e., for each pair for which MAXCRS defines a capacity we create a corresponding edge in G . In particular the weight of the edge between x and y is set to $\alpha - c(x, y)$, where $\alpha = 2 \max_{z \in X_i, t \in X_j, i \neq j} c(z, t)$.

Observe that for any pair of nodes $u \in X_i$, $v \in X_j$, with $i \neq j$, $d(u, v)$ in G is equal to the weight of (u, v) , i.e., it is not convenient to pass through other nodes when going from u to v : in fact, for any path z_1, \dots, z_p from u to v in G , with $p \geq 1$, we always have $\alpha - c(u, v) \leq \alpha - c(u, z_1) + \dots + \alpha - c(z_p, v)$, since $\alpha - c(u, v) \leq \alpha$ and the weight of such a path is at least $\frac{p+1}{2}\alpha \geq \alpha$. Moreover, observe that any optimal solution in G has exactly one element for each set X_i : thus, we have $k(k-1)$ pairs of elements (x, y) , whose distance is always given by the weight of the single edge (x, y) , that is $\alpha - c(x, y)$.

Hence it is easy to see that MAXCRS admits a system of representatives whose capacity is greater than h , if and only if MINDIR admits a solution S such that $f(S)$ is less than $k(k-1)\alpha - h$.

Since MINDIR is a restriction of MINDR we can conclude that also MINDR is NP-complete. \square

B Proof of Theorem 1

Proof. We can rewrite the objective function as follows.

$$\begin{aligned} \sum_{x_i, x_j \in S} d(x_i, x_j) + \sum_{x_i \in S} \sum_{z \in B} d(x_i, z) b(z) &= 2|Y_0| |Y_1| d(t_0, t_1) + \sum_{x_i, x_j \in S \cap V(Y_0)} d(x_i, x_j) + \sum_{x_i, x_j \in S \cap V(Y_1)} d(x_i, x_j) + \\ &\quad 2|Y_1| \sum_{x_i \in S \cap V(Y_0)} d(x_i, t_0) + \sum_{x_i \in S \cap V(Y_0)} \sum_{z \in B} d(x_i, z) b(z) + \\ &\quad 2|Y_0| \sum_{x_j \in S \cap V(Y_1)} d(t_1, x_j) + \sum_{x_i \in S \cap V(Y_1)} \sum_{z \in B} d(x_i, z) b(z). \end{aligned}$$

This is because if $z \in B \cap Z_1$, for any node $x_i \in S \cap V(Y_0)$, we have $d(x_i, z) = d(x_i, t_0) + d(t_0, z)$ (and analogously, if $z \in B \cap Z_0$, for any node $x_i \in S \cap V(Y_1)$, we have $d(x_i, z) = d(x_i, t_1) + d(t_1, z)$). Hence:

$$\sum_{x_i \in S \cap V(Y_0)} \sum_{z \in B} d(x_i, z) b(z) = \sum_{x_i \in S \cap V(Y_0)} \sum_{z \in B \cap Z_0} d(x_i, z) b(z) + \sum_{x_i \in S \cap V(Y_0)} \sum_{z \in B \cap Z_1} d(x_i, t_0) b(z) + d(t_0, z) b(z)$$

and

$$\sum_{x_i \in S \cap V(Y_1)} \sum_{z \in B} d(x_i, z) b(z) = \sum_{x_i \in S \cap V(Y_1)} \sum_{z \in B \cap Z_1} d(x_i, z) b(z) + \sum_{x_i \in S \cap V(Y_1)} \sum_{z \in B \cap Z_0} d(x_i, t_1) b(z) + d(t_1, z) b(z).$$

Observe that t_0 or t_1 might already belong to B : this is why we assumed that B is a multiset.

Then, we have that:

$$f(S_0) = \sum_{x_i, x_j \in S \cap V(Y_0)} d(x_i, x_j) + \sum_{x_i \in S \cap V(Y_0)} \sum_{z \in B \cap Z_0} d(x_i, z) b(z) + \sum_{x_i \in S \cap V(Y_0)} d(x_i, t_0) \cdot \left(2|Y_1| + \sum_{z \in B \cap Z_1} b(z) \right)$$

$$f(S_1) = \sum_{x_i, x_j \in S \cap V(Y_1)} d(x_i, x_j) + \sum_{x_i \in S \cap V(Y_1)} \sum_{z \in B \cap Z_1} d(x_i, z) b(z) + \sum_{x_i \in S \cap V(Y_1)} d(x_i, t_1) \cdot \left(2|Y_0| + \sum_{z \in B \cap Z_0} b(z) \right)$$

Hence, by adding t_s to $B \cap Z_s = B_s$, with weight equal to $b_s = 2|Y_{1-s}| + \sum_{z \in B \cap Z_{1-s}} b(z)$, $f(S)$ can be reduced to $f(S_0)$ and $f(S_1)$. \square

C The algorithm for finding useful edges

Algorithm 2: USEFULEDGE

Input: An instance $G, \{X_1, \dots, X_k\}, B, b$
Output: A useful edge, or null
 Pick a node u of the set X_i of the instance $G, \{X_1, \dots, X_k\}, B, b$
 Mark all the nodes as unseen
 $\text{dfs}[] \leftarrow -1, \text{visited} \leftarrow 0, \text{usefulEdgeFound} \leftarrow \text{false}, \text{usefulEdge} \leftarrow \text{null}$
 $\text{DFS}(u, -1)$
if usefulEdgeFound **then**
 | **return** usefulEdge
else
 | **return** null
end

Algorithm 3: DFS

Input: A node u , its parent p
Output: A pair (t, Y) , where t is an integer and Y is a set of indices
if usefulEdgeFound **then** **return** null Mark u as seen
 $\text{dfs}[u] \leftarrow \text{visited}$
 $\text{visited} \leftarrow \text{visited} + 1$
 $\text{furthestAncestor} \leftarrow \text{visited}$
 $Y \leftarrow \emptyset$
if $t \in X_i$ **then** $Y \leftarrow Y \cup \{i\}$ **for** $v \in N(u)$ s.t. $w \neq p$ **do**
 | **if** v is unseen **then**
 | | $(t', Y') \leftarrow \text{DFS}(v, u)$
 | | **if** $t' > \text{dfs}[u]$ and $\emptyset \neq Y' \neq \{1, \dots, k\}$ **then**
 | | | $\text{usefulEdgeFound} \leftarrow \text{true}$
 | | | $\text{usefulEdge} \leftarrow (u, v)$
 | | | **return** null
 | | **end**
 | | $\text{furthestAncestor} \leftarrow \min(\text{furthestAncestor}, t')$
 | | $Y \leftarrow Y \cup Y'$
 | **else**
 | | $\text{furthestAncestor} \leftarrow \min(\text{furthestAncestor}, \text{dfs}[v])$
 | **end**
end
return $(\text{furthestAncestor}, Y)$

RESEARCH ARTICLE

Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions

Young-Ho Eom¹, Pablo Aragón², David Laniado², Andreas Kaltenbrunner²,
Sebastiano Vigna³, Dima L. Shepelyansky^{1*}

1 Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, F-31062 Toulouse, France, **2** Barcelona Media Foundation, Barcelona, Spain, **3** Dipartimento di Informatica, Università degli Studi di Milano, Milano, Italy

* dima@irsamc.ups-tlse.fr



OPEN ACCESS

Citation: Eom Y-H, Aragón P, Laniado D, Kaltenbrunner A, Vigna S, Shepelyansky DL (2015) Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions. PLoS ONE 10(3): e0114825. doi:10.1371/journal.pone.0114825

Academic Editor: Zhong-Ke Gao, Tianjin University, CHINA

Received: May 30, 2014

Accepted: November 14, 2014

Published: March 4, 2015

Copyright: © 2015 Eom et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All used computational data are publicly available at <http://dumps.wikimedia.org/>. All the raw data necessary to replicate the findings and conclusion of this study are within the paper, supporting information files and this Wikimedia web site.

Funding: This research is supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE number 288956). No additional external or internal funding was received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Wikipedia is a huge global repository of human knowledge that can be leveraged to investigate interwovenements between cultures. With this aim, we apply methods of Markov chains and Google matrix for the analysis of the hyperlink networks of 24 Wikipedia language editions, and rank all their articles by PageRank, 2DRank and CheiRank algorithms. Using automatic extraction of people names, we obtain the top 100 historical figures, for each edition and for each algorithm. We investigate their spatial, temporal, and gender distributions in dependence of their cultural origins. Our study demonstrates not only the existence of skewness with local figures, mainly recognized only in their own cultures, but also the existence of global historical figures appearing in a large number of editions. By determining the birth time and place of these persons, we perform an analysis of the evolution of such figures through 35 centuries of human history for each language, thus recovering interactions and entanglement of cultures over time. We also obtain the distributions of historical figures over world countries, highlighting geographical aspects of cross-cultural links. Considering historical figures who appear in multiple editions as interactions between cultures, we construct a network of cultures and identify the most influential cultures according to this network.

Introduction

The influence of digital media on collective opinions, social relationships, and information dynamics is growing significantly with the advances of information technology. On the other hand, understanding how collective opinions are reflected in digital media has crucial importance. Among such a medium, Wikipedia, the open, free, and online encyclopedia, has crucial importance since it is not only the largest global knowledge repository but also the biggest collaborative knowledge platform on the Web. Thanks to its huge size, broad coverage and ease of use, Wikipedia is currently one of the most widely used knowledge references. However, since its beginning, there have been constant concerns about the reliability of Wikipedia because of

Competing Interests: The authors have declared that no competing interests exist.

its openness. Although professional scholars may not be affected by a possible skewness or bias of Wikipedia, students and the public can be affected significantly [1, 2]. Extensive studies have examined the reliability of contents [1–3], topic coverage [4], vandalism [5], and conflict [6–8] in Wikipedia.

Wikipedia is available in different language editions; 287 language editions are currently active. This indicates that the same topic can be described in hundreds of articles written by different language user groups. Since language is one of the primary elements of culture [9], collective cultural biases may be reflected on the contents and organization of each Wikipedia edition. Although Wikipedia adopts a “neutral point of view” policy for the description of contents, aiming to provide unbiased information to the public [10], it is natural that each language edition presents reality from a different angle. To investigate differences and relationships among different language editions, we develop mathematical and statistical methods which treat the huge amount of information in Wikipedia, excluding cultural preferences of the investigators.

Cultural bias or differences across Wikipedia editions have been investigated in previous research [11–17]. A special emphasis was devoted to persons described in Wikipedia articles [12] and their ranking [18, 19]. Indeed, human knowledge, as well as Wikipedia itself, was created by people who are the main actors of its development. Thus it is rather natural to analyze a ranking of people according to the Wikipedia hyper-link network of citations between articles (see network data description below). A cross-cultural study of biographical articles was presented in [20], by building a network of interlinked biographies. Another approach was proposed recently in [21]: the difference in importance of historical figures across Wikipedia language editions is assessed on the basis of the global ranking of Wikipedia articles about persons. This study, motivated by the question “Is an important person in a given culture also important in other cultures?”, showed that there are strong entanglements and local biases of historical figures in Wikipedia. Indeed, the results of the study show that each Wikipedia edition favors persons belonging to the same culture (language), but also that there are cross-Wikipedia top ranked persons, who can be signs of entanglement between cultures. These cross-language historical figures can be used to generate inter-culture networks demonstrating interactions between cultures [21]. Such an approach provides us novel insights on cross-cultural differences across Wikipedia editions. However, in [21] only 9 Wikipedia editions, mainly languages spoken in European, have been considered. Thus a broader set of language editions is needed to offer a more complete view on a global scale.

We note that the analysis of persons’ importance via Wikipedia becomes more and more popular. This is well visible from the appearance of new recent studies for the English Wikipedia [22] and for multiple languages [23]. The analysis of coverage of researchers and academics via Wikipedia is reported in [24].

Here we investigate interactions and skewness of cultures with a broader perspective, using global ranking of articles about persons in 24 Wikipedia language editions. According to Wikipedia [25] these 24 languages cover 59 percent of world population. Moreover, according to Wikipedia [26], our selection of 24 language editions covers the 68 percent of the total number of 30.9 millions of Wikipedia articles in all 287 languages. These 24 editions also cover languages which played an important role in human history including Western, Asian and Arabic cultures.

On the basis of this data set we analyze spatial, temporal, and gender skewness in Wikipedia by analyzing birth place, birth date, and gender of the top ranked historical figures in Wikipedia. We identified overall Western, modern, and male skewness of important historical figures across Wikipedia editions, a tendency towards local preference (i.e. each Wikipedia edition favors historical figures born in countries speaking that edition’s language), and the existence of

global historical figures who are highly ranked in most of Wikipedia editions. We also constructed networks of cultures based on cross-cultural historical figures to represent interactions between cultures according to Wikipedia.

To obtain a unified ranking of historical figures for all 24 Wikipedia editions, we introduce an average ranking which gives us the top 100 persons of human history. To assess the alignment of our ranking with previous work by historians, we compare it with the Hart's list of the top 100 people who, according to him, most influenced human history [27]. We note that Hart "ranked these 100 persons in order of importance: that is, according to the total amount of influence that each of them had on human history and on the everyday lives of other human beings".

Methods

In this research, we consider each Wikipedia edition as a network of articles. Each article corresponds to a node of the network and hyperlinks between articles correspond to links of the network. For a given network, we can define an adjacency matrix A_{ij} . If there is a link (one or more) from node (article) j to node (article) i then $A_{ij} = 1$, otherwise, $A_{ij} = 0$. The out-degree $k_{out}(j)$ is the number of links from node j to other nodes and the in-degree $k_{in}(j)$ is the number of links to node j from other nodes. The links between articles are considered only inside a given Wikipedia edition, there are no links counted between editions. Thus each language edition is analyzed independently from others by the Google matrix methods described below. The transcriptions of names from English to the other 23 selected languages are harvested from WikiData (<http://dumps.wikimedia.org/wikidatawiki>) and not directly from the text of articles.

To rank the articles of a Wikipedia edition, we use two ranking algorithms based on the articles network structure. Detailed descriptions of these algorithms and their use for Wikipedia editions are given in [18, 19, 28, 29]. The methods used here are described in [21]; we keep the same notations.

Google matrix

First we construct the matrix S_{ij} of Markov transitions by normalizing the sum of the elements in each column of A to unity ($S_{ij} = A_{ij}/\sum_i A_{ij}$, $\sum_i S_{ij} = 1$) and replacing columns with zero elements by elements $1/N$ with N being the matrix size. Then the Google matrix is given by the relation $G_{ij} = \alpha S_{ij} + (1 - \alpha)/N$, where α is the damping factor [30]. As in [21] we use the conventional value $\alpha = 0.85$. It is known that the variation of α in a range $0.5 \leq \alpha < 0.95$ does not significantly affect the probability distribution of ranks discussed below (see e.g. [18, 19, 30]).

PageRank algorithm

PageRank is a widely used algorithm to rank nodes in a directed network. It was originally introduced for Google web search engine to rank web pages of the World Wide Web based on the idea of academic citations [31]. Currently PageRank is used to rank nodes of network systems from scientific papers [32] to social network services [33], world trade [34] and biological systems [35]. Here we briefly outline the iteration method of PageRank computation. The PageRank vector $P(i, t)$ of a node i at iteration t in a network with N nodes is given by

$$P(i, t) = \sum_j G_{ij} P(j, t - 1) = (1 - \alpha)/N + \alpha \sum_j A_{ij} P(j, t - 1)/k_{out}(j). \quad (1)$$

The stationary state $P(i)$ of $P(i, t)$ is the PageRank of node i . More detailed information about the PageRank algorithm is described in [30]. Ordering all nodes by their decreasing probability $P(i)$, we obtain the PageRank ranking index $K(i)$. In qualitative terms, the PageRank probability of a node is proportional to the number of incoming links weighted according to their own probability. A random network surfer spends on a given node a time given on average by the PageRank probability.

CheiRank algorithm

In a directed network, outgoing links can be as important as ingoing links. In this sense, as a complementary to PageRank, the CheiRank algorithm is defined and used in [18, 28, 36]. The CheiRank vector $P^*(i, t)$ of a node at iteration time t is given by

$$P^*(i) = (1 - \alpha)/N + \alpha \sum_j A_{ji} P^*(j)/k_{in}(j) \quad (2)$$

Same as the case of PageRank, we consider the stationary state $P^*(i)$ of $P^*(i, t)$ as the CheiRank probability of node i with $\alpha = 0.85$. High CheiRank nodes in the network have large out-degree. Ordering all nodes by their decreasing probability $P^*(i)$, we obtain the CheiRank ranking index $K^*(i)$. The PageRank probability of an article is proportional to the number of incoming links, while the CheiRank probability of an article is proportional to the number of outgoing links. Thus a top PageRank article is important since other articles refer to it, while a top CheiRank article is highly connected because it refers to other articles.

2DRank algorithm

PageRank and CheiRank algorithms focus only on in-degree and out-degree of nodes, respectively. The 2DRank algorithm considers both types of information simultaneously to rank nodes with a balanced point of view in a directed network. Briefly speaking, nodes with both high PageRank and CheiRank get high 2DRank ranking. Consider a node i which is K_i -th ranked by PageRank and K^*_i ranked by CheiRank. Then we can assign a secondary ranking $K'_i = \max\{K_i, K^*_i\}$ to the node. If $K'_i < K'_j$, then node j has lower 2DRank and vice versa. A detailed illustration and description of this algorithm is given in [18].

We note that the studies reported in [21] show that the overlap between top CheiRank persons of different editions is rather small and due to that the statistical accuracy of this data is not sufficient for determining interactions between different cultures for the CheiRank list. Moreover, CheiRank, based on outgoing links only, selects mainly persons from such activity fields like sports and arts where the historical trace is not so important. Due to these reasons we restrict our study to PageRank and 2DRank. It can be also interesting to use other algorithms of ranking, e.g. LeaderRank [37], but here we restrict ourselves to the methods which we already tested, leaving investigation of other ranking methods for further studies.

Data preparation

We consider 24 different language editions of Wikipedia: English (EN), Dutch (NL), German (DE), French (FR), Spanish (ES), Italian (IT), Portuguese (PT), Greek (EL), Danish (DA), Swedish (SV), Polish (PL), Hungarian (HU), Russian (RU), Hebrew (HE), Turkish (TR), Arabic (AR), Persian (FA), Hindi (HI), Malaysian (MS), Thai (TH), Vietnamese (VI), Chinese (ZH), Korean (KO), and Japanese (JA). The Wikipedia data were collected in middle February 2013. The overview summary of each Wikipedia is represented in Table 1.

We understand that our selection of Wikipedia editions does not represent a complete view of all the 287 languages of Wikipedia editions. However, this selection covers most of the

Table 1. Wikipedia hyperlink networks from the 24 considered language editions. Here N_a is the number of articles. Wikipedia data were collected in middle February 2013.

Edition	Language	N_a	Edition	Language	N_a
EN	English	4212493	RU	Russian	966284
NL	Dutch	1144615	HE	Hebrew	144959
DE	German	1532978	TR	Turkish	206311
FR	French	1352825	AR	Arabic	203328
ES	Spanish	974025	FA	Persian	295696
IT	Italian	1017953	HI	Hindi	96869
PT	Portuguese	758227	MS	Malaysian	180886
EL	Greek	82563	TH	Thai	78953
DA	Danish	175228	VI	Vietnamese	594089
SV	Swedish	780872	ZH	Chinese	663485
PL	Polish	949153	KO	Korean	231959
HU	Hungarian	235212	JA	Japanese	852087

doi:10.1371/journal.pone.0114825.t001

largest language editions and allows us to perform quantitative and statistical analysis of important historical figures. Among the 20 largest editions (counted by their size, taken at the middle of 2014) we have not considered the following editions: Waray-Waray, Cebuano, Ukrainian, Catalan, Bokmal-Riksmaal, and Finnish.

First we ranked all the articles in a given Wikipedia edition by PageRank and 2DRank algorithms, and selected biographical articles about historical figures. To identify biographical articles, we considered all articles belonging to “Category:living people”, or to “Category:Deaths by year” or “Category:Birth by year” or their subcategories in the English Wikipedia. In this way, we obtained a list of about 1.1 million biographical articles. We identified birth place, birth date, and gender of each selected historical figure based on DBpedia [38] or a manual inspection of the corresponding Wikipedia biographical article, when for the considered historical figure no DBpedia data were available. We then started from the list of persons with their biographical article’s title on the English Wikipedia, and found the corresponding titles in other language editions using the inter-language links provided by WikiData. Using the corresponding articles, identified by the inter-languages links in different language editions, we extracted the top 100 persons from the rankings of all Wikipedia articles of each edition. At the end, for each Wikipedia edition and for each ranking algorithm, we have information about the top 100 historical figures with their corresponding name in the English Wikipedia, their birth place and date, and their gender. All 48 lists of the top 100 historical figures in PageRank and 2DRank for the 24 Wikipedia editions and for the two ranking algorithms are represented in [39] and Supporting Information (SI). The original network data for each edition are available at [39]. The automatic extraction of persons from PageRank and 2DRank listings of articles of each edition is performed by using the above whole list of person names in all 24 editions. This method implies a significantly higher recall compared to the manual selection of persons from the ranking list of articles for each edition used in [21].

We attribute each of the 100 historical figures to a birth place at the country level (actual country borders), to a birth date in year, to a gender, and to a cultural group. Historical figures are assigned to the countries currently at the locations where they were born. The cultural group of historical figures is assigned by the most spoken language of their birth place at the current country level. For example, if someone was born in “Constantinople” in the ancient Roman era, since the place is now Istanbul, Turkey, we assign her/his birth place as “Turkey” and since Turkish is the most spoken language in Turkey, we assign this person to the Turkish

Table 2. List of top persons by PageRank and 2DRank for the English Wikipedia. All names are represented by article titles in the English Wikipedia.

Rank	PageRank persons	2DRank persons
1st	Napoleon	Frank Sinatra
2nd	Barack Obama	Michael Jackson
3rd	Carl Linnaeus	Pope Pius XII
4th	Elizabeth II	Elton John
5th	George W. Bush	Elizabeth II
6th	Jesus	Pope John Paul II
7th	Aristotle	Beyoncé Knowles
8th	William Shakespeare	Jorge Luis Borges
9th	Adolf Hitler	Mariah Carey
10th	Franklin D. Roosevelt	Vladimir Putin

doi:10.1371/journal.pone.0114825.t002

cultural group. If the birth country does not belong to any of the 24 cultures (languages) which we consider, we assign WR (world) as the culture of this person. We would like to point out that although a culture can not be defined only by language, we think that language is a suitable first-approximation of culture. All lists of top 100 historical figures with their birth place, birth date, gender, and cultural group for each Wikipedia edition and for each ranking algorithm are represented in [39]. A part of this information is also reported in SI.

To apply PageRank and 2DRank methods, we consider each edition as the network of articles of the given edition connected by hyper-links among the articles (see the details of ranking algorithms in Section [Methods](#)). The full list of considered Wikipedia language editions is given in [Table 1](#). [Table 2](#) represents the top 10 historical figures by PageRank and 2DRank in the English Wikipedia. Roughly speaking, top PageRank articles imply highly cited articles in Wikipedia and top 2DRank articles imply articles which are both highly cited and highly citing in Wikipedia. In total, we identified 2400 top historical figures for each ranking algorithm. However, since some historical figures such as *Jesus*, *Aristotle*, or *Napoleon* appear in multiple Wikipedia editions, we have 1045 unique top PageRank historical figures and 1616 unique top 2DRank historical figures.

We should note that the extraction of persons and their information from a Wikipedia edition is not an easy task even for the English edition, and much more complicated for certain other language editions. Therefore, the above automatic method based on 1.1 million English names and their corresponding names seems to us to be the most adequate approach. Of course, it will miss people who do not have a biographical article on the English Wikipedia. Cross-checking investigation is done for Korean and Russian Wikipedia, which are native languages for two authors, by manually selecting top 100 persons from top lists of all articles ordered by PageRank and 2DRank in both Wikipedia editions. We find that our automatic search misses on average only 2 persons from 100 top persons for these two editions (the missed names are given in SI). The errors appear due to transcription changes of names or missing cases in our name-database based on English Wikipedia. For Western languages the number of errors is presumably reduced since transcription remains close to English. Based on the manual inspection for the Korean and the Russian Wikipedia, we expect that the errors of our automatic recovery of the top people from the whole articles ordered by PageRank and 2DRank are on a level of two percent.

We also note that our study is in compliance with Wikipedia's Terms and Conditions.

Results

Above we described the methods used for the extraction of the top 100 persons in the ranking list of each edition. Below we present the obtained results describing the spatial, temporal and gender distributions of top ranked historical figures. We also determine the global and local persons and obtain the network of cultures based on the ranking of persons from a given language by other language editions of Wikipedia.

Spatial distribution

The birth places of historical figures are attributed to the country containing their geographical location of birth according to the present geographical territories of all world countries. The list of countries appeared for the top 100 persons in all editions is given in [Table 3](#). We also

Table 3. List of country code (CC), countries as birth places of historical figures, and language code (LC) for each country. LC is determined by the most spoken language in the given country. Country codes are based on country codes of Internet top-level domains and language codes are based on language edition codes of Wikipedia; WR represents all languages other than the considered 24 languages.

CC	Country	LC	CC	Country	LC	CC	Country	LC
AE	United Arab Emirates	AR	AF	Afghanistan	FA	AL	Albania	WR
AR	Argentina	ES	AT	Austria	DE	AU	Australia	EN
AZ	Azerbaijan	TR	BE	Belgium	NL	BG	Bulgaria	WR
BR	Brazil	PT	BS	Bahamas	EN	BY	Belarus	RU
CA	Canada	EN	CH	Switzerland	DE	CL	Chile	ES
CN	China	ZH	CO	Colombia	ES	CU	Cuba	ES
CY	Cyprus	EL	CZ	Czech Rep.	WR	DE	Germany	DE
DK	Denmark	DA	DZ	Algeria	AR	EG	Egypt	AR
ES	Spain	ES	FI	Finland	WR	FR	France	FR
GE	Georgia	WR	GR	Greece	EL	HK	Hong Kong	ZH
HR	Croatia	WR	HU	Hungary	HU	ID	Indonesia	WR
IE	Ireland	EN	IL	Israel	HE	IN	India	HI
IQ	Iraq	AR	IR	Iran	FA	IS	Iceland	WR
IT	Italy	IT	JP	Japan	JA	KE	Kenya	EN
KG	Kyrgyzstan	WR	KH	Cambodia	WR	KO	S. Korea	KO
KP	N. Korea	KO	KW	Kuwait	AR	KZ	Kazakhstan	WR
LB	Lebanon	AR	LT	Lithuania	WR	LV	Latvia	WR
LY	Libya	AR	MK	Macedonia	WR	MM	Myanmar	WR
MN	Mongolia	WR	MX	Mexico	ES	MY	Malaysia	MS
NL	Netherlands	NL	NO	Norway	WR	NP	Nepal	WR
NZ	New Zealand	EN	OM	Oman	AR	PA	Panama	ES
PE	Peru	ES	PK	Pakistan	HI	PL	Poland	PL
PS	State of Palestine	AR	PT	Portugal	PT	RO	Romania	WR
RS	Serbia	WR	RU	Russia	RU	SA	Saudi Arabia	AR
SD	Sudan	AR	SE	Sweden	SV	SG	Singapore	ZH
SI	Slovenia	WR	SK	Slovakia	WR	SR	Suriname	NL
SY	Syria	AR	TH	Thailand	TH	TJ	Tajikistan	WR
TN	Tunisia	AR	TR	Turkey	TR	TW	Taiwan	ZH
TZ	Tanzania	WR	UA	Ukraine	WR	UK	United Kingdom	EN
US	United States	EN	UZ	Uzbekistan	WR	VE	Venezuela	ES
VN	Vietnam	VI	XX	Unknown	WR	YE	Yemen	AR
ZA	South Africa	WR						

doi:10.1371/journal.pone.0114825.t003

attribute each country to one of the 24 languages of the considered editions. This attribution is done according to the language spoken by the largest part of population in the given country. Thus e.g. Belgium is attributed to Dutch (NL) since the majority of the population speaks Dutch. If the main language of a country is not among our 24 languages, then this country is attributed to an additional section WR corresponding to the remaining world (e.g. Ukraine, Norway are attributed to WR). If the birth place of a person is not known, then it is also attributed to WR. The choice of attribution of a person to a given country in its current geographic territory, and as a result to a certain language, may have some fluctuations due to historical variations of country borders (e.g. Immanuel Kant was born in the current territory of Russia and hence is attributed to Russian language). However, the number of such cases is small, being on a level of 3.5 percent (see Section “[Network of cultures](#)” below). We think that the way in which a link between person, language and country is fixed by the birth place avoids much larger ambiguity of attribution of a person according to the native language which is not so easy to fix in an automatic manner.

The obtained spatial distribution of historical figures of Wikipedia over countries is shown in [Fig. 1](#). This averaged distribution gives the average number of top 100 persons born in a specific country as birth place, with averaging done over our 24 Wikipedia editions. Thus an average over the 24 editions gives for Germany (DE) approximately 9.7 persons in the top 100 of PageRank, being at the first position, followed by USA with approximately 9.5 persons. For

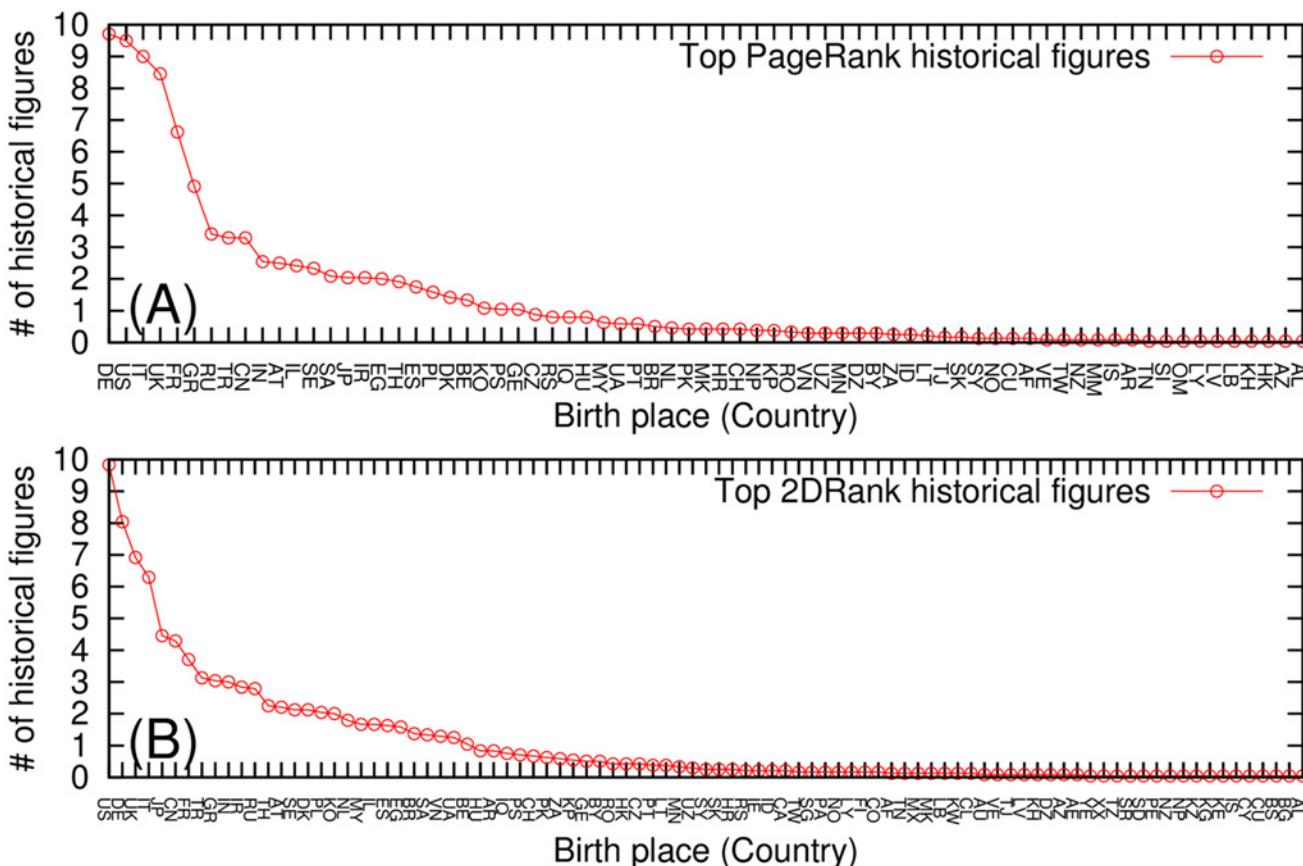


Fig 1. Birth place distribution of top historical figures averaged over 24 Wikipedia edition for (A) PageRank historical figures (71 countries) and (B) 2DRank historical figures (91 countries). Two letter country codes are represented in [Table 3](#).

doi:10.1371/journal.pone.0114825.g001

2DRank we have USA at the first position with an average of 9.8 persons and Germany at the second with an average of 8.0 persons.

Western (Europe and USA) skewed patterns are observed in both top PageRank historical figures (Fig. 1. (A)) and top 2DRank historical figures (Fig. 1. (B)). This Western skewed pattern is remarkable since 11 Wikipedia editions of the 24 considered editions are not European language editions. Germany, USA, Italy, UK and France are the top five birth places of top PageRank historical figures among 71 countries. On the other hand, USA, Germany, UK, Italy and Japan are top five birth places of the top 2DRank historical figures among 91 countries.

In Fig. 2 we show the world map of countries, where color indicates the number of persons from a given country among the 24×100 top persons for PageRank and 2DRank. Additional figures showing these distributions for different centuries are available at [39].

We also observed local skewness in the spatial distribution of the top historical figures for the PageRank (2DRank) ranking algorithm as shown in Fig. 3A (in Fig. 3B). For example, 47 percent of the top PageRank historical figures in the English Wikipedia were born in USA (25 percent) and UK (22 percent) and 56 percent of the top historical figures in the Hindi Wikipedia were born in India. A similar strong locality pattern of the top historical figures was observed in our previous research [21]. However it should be noted that in the previous study we considered the native language of the top historical figure as a criterion of locality, while in the current study we considered 'birth place' as criterion of locality.

Regional skewness, the preferences of Wikipedia editions for historical figures who were born in geographically or culturally related countries, is also observed. For example, 18 (5) of the top 100 PageRank historical figures in the Korean (Japanese) Wikipedia were born in China. Also 9 of the top 100 PageRank historical figures in the Persian Wikipedia were born in Saudi Arabia. The distribution of top persons from each Wikipedia edition over world countries is shown in Fig. 3A and Fig. 3B. The countries on a horizontal axis are grouped by clusters of corresponding language so that the links inside a given culture (or language) become well visible.

To observe patterns in a better way at low numbers of historical figures, we normalized each column of Fig. 3A and Fig. 3B corresponding to a given country. In this way we obtain a re-scaled distribution with better visibility for each birth country level as shown in Fig. 3C and Fig. 3D, respectively. We can observe a clear birth pattern of top PageRank historical figures born in Lebanon, Libya, Oman, and Tunisia in the case of the Arabic Wikipedia, and historical figures born in N. Korea appearing not only in the Korean but also in the Japanese Wikipedia.

In the case of the top 2DRank historical figures shown in Fig. 3B and Fig. 3D, we observe overall patterns of locality and regions being similar to the case of PageRank, but the locality is stronger.

In short, we observed that most of the top historical figures in Wikipedia were born in Western countries, but also that each edition shows its own preference to the historical figures born in countries which are closely related to the corresponding language edition.

Temporal distribution

The analysis of the temporal distribution of top historical figures is done based on their birth dates. As shown in Fig. 4A for PageRank, most of historical figures were born after the 17th century on average, which shows similar pattern with world population growth [40]. However, there are some distinctive peaks around BC 5th century and BC 1st century for the case of PageRank because of Greek scholars (*Socrates, Plato, and Herodotus*), Roman politicians (*Julius Caesar, Augustus*) and Christianity leaders (*Jesus, Paul the Apostle, and Mary (mother of Jesus)*). We also observe that the Arabic and the Persian Wikipedia have more historical figures



Fig 2. Sum of appearances of historical figures from a given country in the 24 lists of top 100 persons for PageRank (top panel) and 2DRank (bottom panel). Color changes from zero (white) to maximum (black). Maximal values are 233 appearances for Germany (top) and 236 for USA (bottom). Values are proportional to the averages per country shown in [Fig. 1](#).

doi:10.1371/journal.pone.0114825.g002

than Western language Wikipedia editions from AD 6th century to AD 12th century. For the case of 2DRank in [Fig. 4B](#), there is only one small peak around BC 1C, which is also smaller than the peak in the case of PageRank, and all the distribution is dominated by a strong growth on the 20th century.

The distributions of the top PageRank historical figures over the 24 Wikipedia editions for each century are shown in [Fig. 4C](#). The same distribution, but normalized to unity over all

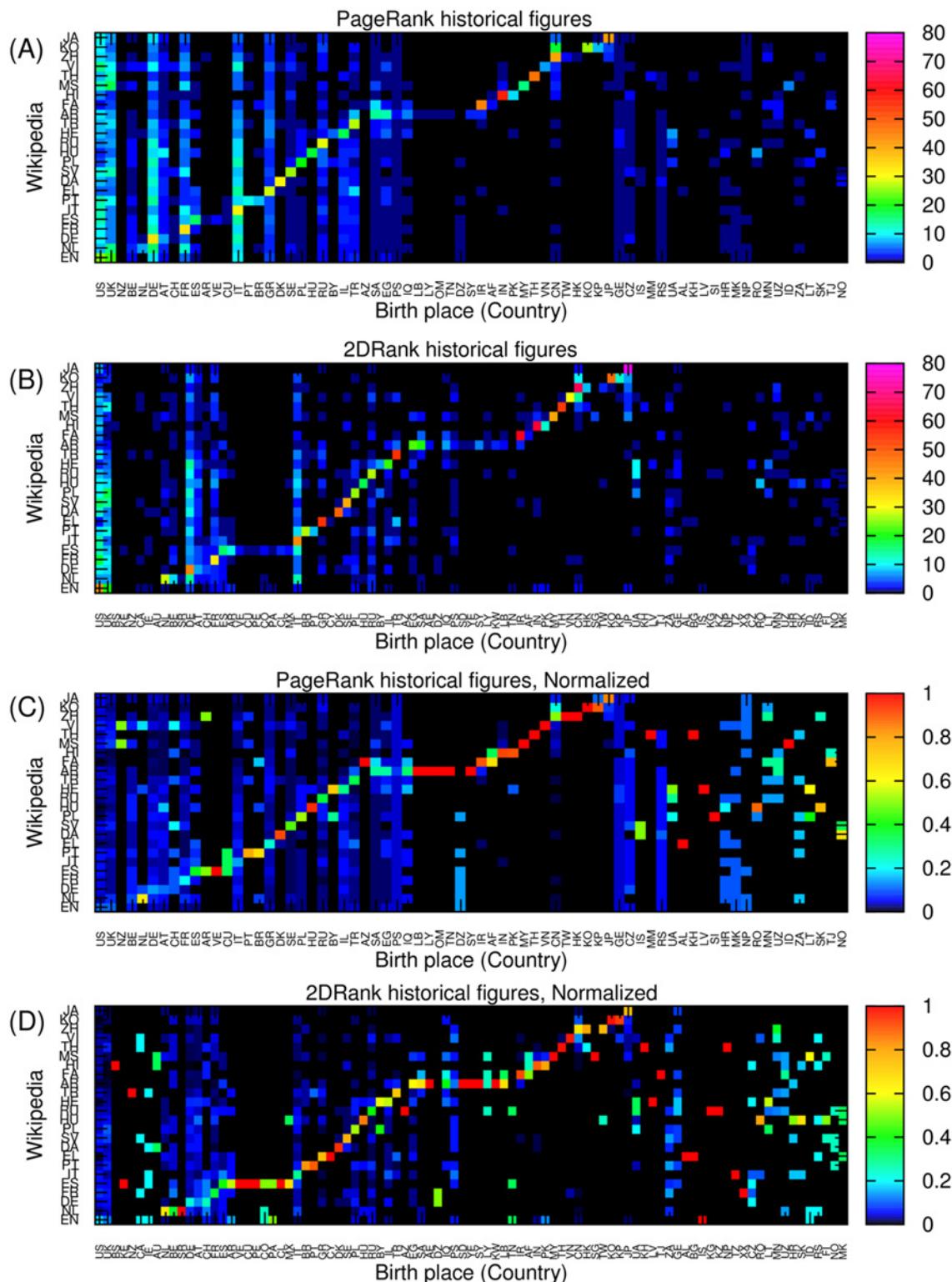


Fig 3. Birth place distributions over countries of top historical figures from each Wikipedia edition; two letter country codes are represented in Table 3. Panels: (A) distributions of PageRank historical figures over 71 countries for each Wikipedia edition; (B) distributions of 2DRank historical figures over 91 countries for each Wikipedia edition; (C) column normalized birth place distributions of PageRank historical figures of panel (A); (D) column normalized birth place distributions of 2DRank historical figures of panel (B).

doi:10.1371/journal.pone.0114825.g003

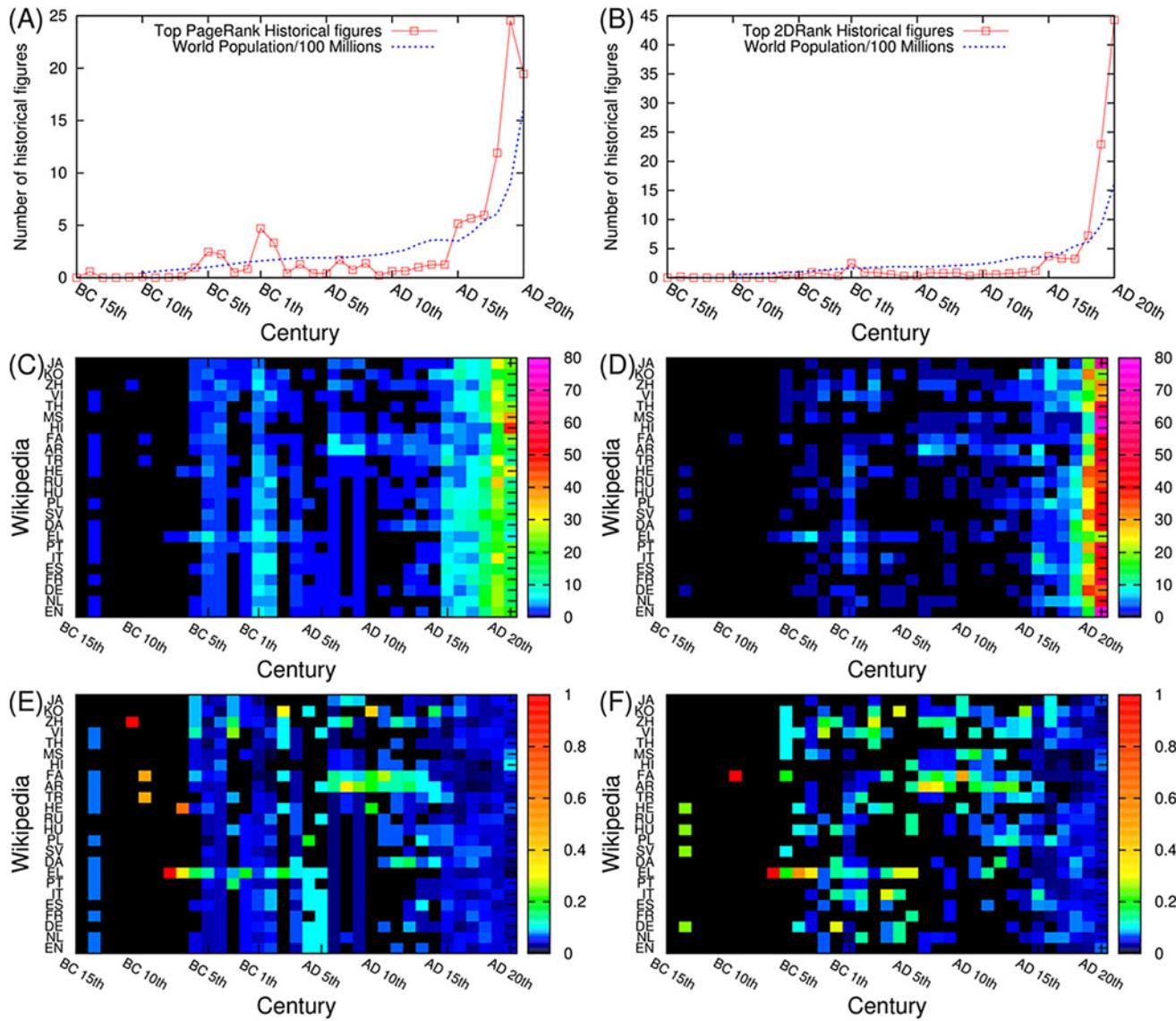


Fig 4. Birth date distributions of top historical figures. (A) Birth date distribution of PageRank historical figures averaged over 24 Wikipedia editions (B) Birth date distribution of 2DRank historical figures averaged over 24 Wikipedia editions (C) Birth date distributions of PageRank historical figures for each Wikipedia edition. (D) Birth date distributions of 2DRank historical figures for each Wikipedia edition. (E) Column normalized birth date distributions of PageRank historical figures for each Wikipedia edition. (F) Column normalized birth date distributions of 2DRank historical figures for each Wikipedia edition.

doi:10.1371/journal.pone.0114825.g004

editions for each century, is shown in Fig. 4E. The Persian (FA) and the Arabic (AR) Wikipedia have more historical figures than other language editions (in particular European language editions) from the 6th to the 12th century due to Islamic leaders and scholars. On the other hand, the Greek Wikipedia has more historical figures in BC 5th century because of Greek philosophers. Also most of western-southern European language editions, including English, Dutch, German, French, Spanish, Italian, Portuguese, and Greek, have more top historical figures because they have *Augustine the Hippo* and *Justinian I* in common. Similar distributions obtained from 2DRank are shown in Fig. 4D and Fig. 4F respectively.

The data of Figs. 4E, F clearly show well pronounced patterns, corresponding to strong interactions between cultures: from BC 5th century to AD 15th century for JA, KO, ZH, VI; from

AD 6th century to AD 12th century for FA, AR; and a common birth pattern in EN, EL, PT, IT, ES, DE, NL (Western European languages) from BC 5th century to AD 6th century. In supporting Figure S1 we show distributions of historical figures over languages according to their birth place. In this case the above patterns become even more pronounced.

At a first glance from [Figs. 4E, F](#) we observe for persons born in AD 20th century a significantly more homogeneous distribution over cultures compared to early centuries. However, as noted in [21], each Wikipedia edition favors historical figures speaking the corresponding language. We investigate how this preference to same-language historical figures changes in time. For this analysis, we define two variables $M_{L, C}$ and $N_{L, C}$ for a given language edition L and a given century C . Here $M_{L, C}$ is the number of historical figures born in all countries being attributed to a given language L , and $N_{L, C}$ is the total number of historical figures for a given century C and a given language edition L . For example, among the 21 top PageRank historical figures from the English Wikipedia, who were born in AD 20th century, two historical figures (Pope John Paul II and Pope Benedict XVI) were not born in English speaking countries. Thus in this case $N_{EN, 20} = 21$ and $M_{EN, 20} = 19$. [Fig. 5](#) represents the ratio $r_{L, C} = M_{L, C}/N_{L, C}$ for each edition and each century. In ancient times (i.e. before AD 5th century), most historical figures for each Wikipedia edition are not born in the same language region except for the Greek, Italian, Hebrew, and Chinese Wikipedia. However, after AD 5th century, the ratio of same language historical figures is rising. Thus, in AD 20th century, most Wikipedia editions have significant numbers of historical figures born in countries speaking the corresponding language. For PageRank persons and AD 20th century, we find that the English edition has the largest fraction of its own language, followed by Arabic and Persian editions while other editions have significantly large connections with other cultures. For the English edition this is related to a significant number of USA presidents appearing in the top 100 list (see [18, 19]). For 2DRank persons the largest fractions were found for Greek, Arabic, Chinese and Japanese cultures. These data show that even in age of globalization there is a significant dominance of local historical figures for certain cultures.

Gender distribution

From the gender distributions of historical figures, we observe a strong male-skewed pattern across many Wikipedia editions regardless of the ranking algorithm. On average, 5.2(10.1) female historical figures are observed among the 100 top PageRank (2DRank) persons for each Wikipedia edition. [Fig. 6](#) shows the number of top female historical figures for each Wikipedia edition. Thai, Hindi, Swedish, and Hebrew have more female historical figures than the average over our 24 editions in the case of PageRank. On the other hand, the Greek and the Korean versions have a lower number of females than the average. In the case of 2DRank, English, Hindi, Thai, and Hungarian Wikipedia have more females than the average while German, Chinese, Korean, and Persian Wikipedia have less females than the average. In short, the top historical figures in Wikipedia are quite male-skewed. This is not surprising since females had little chance to be historical figures for most of human history. We compare the gender skewness to other cases such as the number of female editors in Wikipedia (9 percent) in 2011 [41] and the share of women in parliaments, which was 18.7 percent in 2012 by UN Statistics and indicators on women and men [42], the male skewness for the PageRank list is stronger in the contents of Wikipedia [43]. However, the ratio of females among the top historical figures is growing by time as shown in [Fig. 6 C](#). It is notable that the peak in [Fig. 6C](#) at BC 1st is due to “Mary (mother of Jesus)”. In the 20th century 2DRank gives a larger percentage of women compared to PageRank. This is due to the fact that 2DRank has a larger fraction of singers and artists comparing to PageRank (see [18, 19]) and that the fraction of women in these fields of activity is larger.

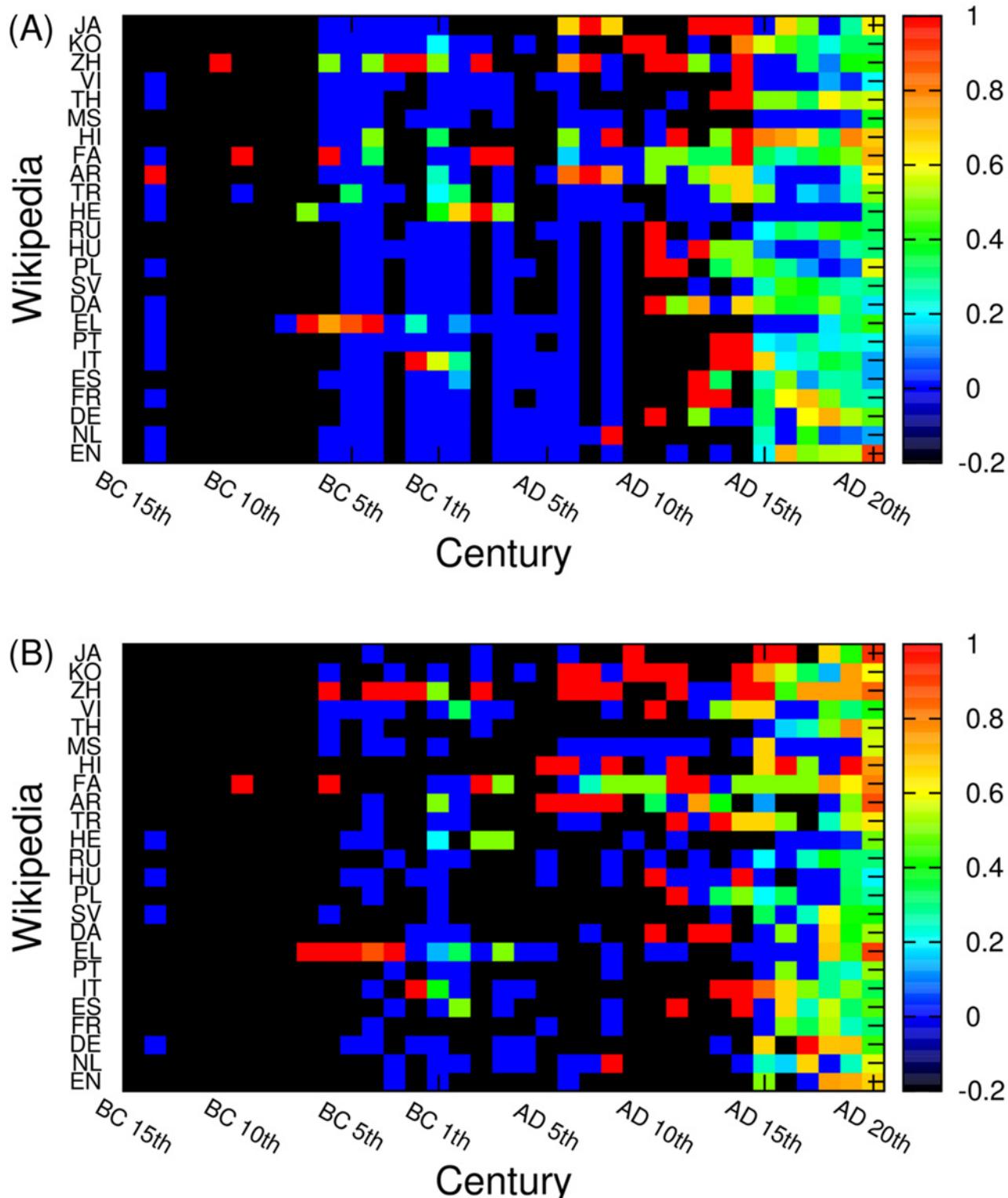


Fig 5. The locality property of cultures represented by the ratio $r_{L,C} = M_{L,C}/N_{L,C}$ for each edition L and each century C . Here $M_{L,C}$ is the number of historical figures born in countries attributed to a given language edition L at century C and $N_{L,C}$ is the total number of historical figures in a given edition at a given century, regardless of language of their birth countries. Black color (-0.2 in the color bars) shows that there is no historical figure at all for a given edition and century; blue (0 in the color bars) shows there are some historical figures but no same language historical figures. Here (A) panel shows PageRank historical figures, and (B) panel shows 2DRank historical figures.

doi:10.1371/journal.pone.0114825.g005

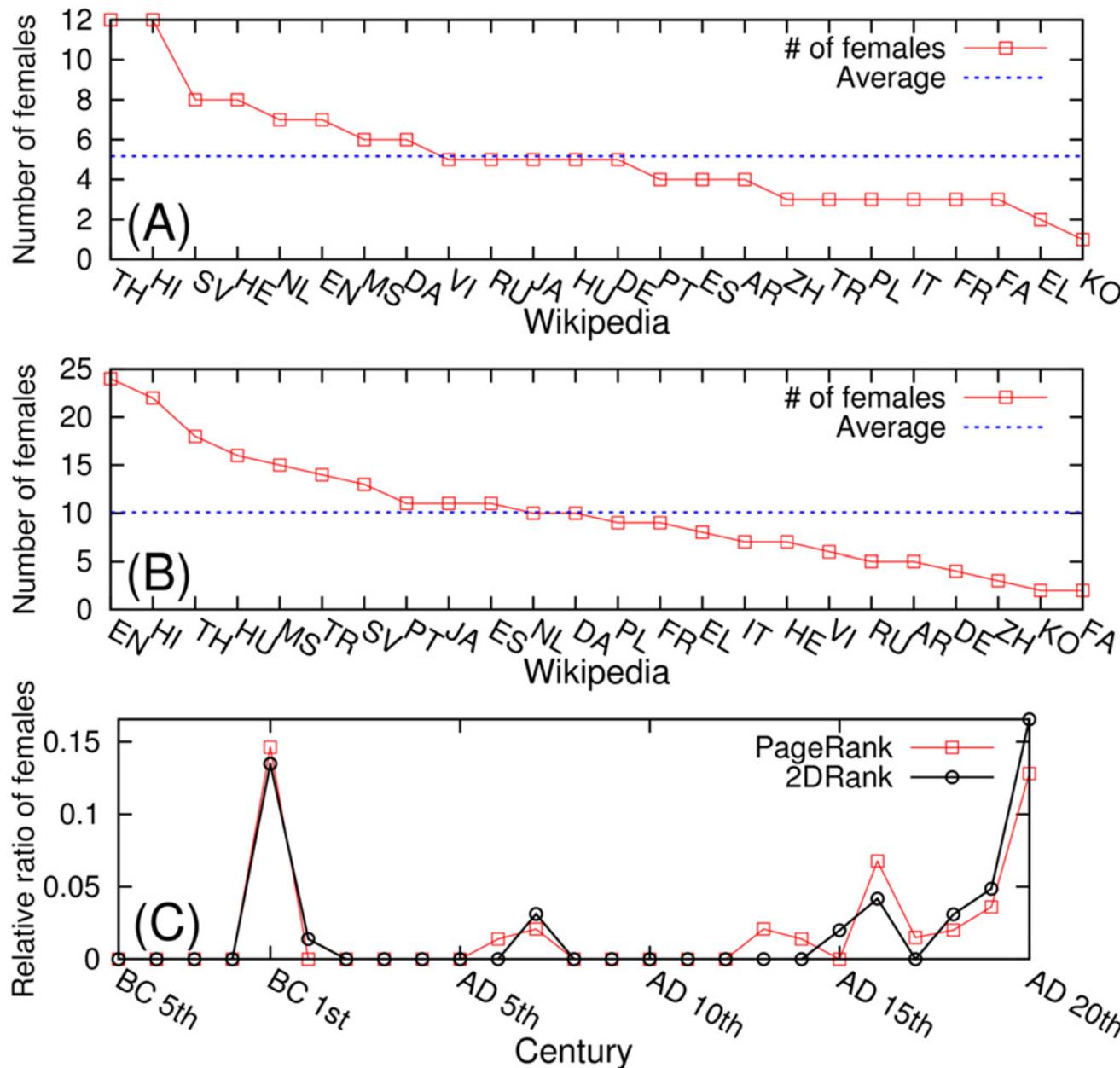


Fig 6. Number of females of top historical figures from each Wikipedia edition (A) Top PageRank historical figures (B) Top 2DRank historical figures. (C) The average female ratio of historical figures in given centuries across 24 Wikipedia editions.

doi:10.1371/journal.pone.0114825.g006

Global historical figures

Above we analyzed how top historical figures in Wikipedia are distributed in terms of space, time, and gender. Now we identify how these top historical figures are distributed in each Wikipedia edition and which are global historical figures. According to previous research [21], there are some global historical figures who are recognized as important historical figures across Wikipedia editions. We identify global historical figures based on the ranking score for a

Table 4. List of global historical figures by PageRank and 2DRank for all 24 Wikipedia editions. All names are represented by the corresponding article titles in the English Wikipedia. Here, Θ_A is the ranking score of algorithm A (3); N_A is the number of appearances of a given person in the top 100 rank for all editions.

Rank	PageRank global figures	Θ_{PR}	N_A	2DRank global figures	Θ_{2D}	N_A
1st	Carl Linnaeus	2284	24	Adolf Hitler	1557	20
2nd	Jesus	2282	24	Michael Jackson	1315	17
3rd	Aristotle	2237	24	Madonna (entertainer)	991	14
4th	Napoleon	2208	24	Jesus	943	14
5th	Adolf Hitler	2112	24	Ludwig van Beethoven	872	14
6th	Julius Caesar	1952	23	Wolfgang Amadeus Mozart	853	11
7th	Plato	1949	24	Pope Benedict XVI	840	12
8th	William Shakespeare	1861	24	Alexander the Great	789	11
9th	Albert Einstein	1847	24	Charles Darwin	773	12
10th	Elizabeth II	1789	24	Barack Obama	754	16

doi:10.1371/journal.pone.0114825.t004

given person determined by her number of appearances and ranking index over our 24 Wikipedia editions.

Following [21], the ranking score $\Theta_{P,A}$ of a historical figure P is given by

$$\Theta_{P,A} = \sum_E (101 - R_{P,E,A}) \quad (3)$$

Here $R_{P,E,A}$ is the ranking of a historical figure P in Wikipedia edition E by ranking algorithm A . According to this definition, a historical figure who appears more often in the lists of top historical figures for the given 24 Wikipedia editions or has higher ranking in the lists gets a higher ranking score. Table 4 represents the top 10 global historical figures for PageRank and 2DRank. *Carl Linnaeus* is the 1st global historical figure by PageRank followed by *Jesus*, *Aristotle*. *Adolf Hitler* is the 1st global historical figure by 2DRank followed by *Michael Jackson*, *Madonna (entertainer)*. On the other hand, the lists of the top 10 local historical figures ordered by our ranking score for each language are represented in supporting Tables S1–S25 and [39].

The reason for a somewhat unexpected PageRank leader *Carl Linnaeus* is related to the fact that he laid the foundations for the modern biological naming scheme so that plenty of articles about animals, insects and plants point to the Wikipedia article about him, which strongly increases the PageRank probability. This happens for all 24 languages where *Carl Linnaeus* always appears on high positions since articles about animals and plants are an important fraction of Wikipedia. Even if in a given language the top persons are often politicians (e.g. *Napoleon*, *Barak Obama* at $K = 1, 2$ in EN), these politicians have mainly local importance and are not highly ranked in other languages (e.g. in ZH *Carl Linnaeus* is at $K = 1$, *Napoleon* at $K = 3$ and *Barak Obama* is at $K = 24$). As a result when the global contribution is counted over all 24 languages *Carl Linnaeus* appears on the top PageRank position.

Our analysis suggests that there might be three groups of historical figures. Fig. 7 shows these three groups of top PageRank historical figures in Wikipedia: (i) global historical figures who appear in most of Wikipedia editions ($N_A \geq 18$) and are highly ranked ($\langle K \rangle \leq 50$) for each Wikipedia such as *Carl Linnaeus*, *Plato*, *Jesus*, and *Napoleon* (Right-Top of the Fig. 7A); (ii) local-highly ranked historical figures who appear in a few Wikipedia editions ($N_A < 18$) but are highly ranked ($\langle K \rangle \leq 50$) in the Wikipedia editions in which they appear, such as *Tycho Brahe*, *Sejong the Great*, and *Sun Yat-sen* (Left-Top of the Fig. 7A); (iii) locally-low

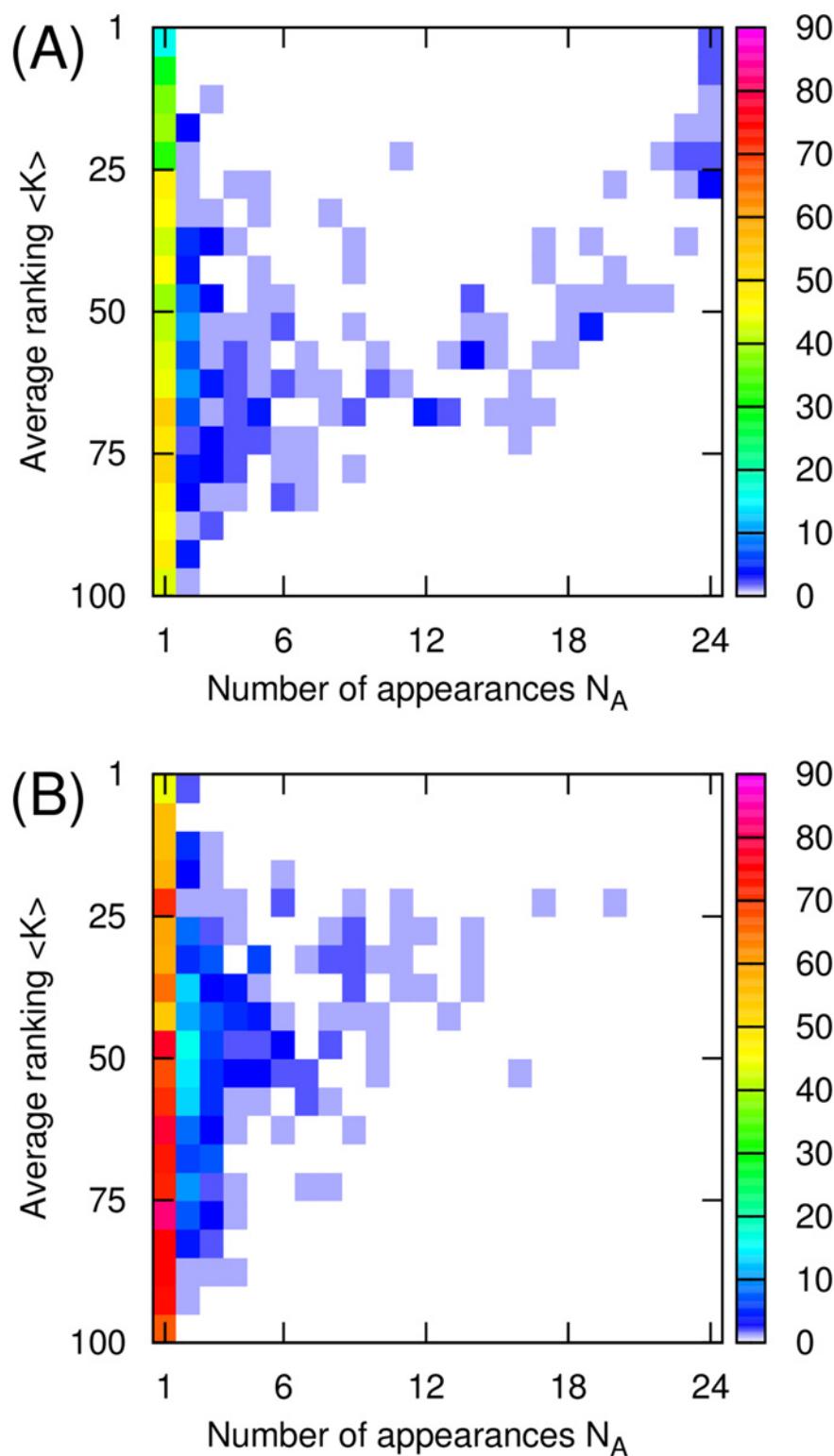


Fig 7. The distribution of 1045 top PageRank persons (A) and 1616 top 2DRank persons (B) as a function of number of appearances N_A of a given person and the rank $\langle K \rangle$ of this person averaged over Wikipedia editions where this person appeared.

doi:10.1371/journal.pone.0114825.g007

Table 5. List of the top 10 global female historical figures by PageRank and 2DRank for all the 24 Wikipedia editions. All names are represented by article titles in the English Wikipedia. Here, Θ_A is the ranking score of the algorithm A (Eq.3); N_A is the number of appearances of a given person in the top 100 rank for all editions. Here CC is the birth country code and LC is the language code of the given historical figure.

Rank	Θ_{PR}	N_A	PageRank female figures	CC	Century	LC
1	1789	24	Elizabeth II	UK	20	EN
2	1094	17	Mary (mother of Jesus)	IL	-1	HE
3	404	12	Queen Victoria	UK	19	EN
4	234	6	Elizabeth I of England	UK	16	EN
5	128	2	Maria Theresa	AT	18	DE
6	100	1	Benazir Bhutto	PK	20	HI
7	94	1	Catherine the Great	PL	18	PL
8	91	1	Anne Frank	DE	20	DE
9	87	1	Indira Gandhi	IN	20	HI
10	86	1	Margrethe II of Denmark	DK	20	DA

Rank	Θ_{2D}	N_A	2DRank female figures	CC	Century	LC
1	991	14	Madonna (entertainer)	US	20	EN
2	664	9	Elizabeth II	UK	20	EN
3	580	8	Mary (mother of Jesus)	IL	-1	HE
4	550	9	Queen Victoria	UK	19	EN
5	225	5	Agatha Christie	UK	19	EN
6	211	4	Mariah Carey	US	20	EN
7	206	7	Britney Spears	US	20	EN
8	200	3	Margaret Thatcher	UK	20	EN
9	191	2	Martina Navratilova	CZ	20	WR
10	175	2	Elizabeth I of England	UK	16	EN

doi:10.1371/journal.pone.0114825.t005

ranked historical figures who appear in a few Wikipedia editions ($N_A < 18$) and who are not highly ranked ($\langle K \rangle > 50$). Here N_A is the number of appearances in different Wikipedia editions for a given person and $\langle K \rangle$ is the average ranking of the given persons across Wikipedia editions for each ranking algorithm. In the case of 2DRank historical figures, due to the absence of global historical figures, most of them belong to two types of local historical figures (i.e. local-highly ranked or local-lowly ranked).

Following ranking of persons via $\Theta_{P, A}$ we determine also the top global female historical figures, presented in Table 5 for PageRank and 2DRank persons. The full lists of global female figures are available at [39] (63 and 165 names for PageRank and 2DRank).

The comparison of our 100 global historical figures with the top 100 from Hart's list [27] gives an overlap of 43 persons for PageRank and 26 persons for 2DRank. We note that for the top 100 from the English Wikipedia we obtain a lower overlap of 37 (PageRank) and 4 (2DRank) persons. Among all editions the highest overlaps with the Hart list are 42 (VI), 37 (EN, ES, PT, TR) and 33 (IT), 32 (DE), 31 (FR) for PageRank; while for 2DRank we find 18 (EL) and 17 (VI). We give the overlap numbers for all editions at [39]. This shows that the consideration of 24 editions provides us the global list of the top 100 persons with a more balanced selection of top historical figures. Our overlap of the top 100 global historical figures by PageRank with the top 100 people from Pantheon MIT ranking list [23] is 44 percent, while the overlap of this Pantheon list with Hart's list is 43 percent. We note that the Pantheon method is significantly based on a number of page views while our approach is based on the network structure of the whole Wikipedia network. The top 100 persons from [22] are not publicly available but nevertheless we present the overlaps between the top 100 persons from the lists of Hart, Pantheon,

Stony-Brook and our global PageRank and 2DRank lists in Figures S2, S3 (we received the Stony-Brook list as a private message from the authors of [22]). We have an average overlap between the 4 methods on a level of 40 percent (2DRank is on average lower by a few percent), we find a larger overlap between our PageRank list and the Stony-Brook list since the Stony-Brook method, applied only for the English Wikipedia, is significantly based on PageRank.

We also compared the distributions of our global top 100 persons of PageRank and 2DRank with the distribution of Hart's top 100 over centuries and over 24 languages with the additional WR category (see Figure S4). We find that these 3 distributions have very similar shapes. Thus the largest number of persons appears in centuries AD 18th, 19th, 20th for the 3 distributions. Among languages, the main peaks for the 3 distributions appear for EN, DE, IT, EL, AR, ZH. The deviations from Hart's distribution are larger for the 2DRank list. Thus the comparison of distributions over centuries and languages shows that the PageRank list has not only a strong overlap with the Hart list in the number of persons but that they also have very similar statistical distributions of the top 100 persons over centuries and languages.

The overlap of the top 100 global persons found here with the previous study [21] gives 54 and 47 percent for PageRank and 2DRank lists, respectively. However, we note that the global list in [21] was obtained from the top 30 persons in each edition while here we use the top 100 persons.

It is interesting to note that for the top 100 PageRank universities from the English Wikipedia edition the overlap with Shanghai top 100 list of universities is on a even higher level of 75 percent [18].

Finally, we note that the ranking of historical figures using the whole PageRank (or 2DRank) list of all Wikipedia articles of a given edition provides a more stable approach compared to the network of biographical articles used in [20]. Indeed, the number of nodes and links in such a biographical network is significantly smaller compared to the whole network of Wikipedia articles and thus the fluctuations become rather large. For example, from the biographical network of the Russian edition one finds as the top person *Napoleon III* (and even not *Napoleon I*) [20], who has a rather low importance for Russia. In contrast to that the present study gives us the top PageRank historical figure of the Russian edition to be *Peter the Great*, that has much more historical grounds. In a similar way for FR the results of [20] give at the first position *Adolf Hitler*, that is rather strange for the French culture, while we find a natural result *Napoleon*.

Network of cultures

We consider the selected top persons from each Wikipedia edition as important historical figures recognized by people who speak the language of that Wikipedia edition. Therefore, if a top person from a language edition *A* appears in another edition *B*, then we can consider this as a 'cultural' influence from culture *A* to *B*. Here we consider each language as a proxy for a cultural group and assign each historical figure to one of these cultural groups based on the most spoken language of her/his birth place at the country level. For example, *Adolf Hitler* was born in modern Austria and since German language is the most spoken language in Austria, he is considered as a German historical figure in our analysis. This method may lead to some misguiding results due to discrepancy between territories of country and cultures, e.g. *Jesus* was born in the modern State of Palestine (Bethlehem), which is an Arabic speaking country. Thus *Jesus* is from the Arabic culture in our analysis while usually one would say that he belongs to the Hebrew culture. Other similar examples we find are: *Charlemagne* (Belgium—Dutch), *Immanuel Kant* (Russia—Russian, while usually he is attributed to DE), *Moses* (Egypt—Arabic), *Catherine the Great* (Poland—Polish, while usually she would be attributed to DE or RU).

Table 6. Numbers of certain historical figures for top 100 list of each language: N_1 is the number of historical figures of a given language among the top 100 PageRank global historical figures; N_2 is the number of historical figures of a given language among the top 100 PageRank historical figures for the given language edition; N_3 is the number of historical figures of a given language among the top 100 2DRank global historical figures; N_4 is the number of historical figures of a given language among the top 100 2DRank historical figures for the given language edition.

Language	N_1	N_2	N_3	N_4	Language	N_1	N_2	N_3	N_4
EN	22	47	27	64	RU	2	29	3	27
NL	2	10	4	38	HE	2	17	2	22
DE	20	41	16	55	TR	2	27	2	54
FR	8	33	3	32	AR	8	42	5	69
ES	2	20	5	39	FA	0	46	1	64
IT	11	31	9	43	HI	1	65	0	76
PT	0	19	0	35	MS	0	15	0	40
EL	5	28	2	55	TH	0	46	0	53
DA	0	31	1	48	VI	0	7	0	30
SV	1	26	1	39	ZH	5	43	6	79
PL	1	20	2	26	KO	0	34	0	59
HU	0	18	0	18	JA	0	41	4	80
WR	8	-	7	-					

doi:10.1371/journal.pone.0114825.t006

In total there are such 36 cases from the global PageRank list of 1045 names (these 36 names are given in SI). However, in our knowledge, the birth place is the best way to assign a given historical figure to a certain cultural background computationally and systematically and with the data we have available. In total we have only about 3.4 percent of cases which can be discussed and where a native speaking language can be a better indicator of belonging to a given culture. For the global 2DRank list of 1616 names we identified 53 similar cases where an attribution to a culture via a native language or a birth place could be discussed (about 3.3 percent). These 53 names are given in SI. About half of such cases are linked with birth places in ancient Russian Empire where people from Belarus, Lithuania and Ukraine moved to RU, IL, PL, WR. However, the percentage of such cases is small and the corresponding errors also remain small.

Based on the above assumption and following the approach developed in [21], we construct two weighted networks of cultures (or language groups) based on the top PageRank historical figures and top 2DRank historical figures respectively. Each culture (i.e. language) is represented as a node of the network, and the weight of a directed link from culture A to culture B is given by the number of historical figures belonging to culture B (e.g. French) appearing in the list of top 100 historical figures for a given culture A (e.g. English). The persons in a given edition, belonging to the language of the edition, are not taken into account since they do not create links between cultures. In Table 6 we give the number of such persons for each language. This table also gives the number of persons of a given language among the top 100 persons of the global PageRank and 2DRank listings.

For example, there are 5 French historical figures among the top 100 PageRank historical figures of the English Wikipedia, so we can assign weight 5 to the link from English to French. Fig. 8A and Fig. 8B represent the constructed networks of cultures defined by appearances of the top PageRank historical figures and top 2DRank historical figures, respectively. In total we have two networks with 25 nodes which include our 24 editions and an additional node WR for all the other world cultures.

The Google matrix G_{ij} for each network is constructed following the standard rules described in [21] and in the Methods Section. In a standard way we determine the PageRank index K and the CheiRank index K^* that order all cultures according to decreasing PageRank

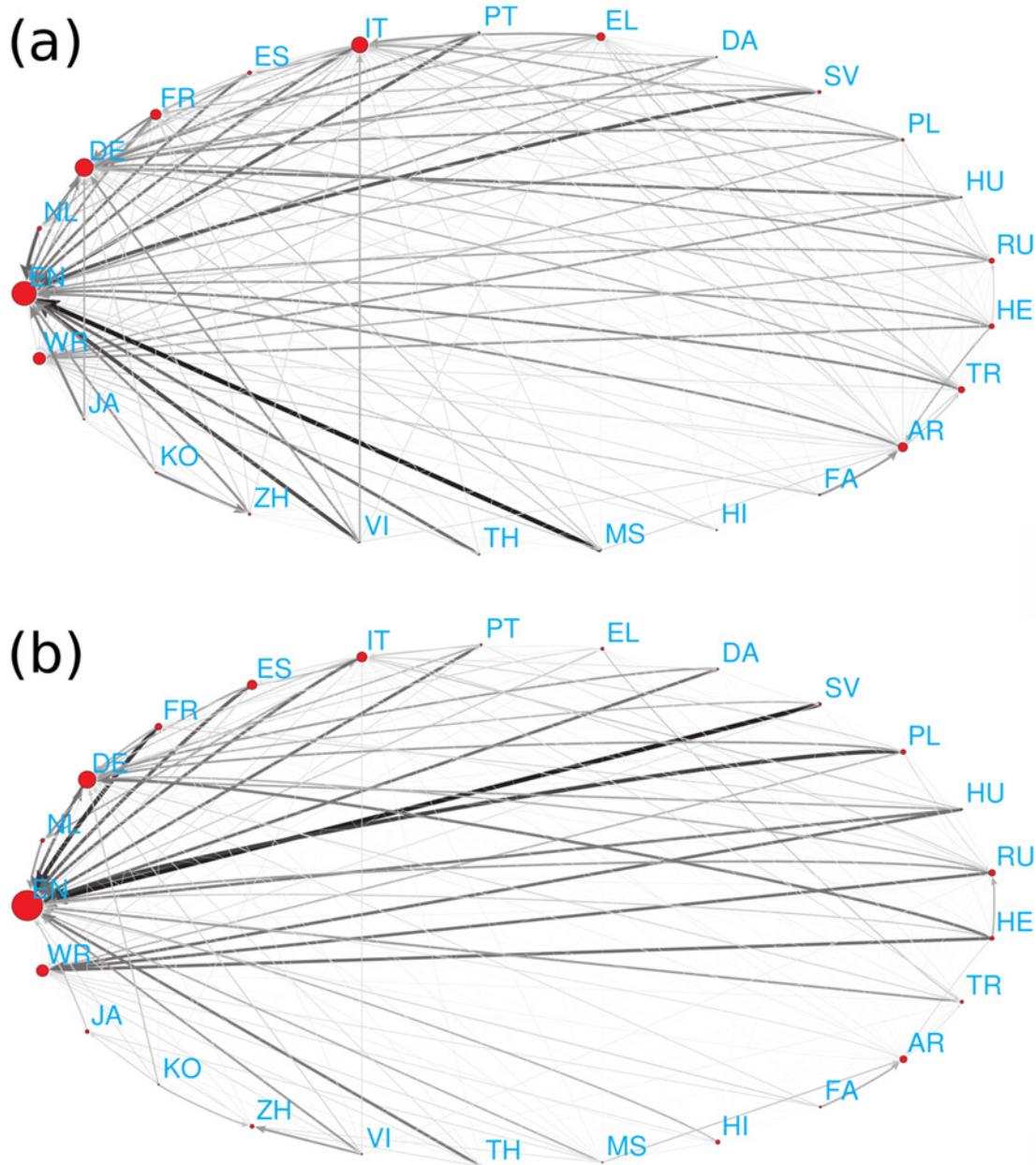


Fig 8. Network of cultures obtained from 24 Wikipedia languages and the remaining world (WR) consider (A) top PageRank historical figures and (B) 2DRank historical figures. The link width and darkness are proportional to a number of foreign historical figures quoted in top 100 of a given culture, the link direction goes from a given culture to cultures of quoted foreign historical figures, links inside cultures are not considered. The size of nodes is proportional to their PageRank.

doi:10.1371/journal.pone.0114825.g008

and CheiRank probabilities (see [Methods](#) and Figure S5). The structure of matrix elements $G_{KK'}$ is shown in [Fig. 9](#).

To identify which cultures (or language groups) are more influential than others, we calculated PageRank and CheiRank of the constructed networks of cultures by considering link weights. Briefly speaking, a culture has high PageRank (CheiRank) if it has many ingoing

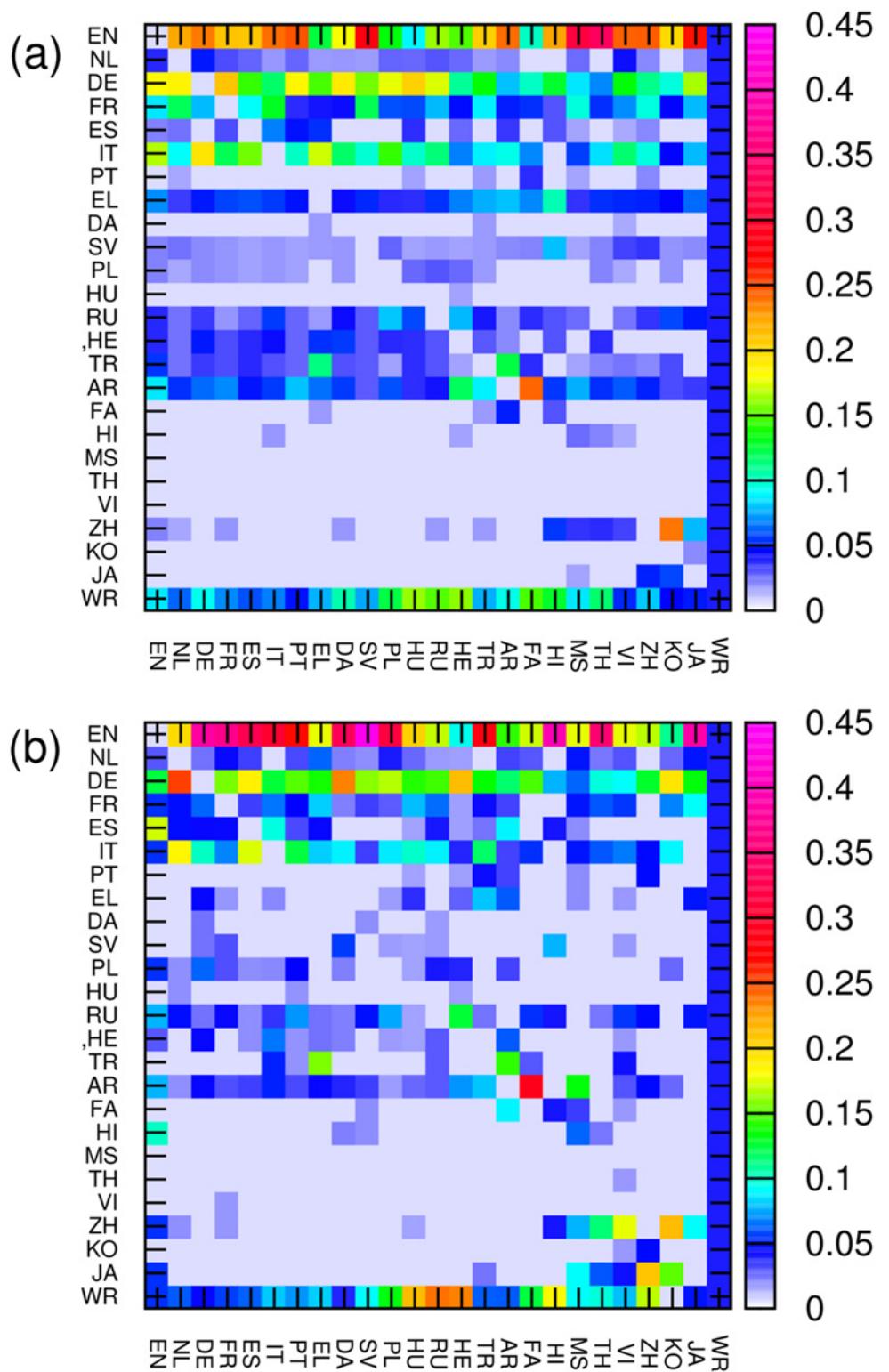


Fig 9. Google matrix of network of cultures shown in Fig. 8 respectively. The matrix elements G_{ij} are shown by color with damping factor $\alpha = 0.85$.

doi:10.1371/journal.pone.0114825.g009

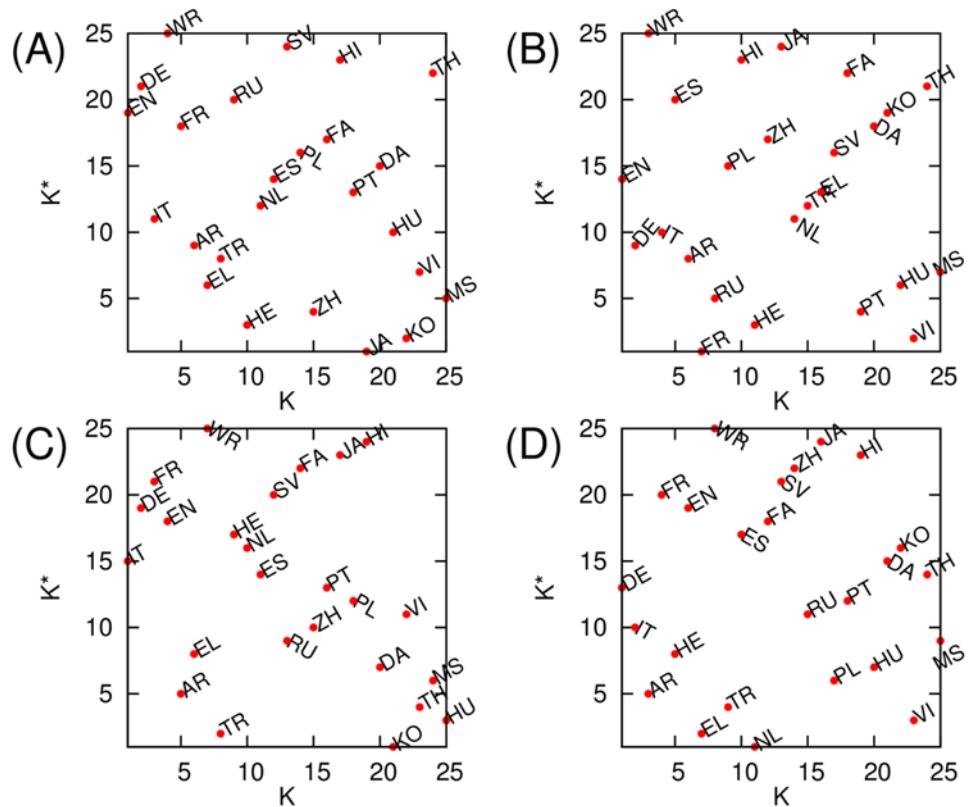


Fig 10. PageRank ranking versus CheiRank ranking plane of cultures with corresponding indexes K and K^* obtained from the network of cultures based on (A) all PageRank historical figures, (B) all 2DRank historical figures, (C) PageRank historical figure born before AD 19th century, and (D) 2DRank historical figure born before AD 19th century, respectively.

doi:10.1371/journal.pone.0114825.g010

(outgoing) links from (to) other cultures (see [Methods](#)). The distribution of cultures on a PageRank-CheiRank plane is shown in [Fig. 10](#). In both cases of PageRank and 2DRank historical figures, historical figures of English culture (i.e. born in English language spoken countries) are the most influential (highest PageRank) and German culture is the second one ([Fig. 10A, B](#)). Here we consider the historical figures for the whole range of centuries. [Fig. 10](#) represents the detailed features of how each culture is located on the plane of PageRank ranking K and CheiRank ranking K^* based on the top PageRank historical figures ([Fig. 10A](#)) and top 2DRank historical figures ([Fig. 10B](#)). Here K indicates the ranking of a given culture ordered by how many of its own top historical figures appear in other Wikipedia editions, and K^* indicates the ranking of a given culture according to how many of the top historical figures in the considered culture are from other cultures. As described above, English is on ($K = 1, K^* = 19$) and German is on ($K = 2, K^* = 21$) in the case of PageRank historical figures ([Fig. 10A](#)). In the case of 2DRank historical figures, English is on ($K = 1, K^* = 14$) and German is on ($K = 2, K^* = 9$).

It is important to note that there is a significant difference compared to the previous study [[21](#)]: there, only 9 editions had been considered and the top positions were attributed to the world node WR which captured a significant fraction of the top persons. This indicated that 9 editions are not sufficient to cover the whole world. Now for 24 editions we see that the importance of the world node WR is much lower (it moves from $K = 1$ for 9 editions [[21](#)] to $K = 4$ and 3 in [Fig. 10A](#) and [Fig. 10B](#)). Thus our 24 editions cover the majority the world. Still it

would be desirable to add a few additional editions (e.g. Ukraine, Baltic Republics, Serbia etc.) to fill certain gaps.

It is interesting to note that the ranking plane of cultures (K, K^*) changes significantly in time. Indeed, if we take into account only persons born before the 19th century then the ranking is modified with EN going to 4th ([Fig. 10C](#) for PageRank figures) and 6th position ([Fig. 10C](#) for 2DRank figures) while the top positions are taken by IT, DE, FR and DE, IT, AR, respectively.

At the same time, we may also argue that for cultures it is important not only to be cited but also to be communicative with other cultures. To characterize communicative properties of nodes on the network of cultures shown in [Fig. 8](#) we use again the concepts of PageRank, CheiRank and 2DRank for these networks as described in Methods and [\[21\]](#). Thus, for the network of cultures of [Fig. 8](#), the 2DRank index of cultures highlights their influence in a more balanced way taking into account their importance (incoming links) and communicative (outgoing links) properties in a balanced manner.

Thus we find for all centuries at the top positions Greek, Turkish and Arabic (for PageRank persons) and French, Russian and Arabic (for 2DRank persons). For historical figures before the 19th century, we find respectively Arabic, Turkish and Greek (for PageRank) and Arabic, Greek and Hebrew (for 2DRank). The high position of Turkish is due to its close links both with Greek culture in ancient times and with Arabic culture in more recent times. We see also that with time the positions of Greek in 2DRank improves due to a global improved ranking of Western cultures closely connected with Greece.

Discussion

By investigating birth place, birth date, and gender of important historical figures determined by the network structure of Wikipedia, we identified spatial, temporal, and gender skewness in Wikipedia. Our analysis shows that the most important historical figures across Wikipedia language editions were born in Western countries after the 17th century, and are male. Also, each Wikipedia edition highlights local figures so that most of its own historical figures are born in the countries which use the language of the edition. The emergence of such pronounced accent to local figures seems to be natural since there are more links and interactions within one culture. This is also visible from the fact that in many editions the main country for the given language is at the first PageRank position among all articles (e.g. Russia in RU edition) [\[21\]](#). Despite such a locality feature, there are also global historical figures who appear in most of the considered Wikipedia editions with very high rankings. Based on the cross-cultural historical figures, who appear in multiple editions, we can construct a network of cultures which describes interactions and entanglement between cultures.

It is very difficult to describe history in an objective way and due to that it was argued that history is “an unending dialogue between the past and present” [\[44\]](#). In a similar way we can say that history is an unending dialogue between different cultural groups.

We use a computational and data mining approach, based on rank vectors of the Google matrix of Wikipedia, to perform a statistical analysis of interactions and entanglement of cultures. We find that this approach can be used for selecting the most influential historical figures through an analysis of collectively generated links between articles on Wikipedia. Our results are coherent with studies conducted by historians [\[27\]](#), with an overlap of 43% of important historical figures. Thus, such a mathematical analysis of local and global historical figures can be a useful step towards the understanding of local and global history and interactions of world cultures. Our approach has some limitations, mainly caused by the data source and by the difficulty of defining culture boundaries across centuries. The ongoing improvement of structured

content in Wikipedia through the WikiData project, eventually in conjunction with additional manual annotation, should allow to deal with these limitations. Furthermore, it would be useful to perform comparisons with other approaches to measure the interactions of cultures, such as the analysis of language crossings of multilingual users [45].

Influence of digital media on information dissemination and social collective opinions among the public is growing fast. Our research across Wikipedia language editions suggests a rigorous mathematical way, based on Markov chains and Google matrix, for the identification of important historical figures and for the analysis of interactions of cultures at different historical periods and in different world regions. We think that a further extension of this approach to a larger number of Wikipedia editions will provide a more detailed and balanced analysis of interactions of world cultures.

Supporting Information

S1 File. Supporting Information file S1 presents Figures S1–S5 with additional information discussed above in the main part of the paper, lists of top 100 global PageRank and 2DRank names; Tables S1–S25 of top 10 names of given language and remained world from the global PageRank and 2DRank ranking lists of persons ordered by the score $\Theta_{P, A}$ of Eq.(3). For a reader convenience the lists of all 100 ranked names for all 24 Wikipedia editions and corresponding network link data for each edition are also given at [39] in addition to Supporting Information file. All used computational data are publicly available at <http://dumps.wikimedia.org/>. All the raw data necessary to replicate the findings and conclusion of this study are within the paper, supporting information files and this Wikimedia web site. (PDF)

Acknowledgments

This research is supported in part by the EC FET Open project “New tools and algorithms for directed network analysis” (NADINE \$No\$ 288956)

Author Contributions

Conceived and designed the experiments: YHE DLS. Performed the experiments: YHE PA DL AK SV. Analyzed the data: YHE PA DLS. Contributed reagents/materials/analysis tools: YHE PA DL AK SV. Wrote the paper: YHE DLS. Conceived and designed the experiments: YHE DLS. Performed the experiments: YHE PA DL AK SV. Analyzed the data: YHE PA DLS. Contributed reagents/materials/analysis tools: YHE PA DL AK SV. Wrote the paper: YHE DLS.

References

1. Rosenzweig R (2006) Can history be open source? Wikipedia and the future and the past, *Journal of American History* 93(1): 117 doi: [10.2307/4486062](https://doi.org/10.2307/4486062)
2. Lavsa SM, Corman SL, Culley CM, Pummer TL (2011) Reliability of Wikipedia as a medication information source for pharmacy students, *Currents in Pharmacy Teaching and Learning* 3(2): 154–158 doi: [10.1016/j.cptl.2011.01.007](https://doi.org/10.1016/j.cptl.2011.01.007)
3. Giles J (2005) Internet encyclopedia go head to head, *Nature*, 438: 900 doi: [10.1038/438900a](https://doi.org/10.1038/438900a) PMID: [16355180](https://pubmed.ncbi.nlm.nih.gov/16355180/)
4. Kittur A, Chi EH, Suh B (2009) What's in Wikipedia?: mapping topics and conflict using socially annotated category structure, In Proc. of SIGCHI Conference on Human Factors in Computing Systems, CHI'09, ACM, New York
5. Priedhorsky R, Chen J, Lam STK, Panciera K, Terveen L et al. (2007). Creating, Destroying, and Restoring Value in Wikipedia, In Proceedings of the Intl. Conf. on Supporting Group Work, 295, ACM, New York

6. Yasseri T, Sumi R, Rung A, Kornai A, Kertész J (2012) Dynamics of Conflicts in Wikipedia, PLoS ONE 7(6): e38869 doi: [10.1371/journal.pone.0038869](https://doi.org/10.1371/journal.pone.0038869) PMID: [22745683](https://pubmed.ncbi.nlm.nih.gov/22745683/)
7. Yasseri T, Spoerri A, Graham M, Kertész J (2013) The most controversial topics in Wikipedia: a multilingual and geographical analysis arXiv:1305.5566 [physics.soc-ph]
8. Laniado D, Tasso R, Volkovich Y, Kaltenbrunner A (2011) When the wikipedians talk: Network and tree structure of Wikipedia discussion pages, Proc. ICWSM 2011: 177–184
9. UNESCO World Report (2009) Investing in cultural diversity and intercultural dialogue, Available: <http://www.unesco.org/new/en/culture/resources/report/the-unesco-world-report-on-cultural-diversity>
10. Wikipedia: Neutral point of view. Retrieved May 12, 2014 from http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view
11. Pfeil U, Zaphiris P, Ang C A, (2006) Cultural Differences in Collaborative Authoring of Wikipedia, J. Computer-Mediated Comm. 12(1): 88 doi: [10.1111/j.1083-6101.2006.00316.x](https://doi.org/10.1111/j.1083-6101.2006.00316.x)
12. Callahan ES, Herring SC (2011) Cultural bias in Wikipedia content on famous persons, Journal of the American society for information science and technology 62: 1899 doi: [10.1002/asi.21577](https://doi.org/10.1002/asi.21577)
13. Hecht B, Gergle D (2009) Measuring self-focus bias in community-maintained knowledge repositories, Proc. of the Fourth Intl Conf. Communities and technologies, ACM, New York 2009: 11
14. Hecht B, Gergle D (2010) The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context, Proc. of SIGCHI Conference on Human Factors in Computing Systems, CHI'10, Atlanta, ACM, New York 291–300p
15. Nemoto K, Gloor PA (2011) Analyzing cultural differences in collaborative innovation networks by analyzing editing behavior in different-language Wikipedias, Procedia—Social and Behavioral Sciences 26: 180 doi: [10.1016/j.sbspro.2011.10.574](https://doi.org/10.1016/j.sbspro.2011.10.574)
16. Warncke-Wang M, Uduwage A, Dong Z, Riedl J (2012) In search of the ur-Wikipedia: universality, similarity, and translation in the Wikipedia inter-language link network, Proceedings of the 8th Intl. Symposium on Wikis and Open Collaboration (WikiSym 2012), ACM, New York
17. Massa P, Scrinzi F (2012) Manypedia: Comparing language points of view of Wikipedia communities, Proceedings of the 8th Intl. Symposium on Wikis and Open Collaboration (WikiSym 2012), ACM, New York
18. Zhirov AO, Zhirov OV, Shepelyansky DL (2010) Two-dimensional ranking of Wikipedia articles, Eur. Phys. J. B 77: 523 doi: [10.1140/epjb/e2010-10500-7](https://doi.org/10.1140/epjb/e2010-10500-7)
19. Eom YH, Frahm KM, Bencur A, Shepelyansky DL (2013) Time evolution of Wikipedia network ranking, Eur. Phys. J. B, 86:482 doi: [10.1140/epjb/e2013-40432-5](https://doi.org/10.1140/epjb/e2013-40432-5)
20. Aragón P, Laniado D, Kaltenbrunner A, Volkovich Y (2012) Biographical social networks on Wikipedia: a cross-cultural study of links that made history, Proc. of the 8th Intl. Symposium on Wikis and Open Collaboration (WikiSym 2012), ACM, New York No 19
21. Eom YH, Shepelyansky DL (2013) Highlighting Entanglement of Cultures via Ranking of Multilingual Wikipedia Articles, PLoS ONE, 8(10): e74554 doi: [10.1371/journal.pone.0074554](https://doi.org/10.1371/journal.pone.0074554) PMID: [24098338](https://pubmed.ncbi.nlm.nih.gov/24098338/)
22. Skiena S, Ward CB (2013) Who is Bigger?: Where Historical Figures Really Rank, Cambridge University Press, Cambridge UK
23. MIT Pantheon project. Available: <http://pantheon.media.mit.edu>. Accessed 2014 May 12.
24. Samoilenco A, Yasseri T (2014) The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics, EPJ Data Sci. 3: 1 doi: [10.1140/epjds20](https://doi.org/10.1140/epjds20)
25. Wikipedia: List of languages by number of native speakers. Retrieved May 12, 2014 from http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers
26. Wikipedia: Wikipedia. Retrieved May 12, 2014 from <http://en.wikipedia.org/wiki/Wikipedia>
27. Hart MH (1992) The 100: ranking of the most influential persons in history, Citadel Press, N.Y.
28. Ermann L, Chepeliantskii AD, Shepelyansky DL (2012) Toward two-dimensional search engines, J. Phys. A: Math. Theor. 45: 275101 doi: [10.1088/1751-8113/45/27/275101](https://doi.org/10.1088/1751-8113/45/27/275101)
29. Ermann L, Frahm KM, Shepelyansky DL (2013) Spectral properties of Google matrix of Wikipedia and other networks, Eur. Phys. J. D 86: 193 doi: [10.1140/epjb/e2013-31090-8](https://doi.org/10.1140/epjb/e2013-31090-8)
30. Langville AM, Meyer CD (2006) Google's PageRank and Beyond: The Science of Search Engine Rankings, Princeton University Press, Princeton
31. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems 30: 107 doi: [10.1016/S0169-7552\(98\)00110-X](https://doi.org/10.1016/S0169-7552(98)00110-X)
32. Chen P, Xie H, Maslov S, Redner S (2007) Finding scientific gems with Google PageRank algorithm, Jour. Informetrics, 1: 8 doi: [10.1016/j.joi.2006.06.001](https://doi.org/10.1016/j.joi.2006.06.001)

33. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media?, Proc. 19th Int. Conf. WWW2010, ACM, New York 591p
34. Ermann L, Shepelyansky DL (2011) Google matrix of the world trade network, *Acta Physica Polonica A* 120(6A), A158
35. Kandiah V, Shepelyansky DL (2013) Google matrix analysis of DNA sequences, *PLoS ONE* 8(5): e61519 doi: [10.1371/journal.pone.0061519](https://doi.org/10.1371/journal.pone.0061519) PMID: [23671568](#)
36. Chepelianskii AD (2010) Towards physical laws for software architecture, *arXiv:1003.5455 [cs.SE]*
37. Lü L, Zhang Y-C, Yeung CH, Zhou T (2011) Leaders in social networks, the delicious case, *PLoS ONE* 6(6): e21202 doi: [10.1371/journal.pone.0021202](https://doi.org/10.1371/journal.pone.0021202) PMID: [21738620](#)
38. <http://dbpedia.org>. Accessed 2014 May 12
39. Top wikipeople. Available: <http://www.quantware.ups-tlse.fr/QWLIB/topwikipeople/>. Accessed 2014 May 12.
40. United States Census Bureau. Retrieved May 12, 2014 from http://www.census.gov/population/international/data/worldpop/table_history.php
41. Wikipedia: Wikipedians. Retrieved May 12, 2014 from <http://en.wikipedia.org/wiki/Wikipedia:Wikipedians>
42. Statistics and indicators on women and men by United Nation. <http://unstats.un.org/unsd/Demographic/products/indwm/> (accesible May 12, 2014)
43. Lam STK, Uduwage A, Dong Z, Sen S (2011) WP:clubhouse?: an exploration of Wikipedia's gender imbalance, Proc. of the 7th Intl. Symposium on Wikis and Open Collaboration, WikiSym'11, Moutain View 1–10p
44. Carr EH (1961) What is History?, Vintage Books, New York
45. Hale SA (2014) Multilinguals and Wikipedia editing, Proc. 6th Annual ACM Web Science Conf. ACM New York 1, 99

SUPPORTING INFORMATION FOR:

Interactions of cultures and top people of Wikipedia from ranking of 24 language editions

Young-Ho Eom¹, Pablo Aragón², David Laniado², Andreas Kaltenbrunner², Sebastiano Vigna³, Dima L. Shepelyansky^{1,*}

¹ *Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, F-31062 Toulouse, France*

² *Barcelona Media Foundation, Barcelona, Spain*

³ *Dipartimento di Informatica, Università degli Studi di Milano, Milano, Italy*

* Corresponding Author E-mail: dima@irsamc.ups-tlse.fr

1 Additional data

Here we present additional figures and tables for the main part of the paper.

Figure S1 is analogous to Figures 4(C,D,E,F), however, now on the vertical axis we plot not the edition to which a given historical figure is attributed from top 100 figures of a given edition but the language, to which this historical figure from the global PageRank (1045 persons) or 2DRank (1616 persons) lists is attributed according to our procedure according to her/his country of birth and then to the major language of this country, if a person does not belong to any of 24 languages then he/she is attributed to the remaining world (WR). The data show that the separation between language (or culture) groups becomes now more distinct. Indeed, attribution to a language related to a birth place is more definite compared to the option where a person appears in one of 24 editions since some global historical figures appear in a few editions while each person is attributed only one language according to our procedure.

Figure S2 shows overlap between the global list of top 100 global PageRank persons and list of Hart [23], PageRank list of English Wikipedia from [15], list of Stony-Brook [19], list of Pantheon MIT project [20].

Figure S3 shows the overlap matrix (in percent) between 5 methods of ranking of top 100 historical figures including Hart, Pantheon, Stony-Brook results and our global PageRank and 2DRank lists. We see that our PageRank has most high correlation with Stony-Brook since the method of Stony-Brook uses significantly the PageRank method.

Figure S4 shows the number of persons from top 100 lists of Hart and our global PageRank and 2DRank lists. The panel (A) shows the number of persons at a given century corresponding to the time dependence and the panel (B) shows distribution of such persons over the language they are attributed according to our method based on the birth place and dominant language of a country of birth. We see that the pattern of Hart ranking is well reproduced from our global ranking, especially for the case of PageRank list.

Figure S5 shows PageRank and CheiRank probabilities for the networks of cultures shown in Figure 8.

The names of persons from top 100 missed by automatic recovery of persons are: Homer, Charles Darwin (RU PageRank); Philipp Kirkorov (RU 2DRank); Alexander the Great, Emperor Gaozu of Han, Homer (KO PageRank); Jinpyeong of Silla, Hyeonjong of Goryeo (KO 2DRank).

Unfortunately, the name of Homer has been missed in the 1.1 million list of English names, other names are missed due to incompleteness and modifications of inter-language translations.

Below we give the list of global top 100 PageRank names from 24 Wikipedia editions. The names are ordered by the ranking score $\Theta_{P,A}$ of Eq.(1). In brackets we give country of birth, century of birth, gender, and language of birth. In the same manner we also give the list of top 100 2DRank names from 24 Wikipedia editions.

We also give 24 names from global 1045 PageRank names and 40 names from 1616 global 2DRank names where a birth place language attribution differs from native language.

We also give the tables of top 10 persons in each language and also world names (tables S1 - S25) extracted from the global PageRank and 2DRank ranking lists of persons ordered by the score $\Theta_{P,A}$ of Eq.(1).

Top 100 of global PageRank names: 1. Carl Linnaeus (SE, 18, M, SV) 2. Jesus (PS, -1, M, AR) 3. Aristotle (GR, -4, M, EL) 4. Napoleon (FR, 18, M, FR) 5. Adolf Hitler (AT, 19, M, DE) 6. Julius Caesar (IT, -1, M, IT) 7. Plato (GR, -5, M, EL) 8. William Shakespeare (UK, 16, M, EN) 9. Albert Einstein (DE, 19, M, DE) 10. Elizabeth II (UK, 20, F, EN) 11. Alexander the Great (GR, -4, M, EL) 12. Isaac Newton (UK, 17, M, EN) 13. Muhammad (SA, 6, M, AR) 14. Karl Marx (DE, 19, M, DE) 15. Joseph Stalin (GE, 19, M, WR) 16. Augustus (IT, -1, M, IT) 17. Christopher Columbus (IT, 15, M, IT) 18. Charlemagne (BE, 8, M, NL) 19. Louis XIV of France (FR, 17, M, FR) 20. George W. Bush (US, 20, M, EN) 21. Immanuel Kant (RU, 18, M, RU) 22. Barack Obama (US, 20, M, EN) 23. Mary (mother of Jesus) (IL, -1, F, HE) 24. Vladimir Lenin (RU, 19, M, RU) 25. Wolfgang Amadeus Mozart (AT, 18, M, DE) 26. Paul the Apostle (TR, 1, M, TR) 27. Charles Darwin (UK, 19, M, EN) 28. Martin Luther (DE, 15, M, DE) 29. Herodotus (TR, -5, M, TR) 30. Franklin D. Roosevelt (US, 19, M, EN) 31. Galileo Galilei (IT, 16, M, IT) 32. Pope John Paul II (PL, 20, M, PL) 33. Constantine the Great (RS, 3, M, WR) 34. Benito Mussolini (IT, 19, M, IT) 35. Cicero (IT, -2, M, IT) 36. Ren Descartes (FR, 16, M, FR) 37. Saint Peter (IL, 1, M, HE) 38. Ludwig van Beethoven (DE, 18, M, DE) 39. George Washington (US, 18, M, EN) 40. Moses (EG, -14, M, AR) 41. Johann Sebastian Bach (DE, 17, M, DE) 42. Bill Clinton (US, 20, M, EN) 43. Leonardo da Vinci (IT, 15, M, IT) 44. Johann Wolfgang von Goethe (DE, 18, M, DE) 45. Gautama Buddha (NP, -6, M, WR) 46. Winston Churchill (UK, 19, M, EN) 47. John F. Kennedy (US, 20, M, EN) 48. Charles V, Holy Roman Emperor (BE, 15, M, NL) 49. Pope Benedict XVI (DE, 20, M, DE) 50. Richard Nixon (US, 20, M, EN) 51. Sigmund Freud (CZ, 19, M, WR) 52. Ronald Reagan (US, 20, M, EN) 53. Abraham Lincoln (US, 19, M, EN) 54. Saddam Hussein (IQ, 20, M, AR) 55. Ptolemy (EG, 1, M, AR) 56. Richard Wagner (DE, 19, M, DE) 57. Diocletian (HR, 3, M, WR) 58. Queen Victoria (UK, 19, F, EN) 59. Napoleon III (FR, 19, M, FR) 60. Charles de Gaulle (FR, 19, M, FR) 61. Mao Zedong (CN, 19, M, ZH) 62. William Herschel (DE, 18, M, DE) 63. Michael Jackson (US, 20, M, EN) 64. Justinian I (MK, 5, M, WR) 65. Augustine of Hippo (DZ, 4, M, AR) 66. Ali (SA, 7, M, AR) 67. Jean-Jacques Rousseau (CH, 18, M, DE) 68. Ernst Haeckel (DE, 19, M, DE) 69. Pliny the Elder (IT, 1, M, IT) 70. Pope Gregory XIII (IT, 16, M, IT) 71. Confucius (CN, -6, M, ZH) 72. Henry VIII of England (UK, 15, M, EN) 73. Thomas Jefferson (US, 18, M, EN) 74. Francisco Franco (ES, 19, M, ES) 75. Georg Wilhelm Friedrich Hegel (DE, 18, M, DE) 76. Pierre Andr Latreille (FR, 18, M, FR) 77. Pope Paul VI (IT, 19, M, IT) 78. Gottfried Wilhelm Leibniz (DE, 17, M, DE) 79. Chiang Kai-shek (CN, 19, M, ZH) 80. John Herschel (UK, 18, M, EN) 81. Elizabeth I of England (UK, 16, F, EN) 82. J. R. R. Tolkien

(ZA, 19, M, WR) 83. Socrates (GR, -5, M, EL) 84. Genghis Khan (MN, 12, M, WR) 85. Qin Shi Huang (CN, -3, M, ZH) 86. Umar (SA, 6, M, AR) 87. Philip II of Spain (ES, 16, M, ES) 88. Frederick the Great (DE, 18, M, DE) 89. Johannes Kepler (DE, 16, M, DE) 90. Emperor Wu of Han (CN, -2, M, ZH) 91. Friedrich Nietzsche (DE, 19, M, DE) 92. Plutarch (GR, 1, M, EL) 93. Thomas Edison (US, 19, M, EN) 94. Max Weber (DE, 19, M, DE) 95. Dante Alighieri (IT, 13, M, IT) 96. Ashoka (IN, -4, M, HI) 97. Tacitus (FR, 1, M, FR) 98. Ernst Mayr (DE, 20, M, DE) 99. Jean-Baptiste Lamarck (FR, 18, M, FR) 100. Elvis Presley (US, 20, M, EN).

Top 100 of global 2DRank names: 1. Adolf Hitler (AT, 19, M, DE) 2. Michael Jackson (US, 20, M, EN) 3. Madonna (entertainer) (US, 20, F, EN) 4. Jesus (PS, -1, M, AR) 5. Ludwig van Beethoven (DE, 18, M, DE) 6. Wolfgang Amadeus Mozart (AT, 18, M, DE) 7. Pope Benedict XVI (DE, 20, M, DE) 8. Alexander the Great (GR, -4, M, EL) 9. Charles Darwin (UK, 19, M, EN) 10. Barack Obama (US, 20, M, EN) 11. Johann Sebastian Bach (DE, 17, M, DE) 12. Napoleon (FR, 18, M, FR) 13. Pope John Paul II (PL, 20, M, PL) 14. Julius Caesar (IT, -1, M, IT) 15. Elizabeth II (UK, 20, F, EN) 16. Albert Einstein (DE, 19, M, DE) 17. Augustus (IT, -1, M, IT) 18. Bob Dylan (US, 20, M, EN) 19. Leonardo da Vinci (IT, 15, M, IT) 20. Mary (mother of Jesus) (IL, -1, F, HE) 21. Charlemagne (BE, 8, M, NL) 22. William Shakespeare (UK, 16, M, EN) 23. Elvis Presley (US, 20, M, EN) 24. Queen Victoria (UK, 19, F, EN) 25. John Lennon (UK, 20, M, EN) 26. George Frideric Handel (DE, 17, M, DE) 27. J. R. R. Tolkien (ZA, 19, M, WR) 28. Muhammad (SA, 6, M, AR) 29. Joseph Stalin (GE, 19, M, WR) 30. Karl Marx (DE, 19, M, DE) 31. Benito Mussolini (IT, 19, M, IT) 32. Franklin D. Roosevelt (US, 19, M, EN) 33. Michael Schumacher (DE, 20, M, DE) 34. Paul McCartney (UK, 20, M, EN) 35. Stephen King (US, 20, M, EN) 36. Henry VIII of England (UK, 15, M, EN) 37. Tokugawa Ieyasu (JP, 16, M, JA) 38. Edgar Allan Poe (US, 19, M, EN) 39. Martin Luther (DE, 15, M, DE) 40. David Bowie (UK, 20, M, EN) 41. Pope Pius XII (IT, 19, M, IT) 42. Alfred Hitchcock (UK, 19, M, EN) 43. Friedrich Nietzsche (DE, 19, M, DE) 44. Vladimir Putin (RU, 20, M, RU) 45. Christopher Columbus (IT, 15, M, IT) 46. Elton John (UK, 20, M, EN) 47. Carl Linnaeus (SE, 18, M, SV) 48. Michelangelo (IT, 15, M, IT) 49. Raphael (IT, 15, M, IT) 50. Roger Federer (CH, 20, M, DE) 51. Cao Cao (CN, 2, M, ZH) 52. Vincent van Gogh (NL, 19, M, NL) 53. Frdric Chopin (PL, 19, M, PL) 54. Steven Spielberg (US, 20, M, EN) 55. Rembrandt (NL, 17, M, NL) 56. Ali (SA, 7, M, AR) 57. Richard Wagner (DE, 19, M, DE) 58. Che Guevara (AR, 20, M, ES) 59. Nelson Mandela (ZA, 20, M, WR) 60. Isaac Asimov (RU, 20, M, RU) 61. Jules Verne (FR, 19, M, FR) 62. Toyotomi Hideyoshi (JP, 16, M, JA) 63. Winston Churchill (UK, 19, M, EN) 64. Paul the Apostle (TR, 1, M, TR) 65. Hirohito (JP, 20, M, JA) 66. 14th Dalai Lama (CN, 20, M, ZH) 67. Franz Liszt (AT, 19, M, DE) 68. Genghis Khan (MN, 12, M, WR) 69. Otto von Bismarck (DE, 19, M, DE) 70. Saint Peter (IL, 1, M, HE) 71. Charlie Chaplin (UK, 19, M, EN) 72. Liu Bei (CN, 2, M, ZH) 73. Oda Nobunaga (JP, 16, M, JA) 74. Suleiman the Magnificent (TR, 15, M, TR) 75. Cyrus the Great (IR, -6, M, FA) 76. George W. Bush (US, 20, M, EN) 77. Agatha Christie (UK, 19, F, EN) 78. Carl Friedrich Gauss (DE, 18, M, DE) 79. Louis XIV of France (FR, 17, M, FR) 80. Saddam Hussein (IQ, 20, M, AR) 81. Pablo Picasso (ES, 19, M, ES) 82. Mariah Carey (US, 20, F, EN) 83. Hans Christian Andersen (DK, 19, M, DA) 84. Plato (GR, -5, M, EL) 85. Britney Spears (US, 20, F, EN) 86. Rafael Nadal (ES, 20, M, ES) 87. George Harrison (UK, 20, M, EN) 88. Margaret Thatcher (UK, 20, F, EN) 89. Jorge Luis Borges (AR, 19, M, ES) 90. Salvador Dal (ES, 20, M, ES) 91. Peter the Great (RU, 17, M, RU) 92. Giuseppe Verdi (IT, 19, M, IT) 93. Sigmund Freud (CZ, 19, M,

WR) 94. Qin Shi Huang (CN, -3, M, ZH) 95. Kangxi Emperor (CN, 17, M, ZH) 96. Martina Navratilova (CZ, 20, F, WR) 97. Charles V, Holy Roman Emperor (BE, 15, M, NL) 98. Zhuge Liang (CN, 2, M, ZH) 99. Constantine the Great (RS, 3, M, WR) 100. Muammar Gaddafi (LY, 20, M, AR)

List of 36 names from the global PageRank list of 1045 names where the birth place in modern geography of countries differs from native language: Jesus (PS AR), Charlemagne (Belgium NL), Immanuel Kant (Russia RU), Moses (Egypt AR), Catherine the Great (Poland PL), Mustafa Kemal Atatürk (Greece EL), Bhumibol Adulyadej (USA EN), Christian V of Denmark (Germany DE), Józef Pilsudski (Litvania WR), Christian IX of Denmark (Germany DE), Philip V of Spain (France FR), Giuseppe Garibaldi (France FR), Muhammad al-Idrisi (Spain ES), Charles XIV John of Sweden (France FR), Leonid Brezhnev (Ukraine WR), George I of Greece (Denmark DA), Juan Carlos I of Spain (Italy IT), Leon Trotsky (Ukraine WR), Golda Meir (Ukraine WR), Valéry Giscard d'Estaing (Germany DE), Magnus IV of Sweden (Noroway WR), Christian I of Denmark (Germany DE), Yitzhak Ben-Zvi (Ukraine WR), Mikhail Bulgakov (Ukraine WR); Kim Jong-il (Russia RU). Lee Myung-bak (Japan JA), Jangsu of Goguryeo (China ZH); Galyani Vadhana (UK EN), Abhisit Vejjajiva (UK EN); Matthias Corvinus (Romania WR), Ferenc Kazinczy (Romania WR), György Kulin (Romania WR), Gabriel Bethlen (Romania WR), Endre Ady (Romania WR), János Arany (Romania WR), Béla Bartók (Romania WR).

List of 53 names from the global 2DRank list of 1616 names where the birth place in modern geography of countries differs from native language: Jesus (PS AR), Charlemagne (BE NL), Isaac Asimov (RU RU), Paul the Apostle (TR TR), Peter Paul Rubens (DE DE), Catherine the Great (PL PL), Julian (emperor) (TR TR), Józef Pilsudski (LT WR), Muhammad Ali of Egypt (GR EL), Juan Carlos I of Spain (IT IT), Shmuel Yosef Agnon (UA WR), Saint Joseph (PS AR), Golda Meir (UA WR), Baibars (UA WR), Levi Eshkol (UA WR), Augustine of Hippo (DZ AR), Yitzhak Ben-Zvi (UA WR), Natan Yonatan (UA WR), Edward Rydz-migy (UA WR), Immanuel Kant (RU RU), Pyotr Stolypin (DE DE), Czeslaw Niemen (BY RU), Moses (EG AR), Albert Camus (DZ AR), Leonid Brezhnev (UA WR), Aharon Barak (LT WR), George Orwell (IN HI), Sergei Korolev (UA WR), Garry Kasparov (AZ TR), Ibn 'Abd al-Barr (ES ES), Georges Simenon (BE NL), Ryszard Kapuściński (BY RU), Mihly Munkácsy (UA WR), Juliusz Slowacki (UA WR), Tadeusz Kościuszko (BY RU), John McCain (PA ES), Maurice, Prince of Orange (DE DE), Zbigniew Herbert (UA WR), Leon Trotsky (UA WR), Charles XIV John of Sweden (FR FR). Lee Myung-bak (JA JA), Jangsu of Goguryeo (CN ZH), Gwanggaeto the Great (CN ZH); Galyani Vadhana (UK EN), Abhisit Vejjajiva (UK EN); Matthias Corvinus (RO WR), Károly Kós (RO WR), László Németh (RO WR), Sándor Körösi Csoma (RO WR), János Bolyai (RO WR), György Kulin (RO WR), Ferenc Kazinczy (RO WR), Béla Bartók (RO WR).

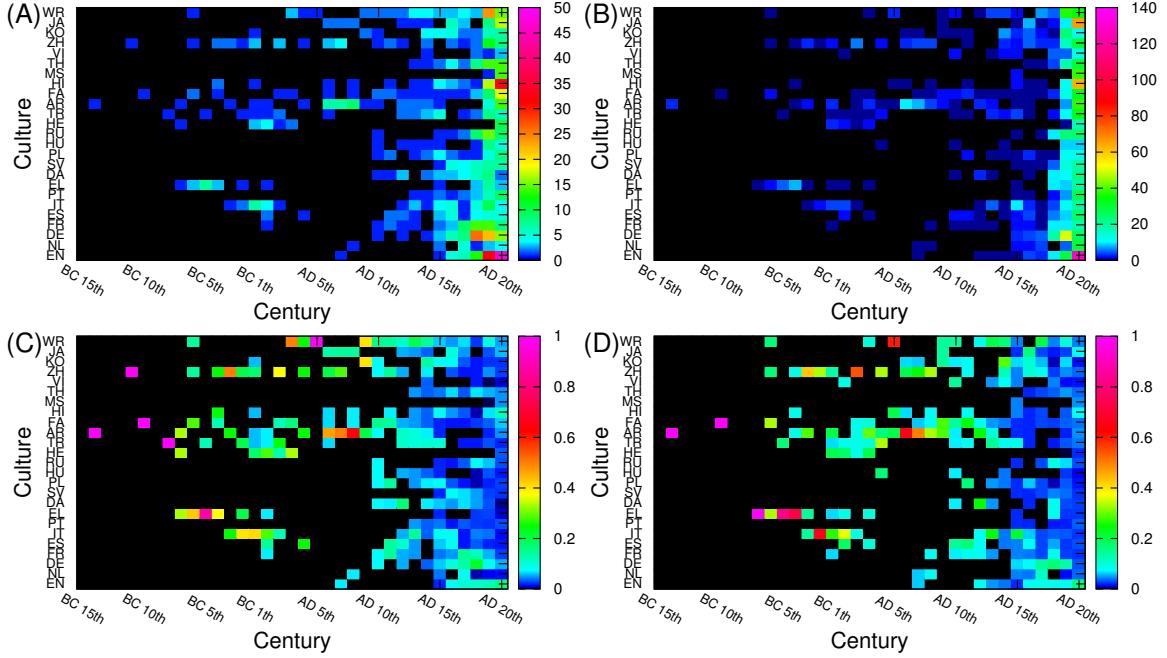


Figure S1. Birth date distribution of historical figures from the global PageRank list (A,C, 1045 persons) and 2DRank list (B,D, 1616 persons). Each historical figure is attributed to her/his own language according to her/his birth place as described in the paper (if the birth place is not among our 24 languages then a person is attributed to the remaining world (WR)). Color in panels (A,B) shows the total number of persons for a given century, while in panels (C,D) color shows a percent for a given century (normalized to unity in each column). This figure give a more distinct separation of cultures (languages) compared to a similar Fig.4 where the distribution over Wikipedia editions is shown on the vertical axis.

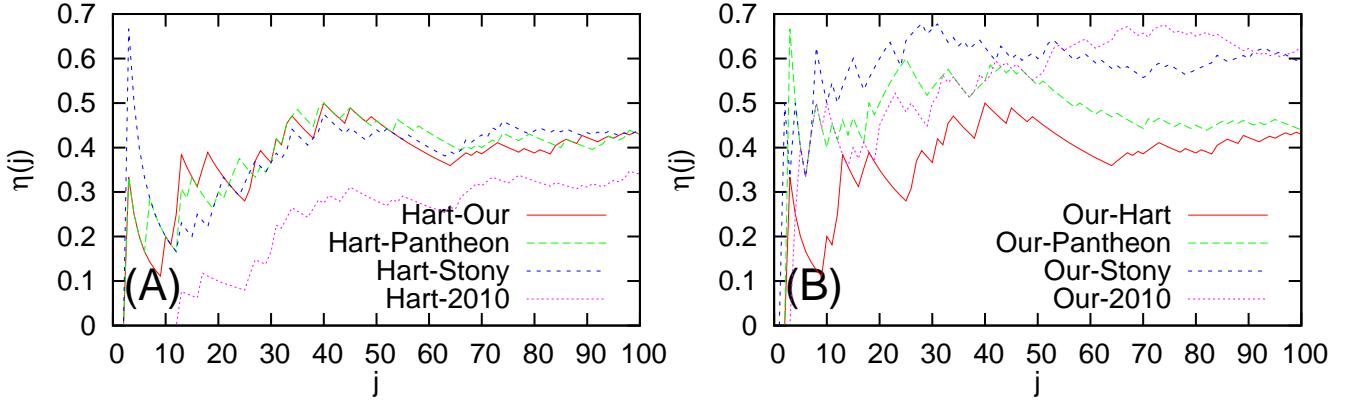


Figure S2. Dependence of fraction η of overlaped persons on rank index of person j . (A) Comparison is done of present study (“our”), PageRank list of English Wikipedia of [15] (“2010”), Stony-Brook list [19], Pantheon MIT project [20] in respect to Hart top 100 list. (B) Same as in (A) but comparison is done in respect to present study.

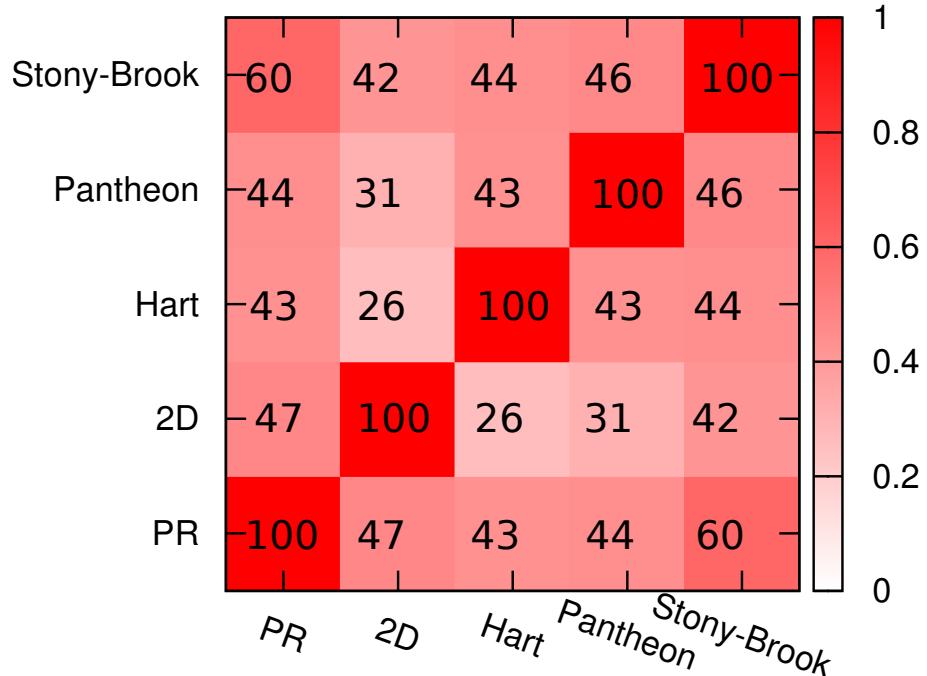


Figure S3. The overlap matrix (in percent) between 5 methods of ranking of top 100 historical figures from lists of Hart, Pantheon, Stony-Brook results and our global PageRank and 2DRank lists; percent or number of persons common for two lists is shown by color and numbers.

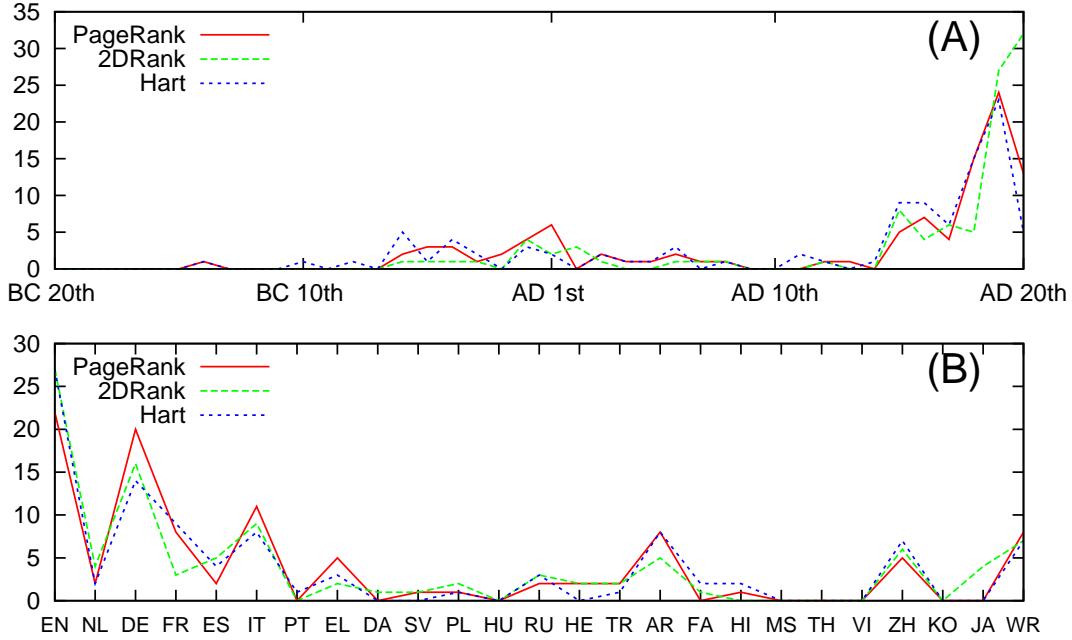


Figure S4. The number of top 100 historical figures, from the list of Hart and our global PageRank and 2DRank lists, are shown as a function of time (for a given century, panel A; one person from Hart’s list *Menes*, born in Egypt at BC 32nd and thus attributed to AR, is outside of time range in this panel but he is counted in panel B) and for a given language to which a person is attributed according to her/his birth place (panel B).

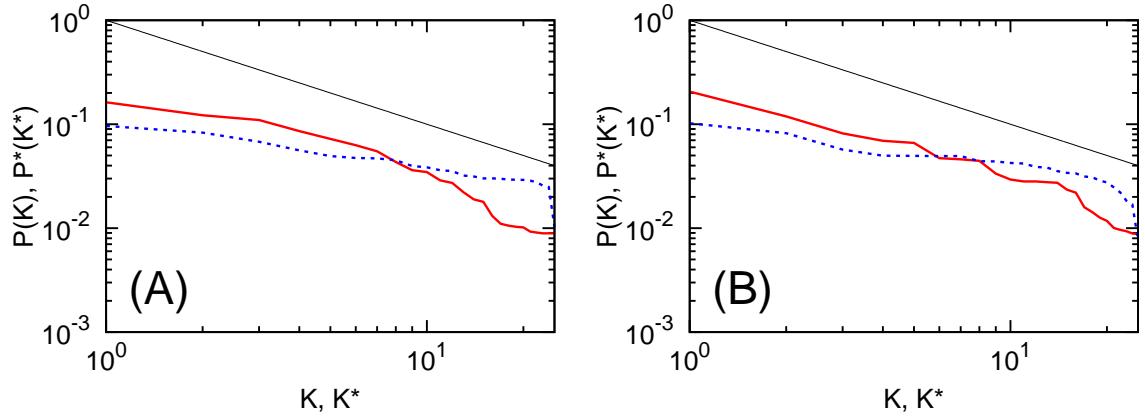


Figure S5. Dependence of probabilities of PageRank P (red) and CheiRank P^* (blue) on corresponding indexes K and K^* . The probabilities are obtained from the network shown in Fig.7 for corresponding panels (A), (B). The straight lines indicate the Zipf’s law $P \sim 1/K$; $P^* \sim 1/K^*$.

Table S1. List of local historical figures for EN category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	1861	William Shakespeare	1315	Michael Jackson
2	1789	Elizabeth II	991	Madonna (entertainer)
3	1756	Isaac Newton	773	Charles Darwin
4	1173	George W. Bush	754	Barack Obama
5	1101	Barack Obama	664	Elizabeth II
6	932	Charles Darwin	624	Bob Dylan
7	910	Franklin D. Roosevelt	556	William Shakespeare
8	656	George Washington	555	Elvis Presley
9	596	Bill Clinton	550	Queen Victoria
10	564	Winston Churchill	541	John Lennon

Table S2. List of local historical figures for NL category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	1476	Charlemagne	569	Charlemagne
2	556	Charles V, Holy Roman Emperor	297	Vincent van Gogh
3	83	Maurice Maeterlinck	294	Rembrandt
4	81	William I of the Netherlands	190	Charles V, Holy Roman Emperor
5	78	Beatrix of the Netherlands	138	Beatrix of the Netherlands
6	61	Baruch Spinoza	98	Baruch Spinoza
7	61	Rembrandt	94	Hugo Claus
8	51	Wilhelmina of the Netherlands	91	Johan Cruyff
9	47	Juliana of the Netherlands	76	Louis Couperus
10	39	Christiaan Huygens	75	Pierre Cuypers

Table S3. List of local historical figures for DE category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1)

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	2112	Adolf Hitler	1557	Adolf Hitler
2	1847	Albert Einstein	872	Ludwig van Beethoven
3	1730	Karl Marx	853	Wolfgang Amadeus Mozart
4	996	Wolfgang Amadeus Mozart	840	Pope Benedict XVI
5	925	Martin Luther	733	Johann Sebastian Bach
6	700	Ludwig van Beethoven	651	Albert Einstein
7	610	Johann Sebastian Bach	540	George Frideric Handel
8	570	Johann Wolfgang von Goethe	465	Karl Marx
9	528	Pope Benedict XVI	446	Michael Schumacher
10	417	Richard Wagner	344	Martin Luther

Table S4. List of local historical figures for FR category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	2208	Napoleon	720	Napoleon
2	1207	Louis XIV of France	268	Jules Verne
3	724	René Descartes	221	Louis XIV of France
4	397	Napoleon III	168	Giuseppe Garibaldi
5	385	Charles de Gaulle	146	Denis Diderot
6	260	Pierre André Latreille	144	François Mitterrand
7	167	Tacitus	127	Napoleon III
8	165	Jean-Baptiste Lamarck	121	Nicolas Sarkozy
9	157	Molière	113	Claudius
10	112	Francis I of France	112	Henry IV of France

Table S5. List of local historical figures for ES category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	276	Francisco Franco	285	Che Guevara
2	195	Philip II of Spain	216	Pablo Picasso
3	119	Pablo Picasso	206	Rafael Nadal
4	82	Lionel Messi	199	Jorge Luis Borges
5	74	Charles III of Spain	198	Salvador Dalí
6	72	Teresa of Ávila	178	Hadrian
7	71	Miguel de Cervantes	105	Shakira
8	70	Ferdinand VII of Spain	100	Francisco Goya
9	66	Alfonso X of Castile	95	Juan Perón
10	65	Ferdinand I, Holy Roman Emperor	94	Augusto Pinochet

Table S6. List of local historical figures for IT category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	1952	Julius Caesar	689	Julius Caesar
2	1662	Augustus	647	Augustus
3	1476	Christopher Columbus	616	Leonardo da Vinci
4	893	Galileo Galilei	464	Benito Mussolini
5	758	Benito Mussolini	339	Pope Pius XII
6	753	Cicero	330	Christopher Columbus
7	594	Leonardo da Vinci	326	Michelangelo
8	292	Pliny the Elder	322	Raphael
9	288	Pope Gregory XIII	197	Giuseppe Verdi
10	250	Pope Paul VI	172	Galileo Galilei

Table S7. List of local historical figures for PT category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	91	Getúlio Vargas	109	Ronaldo
2	83	Cristiano Ronaldo	100	Getúlio Vargas
3	74	John VI of Portugal	92	Juscelino Kubitschek
4	71	Luiz Inácio Lula da Silva	91	Rubens Barrichello
5	70	Pedro I of Brazil	90	Joaquim Maria Machado de Assis
6	67	Ferdinand Magellan	89	Fernando Henrique Cardoso
7	66	Maria I of Portugal	82	Luís de Camões
8	64	John I of Portugal	80	José Saramago
9	63	Pedro II of Brazil	79	John VI of Portugal
10	62	Juscelino Kubitschek	77	Oscar Niemeyer

Table S8. List of local historical figures for EL category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	2237	Aristotle	789	Alexander the Great
2	1949	Plato	207	Plato
3	1771	Alexander the Great	167	Aristotle
4	213	Socrates	108	Pericles
5	178	Plutarch	100	Mustafa Kemal Atatürk
6	153	Mustafa Kemal Atatürk	98	Eleftherios Venizelos
7	123	Sophocles	95	Andreas Papandreou
8	93	Aeschylus	94	Muhammad Ali of Egypt
9	86	Euripides	94	Ioannis Kapodistrias
10	84	Ioannis Kapodistrias	93	Plutarch

Table S9. List of local historical figures for DA category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	99	Tycho Brahe	210	Hans Christian Andersen
2	94	Ole Rømer	98	Margrethe II of Denmark
3	93	Christian IV of Denmark	95	N. F. S. Grundtvig
4	86	Margrethe II of Denmark	92	Sren Kierkegaard
5	85	Hans Christian Andersen	89	Christian IV of Denmark
6	84	Frederick IV of Denmark	88	Hans Christian Ørsted
7	80	Frederick II of Denmark	86	Anders Fogh Rasmussen
8	78	John Louis Emil Dreyer	84	Carl Nielsen
9	77	Christian VII of Denmark	83	Christian X of Denmark
10	76	Frederick III of Denmark	82	Niels Bohr

Table S10. List of local historical figures for SV category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	2284	Carl Linnaeus	326	Carl Linnaeus
2	125	August Strindberg	151	Ingmar Bergman
3	98	Alfred Nobel	146	Charles XII of Sweden
4	94	Gustav I of Sweden	116	Astrid Lindgren
5	93	Gustav III of Sweden	100	August Strindberg
6	86	Charles XII of Sweden	98	Carl XVI Gustaf of Sweden
7	82	Gustavus Adolphus of Sweden	92	Evert Taube
8	72	Carl XVI Gustaf of Sweden	89	Jan Myrdal
9	71	Charles XI of Sweden	88	Carl Jonas Love Almqvist
10	67	Charles IX of Sweden	83	Gustav I of Sweden

Table S11. List of local historical figures for PL category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	864	Pope John Paul II	693	Pope John Paul II
2	94	Catherine the Great	296	Frédéric Chopin
3	88	David Ben-Gurion	135	Catherine the Great
4	80	Casimir III the Great	98	David Ben-Gurion
5	72	Nathan Alterman	95	Bolesaw III Wrymouth
6	69	Lech Walesa	94	Andrzej Wajda
7	66	Lech Kaczyński	93	Nathan Alterman
8	63	Frédéric Chopin	91	Gerhart Hauptmann
9	60	Henryk Sienkiewicz	88	Anton Denikin
10	58	Sigismund I the Old	83	Lech Kaczyński

Table S12. List of local historical figures for HU category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	93	János Szentágothai	100	Stephen I of Hungary
2	91	Stephen I of Hungary	99	Sándor Petöfi
3	87	Lajos Kossuth	94	Kati Kovács
4	86	Miklós Réthelyi	93	Miklós Horthy
5	80	Béla IV of Hungary	92	Attila József
6	79	Louis I of Hungary	89	Sándor Weöres
7	75	Sándor Petöfi	86	Theodor Herzl
8	67	Miklós Horthy	83	Lajos Kossuth
9	56	Theodor Herzl	81	Miklós Radnóti
10	53	Andrew II of Hungary	77	János Kodolányi

Table S13. List of local historical figures for RU category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	1123	Immanuel Kant	334	Vladimir Putin
2	1022	Vladimir Lenin	274	Isaac Asimov
3	156	Peter the Great	198	Peter the Great
4	130	Mikhail Gorbachev	171	Vladimir Lenin
5	101	Pyotr Ilyich Tchaikovsky	127	Yuri Gagarin
6	97	Yuri Gagarin	109	Igor Stravinsky
7	97	Alexander Pushkin	100	Menachem Begin
8	91	Vladimir Putin	99	Dmitri Mendeleev
9	89	Nikita Khrushchev	96	Aleksander Griboyedov
10	88	Alexander II of Russia	95	Shimon Peres

Table S14. List of local historical figures for HE category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	1094	Mary (mother of Jesus)	580	Mary (mother of Jesus)
2	724	Saint Peter	240	Saint Peter
3	138	John the Baptist	171	John the Baptist
4	99	Yitzhak Rabin	99	Saint George
5	95	Yigal Amir	99	Yitzhak Rabin
6	84	Josephus	96	Ariel Sharon
7	81	Tom Segev	92	Benjamin Netanyahu
8	75	Ariel Sharon	85	Ehud Barak
9	65	Benjamin Netanyahu	82	Roni Dalumi
10	54	Herod the Great	79	Moshe Dayan

Table S15. List of local historical figures for TR category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	973	Paul the Apostle	252	Paul the Apostle
2	925	Herodotus	231	Suleiman the Magnificent
3	133	Strabo	172	Mehmed the Conqueror
4	117	Mehmed the Conqueror	169	Selim I
5	106	Suleiman the Magnificent	142	Abdul Hamid II
6	96	Abdul Hamid II	111	Julian (emperor)
7	93	Pausanias (geographer)	90	Recep Tayyip Erdoğan
8	83	İsmet İnönü	87	Adnan Menderes
9	79	Selim I	85	Lucian
10	79	Hesiod	84	Blent Ecevit

Table S16. List of local historical figures for AR category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	2282	Jesus	943	Jesus
2	1735	Muhammad	499	Muhammad
3	629	Moses	291	Ali
4	426	Saddam Hussein	219	Saddam Hussein
5	424	Ptolemy	181	Muammar Gaddafi
6	329	Augustine of Hippo	143	Hannibal
7	328	Ali	128	Saladin
8	196	Umar	128	Anwar Sadat
9	147	Anwar Sadat	117	Hosni Mubarak
10	134	Euclid	108	Yasser Arafat

Table S17. List of local historical figures for FA category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	110	Zoroaster	229	Cyrus the Great
2	101	Darius I	99	Zoroaster
3	100	Mahmoud Ahmadinejad	98	Mohammad Reza Pahlavi
4	97	Mohammad Reza Pahlavi	97	Mohammad Khatami
5	96	Rez Shh	96	Mir-Hossein Mousavi
6	94	Cyrus the Great	95	Ruhollah Khomeini
7	92	Ferdowsi	94	Naser al-Din Shah Qajar
8	90	Ruhollah Khomeini	93	Ali Khamenei
9	89	Naser al-Din Shah Qajar	92	Mohammad Mosaddegh
10	86	Mohammad Khatami	91	Ardashir I

Table S18. List of local historical figures for HI category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	168	Ashoka	126	Ashoka
2	106	Mahatma Gandhi	108	Akbar
3	100	Benazir Bhutto	99	Indira Gandhi
4	91	Vikramditya	98	Mahadevi Varma
5	90	Shivaji	96	Sanjeev Kumar
6	89	Jawaharlal Nehru	93	Amitabh Bachchan
7	88	Akbar	91	Premchand
8	87	Indira Gandhi	90	Dayananda Saraswati
9	86	Adi Shankara	89	Jaishankar Prasad
10	85	Vishnu Prabhakar	86	Adi Shankara

Table S19. List of local historical figures for MS category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	96	Mahathir Mohamad	100	Mahathir Mohamad
2	85	Najib Razak	99	Najib Razak
3	84	P. Ramlee	98	Anwar Ibrahim
4	81	Tunku Abdul Rahman	93	Mizan Zainal Abidin of Terengganu
5	79	Abdullah Ahmad Badawi	92	Sudirman Arshad
6	77	Muhyiddin Yassin	91	Tunku Abdul Rahman
7	74	Abdul Razak Hussein	90	Siti Nurhaliza
8	62	Anwar Ibrahim	89	Abdullah Ahmad Badawi
9	58	Hussein Onn	88	Abdul Taib Mahmud
10	37	Mizan Zainal Abidin of Terengganu	84	P. Ramlee

Table S20. List of local historical figures for TH category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	100	Chulalongkorn	100	Sirindhorn
2	97	Vajiravudh	98	Sirikit
3	96	Mongkut	97	Thaksin Shinawatra
4	94	Buddha Yodfa Chulaloke	94	Taksin
5	92	Nangklao	91	Pridi Banomyong
6	91	Thaksin Shinawatra	90	Yingluck Shinawatra
7	90	Damrong Rajanubhab	88	Srinagarindra
8	89	Taksin	86	Samak Sundaravej
9	88	Plaek Phibunsongkhram	82	Vajiralongkorn
10	87	Prajadhipok	80	Chao Keo Naovarat

Table S21. List of local historical figures for VI category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	91	Ho Chi Minh	98	Ho Chi Minh
2	71	Ngo Dinh Diem	97	Gia Long
3	62	Minh Mng	96	Minh Mng
4	46	Gia Long	94	Nguyen Hue
5	44	Bo i	86	Le Loi
6	22	Le Loi	84	Tran Hung Dao
7	15	Nhat Linh	83	Vo Nguyen Giap
8	N/A	N/A	82	Tu Duc
9	N/A	N/A	81	Le Thánh Tông
10	N/A	N/A	80	Trung Sisters

Table S22. List of local historical figures for ZH category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	375	Mao Zedong	306	Cao Cao
2	285	Confucius	243	14th Dalai Lama
3	244	Chiang Kai-shek	234	Liu Bei
4	197	Qin Shi Huang	192	Qin Shi Huang
5	186	Emperor Wu of Han	191	Kangxi Emperor
6	135	Cao Cao	188	Zhuge Liang
7	129	Hongwu Emperor	179	Qianlong Emperor
8	119	Qianlong Emperor	154	Mao Zedong
9	119	Kangxi Emperor	147	Hongwu Emperor
10	94	Sun Yat-sen	146	Sun Yat-sen

Table S23. List of local historical figures for KO category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	100	Gojong of the Korean Empire	114	Gojong of the Korean Empire
2	98	Kim Il-sung	106	Kim Il-sung
3	95	Sejong the Great	100	Park Chung-hee
4	94	Park Chung-hee	99	Kim Dae-jung
5	93	Taejong of Joseon	97	Roh Moo-hyun
6	92	Syngman Rhee	95	Sejong the Great
7	91	Yeongjo of Joseon	94	Taejo of Goryeo
8	90	Kim Dae-jung	93	Kim Young-sam
9	89	Seonjo of Joseon	92	Jeongjo of Joseon
10	86	Taejo of Joseon	90	Syngman Rhee

Table S24. List of local historical figures for JA category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	154	Toyotomi Hideyoshi	346	Tokugawa Ieyasu
2	153	Tokugawa Ieyasu	266	Toyotomi Hideyoshi
3	108	Hirohito	252	Hirohito
4	97	Oda Nobunaga	233	Oda Nobunaga
5	86	Emperor Meiji	140	Junichiro Koizumi
6	81	Minamoto no Yoritomo	131	Shinzō Abe
7	76	Junichiro Koizumi	112	Tsunku
8	73	Emperor Tenmu	106	Emperor Meiji
9	70	Natsume Sōseki	100	Koxinga
10	69	Akihito	97	Osamu Tezuka

Table S25. List of local historical figures for WR category. Here Θ_A is the ranking score of the algorithm A defined in Eq.(1).

	Θ_A	PageRank local figures	Θ_A	2DRank local figures
1	1686	Joseph Stalin	529	J. R. R. Tolkien
2	842	Constantine the Great	477	Joseph Stalin
3	564	Gautama Buddha	276	Nelson Mandela
4	506	Sigmund Freud	241	Genghis Khan
5	405	Diocletian	195	Sigmund Freud
6	351	Justinian I	191	Martina Navratilova
7	219	J. R. R. Tolkien	186	Constantine the Great
8	203	Genghis Khan	173	Justinian I
9	138	Avicenna	127	Nikola Tesla
10	129	Rumi	123	Kublai Khan

A Weighted Correlation Index for Rankings with Ties

Sebastiano Vigna*
 Università degli Studi di Milano, Italy

April 28, 2014

Abstract

Understanding the correlation between two different scores for the same set of items is a common problem in information retrieval, and the most commonly used statistics that quantifies this correlation is Kendall’s τ . However, the standard definition fails to capture that discordances between items with high rank are more important than those between items with low rank. Recently, a new measure of correlation based on *average precision* has been proposed to solve this problem, but like many alternative proposals in the literature it assumes that there are *no ties* in the scores. This is a major deficiency in a number of contexts, and in particular while comparing centrality scores on large graphs, as the obvious baseline, indegree, has a very large number of ties in web and social graphs. We propose to extend Kendall’s definition in a natural way to take into account weights in the presence of ties. We prove a number of interesting mathematical properties of our generalization and describe an $O(n \log n)$ algorithm for its computation. We also validate the usefulness of our weighted measure of correlation using experimental data.

1 Introduction

In information retrieval, one is often faced with different scores¹ for the same set of items. This includes the lists of documents returned by different search engines and their associated relevance scores, the lists of query recommendation returned by different algorithms, and also the score associated to each node of a graph by different centrality measures (e.g., indegree and Bavelas’s closeness [1]).

In most of the literature, the scores are assumed to be without ties, thus inducing a *ranking* of the elements. At that point, correlation statistics such as Spearman’s rank correlation coefficient [24] and Kendall’s τ [12] can be used to evaluate the similarity of the rankings. Spearman’s correlation coefficient is equivalent to the traditional linear correlation coefficient computed on ranks of items. Kendall’s τ , instead, is proportional to the number of pairwise adjacent swaps needed to convert one ranking into the other.

For a number of reasons, Kendall’s τ has become a standard statistic to compare the correlation between two ranked lists. Such reasons include fast computation ($O(n \log n)$, where n is the length of the list, using Knight’s algorithm [14]), and the existence of a variant that takes care of ties [13].

The explicit treatment of ties is of great importance when comparing global *exogenous* relevance scores in large collections of web documents. The baseline of such scores is indegree—the number of documents containing hypertextual link to a given document. More sophisticated approaches include Katz’s index [10], PageRank [21], and countless variants. Due to the highly skewed indegree distribution, a very large number of documents share the same indegree, and the same happens of many other scores: it is thus of uttermost importance that the evaluation of correlation takes into account ties as first-class citizens.

On the other hand, Kendall’s τ has some known problems that motivated the introduction of several weighted variants. In particular, a striking difference often emerges between the anecdotal evidence of the top elements by different scores being almost identical, and the τ value being quite low. This is due to a known phenomenon: the scores of important items tend to be highly correlated in all reasonable rankings, whereas most of the remaining items are ranked in slightly different ways, introducing a large amount of noise, yielding a low τ value.

*Sebastiano Vigna has been supported by the EU-FET grant NADINE (GA 288956).

¹We purposely and consistently use “score” to denote real numbers associated to items, and “rank” to denote ordinal positions. The two terms are used somewhat interchangeably in the literature, but in this paper the distinction is important as we assume that scores of different items can be identical.

This problem motivates the definition of correlation statistics that consider more important correlation between highly ranked items. In particular, recently Yilmaz, Aslam and Robertson introduced a statistics, named *AP (average precision) correlation* [27], which aims at considering more important swaps between highly ranked items. The need for such a measure is very well motivated in the introduction of their paper, and we will not repeat here their detailed discussion.

In this paper, we aim at providing a measure of correlation in the same spirit of the definition of Yilmaz, Aslam and Robertson, but taking smoothly ties into account. We will actually define a general notion of weighting for Kendall's τ , and develop its mathematical properties. Since it is important that such a statistics is computable on very large data sets, we will provide a generalization of Knight's algorithm that can be applied whenever the weighting depends additively or multiplicatively on a weight assigned to each item. The same algorithm can be used to compute AP correlation in time $O(n \log n)$.

All data and software used in this paper are available as part of the LAW software library under the GNU General Public License.²

2 Related work

Shieh [23] wrote the one of the first papers proposing a generic weighting of Kendall's τ . She assumes from the very start that there are no ties, and assign to the exchange between i and j a weight w_{ij} . Her motivation is the *fidelity evaluation of software packages for structural engineering*, in which a set of variables is ranked in two different ways, and one would like to emphasize agreement on the most important ones. In particular, she concentrates on weights given by the product of two weights associated with the elements participating in the exchange. Our work can be seen as a generalization of her approach, albeit we combine weights differently.

Kumar and Vassilvitskii [16] study a definition that extends Shieh's taking into account *position weights* and *similarity between elements*. Again, they assume that ties are broken arbitrarily, which is an unacceptable assumption if large sets of elements have the same score. Fagin, Kumar and Sivakumar [6] use instead *penalty weights* to apply Kendall's τ just to the top k elements of two ranked lists (with no ties). Exchanges partially or completely outside the top k elements obtain different weights.

Finally, the recent quoted work of Yilmaz, Aslam and Robertson [27] on AP correlation is the closest to ours in motivation and methodology, albeit targeted at ranked lists with no ties.

We remark that analogous research exists in association with Spearman's correlation: Iman and Conover [9], for example, study the usage of *Savage scores* [22] instead of ranks when comparing ranked lists. Savage scores for a ranked list of n elements are given by $\sum_{j=i}^n 1/j$, where i is the rank (starting at one) of an element. Spearman's correlation applied to Savage scores considers more important elements at the top of a ranked list.

Recently, Webber, Moffat and Zobel [26] have described a similarity measure for *indefinite rankings*—rankings that might have different lengths and contain different elements. Their work has some superficial resemblance with the approach of [16, 27] and our work, as it give preminence to differences at the top of ranked lists, but it is not technically a correlation index, as it is based on measuring overlaps of infinite lists, rather than on exchanges. Thus, the basic condition for a correlation index (i.e., that inverting the list one obtains the minimum possible correlation, usually standardized to -1), is not even expressible in their framework. Moreover, their measure, being defined on infinite lists, needs the fundamental assumption that the weight function applied to overlaps must be *summable*; in particular, they make importance decrease exponentially. As we will discuss in Section 4.2, and verify experimentally in Section 6, such a choice is a reasonable framework for very short lists, or when only very first elements are relevant (e.g., because one is modelling user behavior), but it would completely flatten the results of our correlation index on large examples, depriving it from its discriminatory power, even if the weight function would decrease just quadratically.

A fascinating proposal, entirely orthogonal to the ones we discussed, is the idea of weighting Kemeny's distance between permutations proposed by Farnoud and Milenkovic [7]. In this proposal, Kemeny's distance between two permutations π and σ is characterized as the minimum number of *adjacent transpositions* (i.e., transpositions of the form $(i \ i+1)$) that turn π into σ . At this point, one can define a *weight* associated to each adjacent transposition, and by assigning larger weights to adjacent transpositions with smaller indices one can make differences in the top part of the permutations more important than differences in the bottom part. The right notion of weighted distance turns out to be the minimum sum of weights of a sequence of adjacent transposition that turn π into σ . The interesting property of this approach is that avoids the need for a *ground truth* (an intrinsic notion of importance of an element), which is necessary,

²<http://law.di.unimi.it/>

implicitly or explicitly, to weigh an exchange in the approaches of [23, 27] and in the one discussed in this paper. The main drawback, presently, is that even in the presence of weighting functions that are monotonically decreasing in i the time necessary to compute the distance is $O(n^2)$ instead of $O(n \log n)$. It is also necessary more tuning to extend the distance to the case of ties, and to turn in this case the distance into a proper correlation index with range in $[-1 \dots 1]$.

3 Motivation

The need for weighted correlation measures in the case of ranked list has been articulated in detail in previous work. Here we will focus on the case of centrality measures for graphs. Consider the graph of English Wikipedia³, which has about four million nodes and one hundred million arcs. In this graph, 99.95% of the nodes have the same indegree of some other node—for example, more than a half million node has indegree one. It is clearly mandatory, when computing the correlation of other scores with indegree, that ties are taken into consideration in a systematic way (e.g., not broken arbitrarily).

We will consider four other commonly used scores based on the adjacency matrix A of the Wikipedia graph. One is PageRank [21], which is defined by

$$\mathbf{1}/n \sum_{k \geq 0} (\alpha \bar{A})^k,$$

where $\alpha \in [0 \dots 1)$ is a *damping factor* and \bar{A} is a stochasticization of A : every row not entirely made of zeroes is divided by its sum, so to have ℓ_1 norm one.

The other index we consider is Katz's [10], which is defined by

$$\mathbf{1} \sum_{k \geq 0} (\alpha A)^k,$$

where $\alpha \in [0 \dots 1/\lambda)$ is an attenuation factor depending on λ , the dominant eigenvalue of A [19]. In both cases, we take α in the middle of the allowed interval (using different values does not change the essence of what follows, unless they are extreme).

A different kind of score is provided by Bavelas's *closeness*. The closeness of x is defined by

$$\frac{1}{\sum_{d(y,x) < \infty} d(y, x)},$$

where $d(-, -)$ denotes the usual graph distance. Note that we have to eliminate nodes at infinite distance to avoid zeroing all scores. By definition the closeness of a node with indegree zero is zero. Finally, we consider *harmonic centrality* [2], a modified version of Bavelas's closeness designed for directed graphs that are not strongly connected; the harmonic centrality of x is defined by

$$\sum_{y \neq x} \frac{1}{d(y, x)}.$$

These scores provide an interesting mix: indegree is an obvious baseline, and entirely local. PageRank and Katz are similar in their definition, but the normalization applied to A makes the scores quite different (at least in theory). Finally, closeness and harmonic centrality are of a completely different nature, having no connection with dominant eigenvectors or Markov chains.

Our first empirical observation is that, looking just at the very top pages of Wikipedia (Table 1; entries in boldface are unique to the list they belong to, here and in the following), we perceive these scores as almost identical, except for closeness, which displays almost random values. The latter behavior is a known phenomenon: nodes that are almost isolated obtain a very high closeness score (this is why harmonic centrality was devised). We note also that harmonic centrality has a slightly different slant, as it is the only ranking including Latin, Europe, Russia and the Catholic Church in the top 20.

The problem is that these facts are not reflected in any way in the values of Kendall's τ shown in Table 3. If we exclude closeness, with the exception of the correlation between indegree and Katz, all other correlation value fail to surpass the 0.9 threshold, usually considered the threshold for considering two rankings equivalent [25]. Actually, they

³More precisely, a specific snapshot of Wikipedia that will be made public by the author. The graph does not contain template pages.

Indegree	PageRank	Katz	Harmonic	Closeness
United States	United States	United States	United States	Kharqan Rural District
List of sovereign states	Animal	List of sovereign states	United Kingdom	Talageh-ye Sofla
Animal	List of sovereign states	United Kingdom	World War II	Talageh-ye Olya
England	France	France	France	Greatest Remix Hits (Whigfield album)
France	Germany	Animal	Germany	Suzhou HSR New Town
Association football	Association football	World War II	Association football	Suzhou Lakeside New City
United Kingdom	England	England	English language	Mepirodipine
Germany	India	Association football	China	List of MPs ... M–N
Canada	United Kingdom	Germany	Canada	List of MPs ... O–R
World War II	Canada	Canada	India	List of MPs ... S–T
India	Arthropod	India	Latin	List of MPs ... U–Z
Australia	Insect	Australia	World War I	List of MPs ... J–L
London	World War II	London	England	List of MPs ... C
Japan	Japan	Italy	Italy	List of MPs ... F–I
Italy	Australia	Japan	Russia	List of MPs ... A–B
Arthropod	Village	New York City	Europe	List of MPs ... D–E
Insect	Italy	English language	Australia	Esmaili-ye Sofla
New York City	Poland	China	European Union	Esmaili-ye Olya
English language	English language	Poland	Catholic Church	Levels of organization (ecology)
Village	National Reg. of Hist. Places	World War I	London	Jacques Moeschal (architect)

Table 1: Top 20 pages of the English version of Wikipedia following five different centrality measures.

are below the threshold 0.8, under which we are supposed to see considerable changes. The correlation of closeness with harmonic centrality, moreover, is even more pathological: it is the *largest* correlation.

An obvious observation is that, maybe, the score is lowered by a large discordance in the rest of the rankings. Table 2 tries to verify this intuition by listing the top pages that are associated with the Wordnet category “scientist” in the Yago2 ontology data [8]. These pages have considerably lower score (their rank is below 300), yet the first three rankings are almost identical. Harmonic centrality is still slightly different (Linnaeus is absent, and actually ranks 21), which tells us that the Kendall’s τ is not giving completely unreasonable data. Nonetheless, closeness continues to provide apparently random results.

We have actually to delve deep into Wikipedia, beyond rank 100 000 using the category “cocktail” to see that, finally, things settle down (Table 5). While closeness still displays a few quirks, the rankings start to stabilize.

To understand what happens in the very low-rank region, in Table 4 we provide Kendall’s τ as in Table 3, but *restricting the computation to nodes of indegree 1 and 2*. As it is immediately evident, after stabilization the low-rank region is fraught with noise and all correlation values drop significantly.

The very high correlation between closeness and harmonic centrality is, actually, not strange: on the nodes reachable from giant connected component of our Wikipedia snapshot (89% of the nodes) they agree almost exactly, as closeness is the reciprocal of a denormalized *arithmetic* mean, whereas harmonic centrality is the reciprocal of a denormalized *harmonic* mean [2]. Even if the remaining 11% of the nodes is completely out of place, making closeness useless, Kendall’s τ tells us that it should be interchangeable with harmonic centrality. At the same time, Kendall’s τ tells us that indegree is very different from PageRank, which again goes completely against our empirical evidence.

In the rest of the paper, we will try to approach in a systematic manner these problems by defining a new weighted correlation index for scores with ties.

4 Definitions and Tools

In his 1945 paper about ranking with ties [13], Kendall, starting from an observation of Daniels [4], reformulates his correlation index using a definition similar in spirit to that of an inner product, which will be the starting point of our proposal: we consider two real-valued vectors r and s (to be thought as scores) with indices in $[n]$; then, let us define

$$\langle r, s \rangle := \sum_{i < j} \operatorname{sgn}(r_i - r_j) \operatorname{sgn}(s_i - s_j),$$

Indegree	PageRank	Katz	Harmonic	Closeness
Carl Linnaeus	Carl Linnaeus	Carl Linnaeus	Aristotle	Noël Bernard (botanist)
Aristotle	Aristotle	Aristotle	Albert Einstein	Charles Coquelin
Thomas Jefferson	Thomas Jefferson	Thomas Jefferson	Thomas Jefferson	Markku Kivinen
Margaret Thatcher	Charles Darwin	Albert Einstein	Charles Darwin	Angiolo Maria Colomboni
Plato	Plato	Charles Darwin	Thomas Edison	Om Prakash (historian)
Charles Darwin	Albert Einstein	Karl Marx	Alexander Graham Bell	Michel Mandjes
Karl Marx	Karl Marx	Plato	Nikola Tesla	Kees Posthumus
Albert Einstein	Pliny the Elder	Margaret Thatcher	William James	F. Wolfgang Schnell
Vladimir Lenin	Vladimir Lenin	Vladimir Lenin	Isaac Newton	Christof Ebert
Sigmund Freud	Johann Wolfgang von Goethe	Isaac Newton	Karl Marx	Reese Prosser
J. R. R. Tolkien	Margaret Thatcher	Ptolemy	Charles Sanders Peirce	David Tulloch
Johann Wolfgang von Goethe	Ptolemy	Johann Wolfgang von Goethe	Noam Chomsky	Kim Hawtrey
Spider-Man	Sigmund Freud	Pliny the Elder	Enrico Fermi	Patrick J. Miller
Pliny the Elder	Isaac Newton	Benjamin Franklin	Ptolemy	Mikel King
Benjamin Franklin	Benjamin Franklin	J. R. R. Tolkien	John Dewey	Albert Perry Brigham
Leonardo da Vinci	J. R. R. Tolkien	Thomas Edison	Johann Wolfgang von Goethe	Gordon Wagner (economist)
Isaac Newton	Immanuel Kant	Sigmund Freud	Bertrand Russell	George Henry Chase
Ptolemy	Leonardo da Vinci	Immanuel Kant	Plato	Charles C. Horn
Immanuel Kant	Pierre André Latreille	Leonardo da Vinci	John von Neumann	Paul Goldstene
George Bernard Shaw	Thomas Edison	Noam Chomsky	Vladimir Lenin	Robert Stanton Avery

Table 2: Top 20 pages of Wikipedia following five different centrality measures and restricting pages to Yago2 Wordnet category “scientist”. The global rank of these items is beyond 300.

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.75	0.90	0.62	0.55
PageRank	0.75	1	0.75	0.61	0.56
Katz	0.90	0.75	1	0.70	0.62
Harmonic	0.62	0.61	0.70	1	0.92
Closeness	0.55	0.56	0.62	0.92	1

Table 3: Kendall’s τ between Wikipedia centrality measures.

where

$$\text{sgn}(x) := \begin{cases} 1 & \text{if } x > 0; \\ 0 & \text{if } x = 0; \\ -1 & \text{if } x < 0. \end{cases}$$

Indices of score vectors in summations belong to $[n]$ throughout the paper. Note that

$$\langle \mathbf{r}, \alpha \mathbf{s} \rangle = \langle \alpha \mathbf{r}, \mathbf{s} \rangle = \text{sgn}(\alpha) \langle \mathbf{r}, \mathbf{s} \rangle,$$

which reminds of the analogous property for inner products, and that $\langle \mathbf{r}, - \rangle = \langle - , \mathbf{r} \rangle = 0$ if \mathbf{r} is constant. Following the analogy, we can define

$$\|\mathbf{r}\| := \sqrt{\langle \mathbf{r}, \mathbf{r} \rangle},$$

so

$$\|\alpha \mathbf{r}\| = |\text{sgn}(\alpha)| \cdot \|\mathbf{r}\|.$$

The norm thus defined measures the “untierness” of \mathbf{r} : it is zero if and only if \mathbf{r} is a constant vector, and it has maximum value $\sqrt{n(n-1)/2}$ when all components of \mathbf{r} are distinct.

We can now define Kendall’s τ between two vectors \mathbf{r} and \mathbf{s} with nonnull norm as a normalized inner product, in a way formally identical to cosine similarity:

$$\tau(\mathbf{r}, \mathbf{s}) := \frac{\langle \mathbf{r}, \mathbf{s} \rangle}{\|\mathbf{r}\| \cdot \|\mathbf{s}\|}. \quad (1)$$

We recall that if \mathbf{r} and \mathbf{s} have no ties, the definition reduces to the classical “normalized difference of concordances and discordances”, as the denominator is exactly $n(n-1)/2$. The definition above is exactly that proposed by Kendall [13], albeit we use a different formalism.

The form of (1) suggests that to obtain a weighted correlation index it would be natural to define a *weighted* inner product

$$\langle \mathbf{r}, \mathbf{s} \rangle_w := \sum_{i < j} \text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j) w(i, j),$$

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.31	0.63	0.24	0.06
PageRank	0.31	1	0.27	0.10	0.10
Katz	0.63	0.27	1	0.50	0.20
Harmonic	0.24	0.10	0.50	1	0.65
Closeness	0.06	0.10	0.20	0.65	1

Table 4: Kendall's τ between Wikipedia centrality measures, restricted to nodes of indegree 1 and 2.

where $w(-, -) : [n] \times [n] \rightarrow \mathbf{R}_{\geq 0}$ is some nonnegative weight function. We would have then a new norm $\|\mathbf{r}\|_w = \sqrt{\langle \mathbf{r}, \mathbf{r} \rangle_w}$ and a new correlation index

$$\tau_w(\mathbf{r}, \mathbf{s}) := \frac{\langle \mathbf{r}, \mathbf{s} \rangle_w}{\|\mathbf{r}\|_w \cdot \|\mathbf{s}\|_w}.$$

Note that still $\langle \mathbf{r}, - \rangle_w = \langle -, \mathbf{r} \rangle_w = 0$ if \mathbf{r} is constant.

We say that two score vectors \mathbf{r} and \mathbf{s} are *equivalent* if $\text{sgn}(r_i - r_j) = \text{sgn}(s_i - s_j)$, *opposite* if $\text{sgn}(r_i - r_j) = -\text{sgn}(s_i - s_j)$ for all i and j .

Lemma 1 *We have*

$$\sum_{i < j} |\text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j)| w(i, j) \leq \|\mathbf{r}\|_w \|\mathbf{s}\|_w. \quad (2)$$

A sufficient condition for equality to hold is that the two vectors are equivalent or opposite.

Proof. Let $R_{ij} = |\text{sgn}(r_i - r_j)|$ and $S_{ij} = |\text{sgn}(s_i - s_j)|$. Then,

$$\begin{aligned} & \left(\sum_{i < j} R_{ij} S_{ij} w(i, j) \right)^2 \\ &= \left(\sum_{i < j} R_{ij}^2 S_{ij}^2 w(i, j)^2 \right) + \left(\sum_{\substack{i < j, k < \ell \\ i \neq k \vee j \neq \ell}} R_{ij} S_{ij} R_{k\ell} S_{k\ell} w(i, j) w(k, \ell) \right) \\ &\leq \left(\sum_{i < j} R_{ij}^2 S_{ij}^2 w(i, j)^2 \right) + \left(\sum_{\substack{i < j, k < \ell \\ i \neq k \vee j \neq \ell}} R_{ij}^2 S_{k\ell}^2 w(i, j) w(k, \ell) \right) \\ &= \left(\sum_{i < j} R_{ij}^2 w(i, j) \right) \left(\sum_{i < j} S_{ij}^2 w(i, j) \right) = \|\mathbf{r}\|_w^2 \|\mathbf{s}\|_w^2. \end{aligned}$$

Note that if the vectors are equivalent or opposite then

$$R_{ij} S_{ij} R_{k\ell} S_{k\ell} = R_{ij}^2 S_{k\ell}^2$$

for all i, j, k and ℓ , so we obtain equality. ■

We now prove a fundamental Cauchy–Schwartz-like inequality:

Theorem 1 $|\langle \mathbf{r}, \mathbf{s} \rangle_w| \leq \|\mathbf{r}\|_w \|\mathbf{s}\|_w$. *A sufficient condition for equality to hold is that the two vectors are equivalent or opposite. The condition is necessary if w is strictly positive and $|\langle \mathbf{r}, \mathbf{s} \rangle_w| \neq 0$.*

Proof. The first two statements are immediate from Lemma 1, as

$$|\langle \mathbf{r}, \mathbf{s} \rangle_w| \leq \sum_{i < j} |\text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j)| w(i, j)$$

and in the case of equivalent or opposite vectors we have equality. On the other hand, if we let $R_{ij} = \text{sgn}(r_i - r_j)$ and $S_{ij} = \text{sgn}(s_i - s_j)$ the chain of equalities and inequalities at the beginning of the proof of Lemma 1 continues to be true. To have equality, however, assuming that w is strictly positive we must have

$$R_{ij} S_{ij} R_{k\ell} S_{k\ell} w(i, j) w(k, \ell) = R_{ij}^2 S_{k\ell}^2 w(i, j) w(k, \ell)$$

for all i, j, k and ℓ , that is,

$$R_{ij}S_{ij}R_{k\ell}S_{k\ell} = R_{ij}^2S_{k\ell}^2.$$

Now, since $|\langle \mathbf{r}, \mathbf{s} \rangle_w| \neq 0$ there must be a pair \bar{i}, \bar{j} such that $R_{\bar{i}\bar{j}} \neq 0$ and $S_{\bar{i}\bar{j}} \neq 0$. Letting $\sigma = R_{\bar{i}\bar{j}}S_{\bar{i}\bar{j}}$ we have

$$R_{k\ell}S_{k\ell} = \sigma S_{k\ell}^2$$

and

$$R_{ij}S_{ij} = \sigma R_{ij}^2$$

for all i, j, k and ℓ . In particular, if $R_{k\ell} = 0$ we have necessarily $S_{k\ell} = 0$, and *vice versa*. If $R_{k\ell} \neq 0$, then $S_{k\ell} = \sigma R_{k\ell}$, which completes the proof. ■

Another application of Lemma 1 gives the triangular inequality:

Theorem 2 $\|\mathbf{r} + \mathbf{s}\|_w \leq \|\mathbf{r}\|_w + \|\mathbf{s}\|_w$.

Proof.

$$\begin{aligned} \|\mathbf{r} + \mathbf{s}\|_w^2 &= \langle \mathbf{r} + \mathbf{s}, \mathbf{r} + \mathbf{s} \rangle_w \\ &= \sum_{i < j} \operatorname{sgn}(r_i + s_i - r_j - s_j)^2 w(i, j) \\ &= \sum_{i < j} |\operatorname{sgn}(r_i + s_i - r_j - s_j)|^2 w(i, j) \\ &\leq \sum_{i < j} (|\operatorname{sgn}(r_i - r_j)| + |\operatorname{sgn}(s_i - s_j)|)^2 w(i, j) \\ &= \langle \mathbf{r}, \mathbf{r} \rangle_w + \langle \mathbf{s}, \mathbf{s} \rangle_w + 2 \sum_{i < j} |\operatorname{sgn}(r_i - r_j) \operatorname{sgn}(s_i - s_j)| w(i, j) \\ &\leq \|\mathbf{r}\|_w^2 + \|\mathbf{s}\|_w^2 + 2\|\mathbf{r}\|_w\|\mathbf{s}\|_w \\ &= (\|\mathbf{r}\|_w + \|\mathbf{s}\|_w)^2. \end{aligned}$$

■

The triangular inequality has a nice combinatorial interpretation: adding score vectors can only *decrease* the amount of “untieness”. There is no way to induce in a sum vector more untieness than the amount present in the summands.

Finally, an easy application of Theorem 1 shows that τ_w is sensible and works as expected:

Theorem 3 Let $w : [n] \times [n] \rightarrow \mathbf{R}$ be a nonnegative weight function. The following properties hold for every score vector \mathbf{t} and for every \mathbf{r}, \mathbf{s} with nonnull norm:

- if \mathbf{t} is constant, $\|\mathbf{t}\|_w = 0$;
- $-1 \leq \tau_w(\mathbf{r}, \mathbf{s}) \leq 1$;
- if \mathbf{r} and \mathbf{s} are equivalent, $\tau_w(\mathbf{r}, \mathbf{s}) = 1$;
- if \mathbf{r} and \mathbf{s} are opposite, $\tau_w(\mathbf{r}, \mathbf{s}) = -1$;

Moreover, if w is strictly positive:

- if $\|\mathbf{t}\|_w = 0$, \mathbf{t} is constant;
- if $\tau_w(\mathbf{r}, \mathbf{s}) = 1$, \mathbf{r} and \mathbf{s} are equivalent;
- if $\tau_w(\mathbf{r}, \mathbf{s}) = -1$, \mathbf{r} and \mathbf{s} are opposite.

As a result, if w is strictly positive and we obtain correlation ± 1 the equivalence classes formed by tied scores are necessarily in a size-preserving bijection that is monotone decreasing on the scores.

Indegree	PageRank	Katz	Harmonic	Closeness
Martini (cocktail)	Martini (cocktail)	Irish coffee	Irish coffee	Magie Noir
Piña colada	Caipirinha	Caipirinha	Caipirinha	Batini (drink)
Mojito	Mojito	Martini (cocktail)	Kir (cocktail)	Scorpion bowl
Caipirinha	Piña colada	Piña colada	Martini (cocktail)	Poinsettia (cocktail)
Cuba Libre	Irish coffee	Kir (cocktail)	Piña colada	Irish coffee
Irish coffee	Kir (cocktail)	Mojito	Mojito	Caipirinha
Singapore Sling	Cosmopolitan (cocktail)	Mai Tai	Beer cocktail	Kir (cocktail)
Manhattan (cocktail)	Manhattan (cocktail)	Cuba Libre	Shaken, not stirred	Martini (cocktail)
Windle (sidecar)	IBA Official Cocktail	Singapore Sling	Pisco Sour	Piña colada
Cosmopolitan (cocktail)	Beer cocktail	Long Island Iced Tea	Mai Tai	Mojito
Mai Tai	Mai Tai	Shaken, not stirred	Spritz (alcoholic beverage)	Beer cocktail
IBA Official Cocktail	Singapore Sling	Beer cocktail	Long Island Iced Tea	Shaken, not stirred
Kir (cocktail)	Cuba Libre	Manhattan (cocktail)	Sazerac	Mai Tai
Shaken, not stirred	Tom Collins	Cosmopolitan (cocktail)	Fizz (cocktail)	Spritz (alcoholic beverage)
Beer cocktail	Long Island Iced Tea	Windle (sidecar)	Flaming beverage	Pisco Sour
Pisco Sour	Sour (cocktail)	Pisco Sour	Cuba Libre	Long Island Iced Tea
Long Island Iced Tea	Shaken, not stirred	White Russian (cocktail)	Wine cocktail	Sazerac
Sour (cocktail)	Negroni	IBA Official Cocktail	Singapore Sling	Flaming beverage
White Russian (cocktail)	Flaming beverage	Moscow mule	Moscow mule	Fizz (cocktail)
Vesper (cocktail)	Lillet	Vesper (cocktail)	White Russian (cocktail)	Wine cocktail

Table 5: Top 20 pages of Wikipedia following five different centrality measures and restricting pages to Yago2 Wordnet category “cocktail”. The global rank of these items is beyond 100 000.

4.1 Decoupling rank and weight

The reader has probably already noticed that the dependence on the weight on the *indices* associated to the elements has no meaning: a trivial request (see, for instance [11]) on a correlation measure is that, like Kendall’s τ , it is *invariant by isomorphism*, that is, it does not change if we permute the indices of the vector. This currently doesn’t happen because we are using the numbering of the element as *ground truth* to weigh the correlation between \mathbf{r} and \mathbf{s} . While there is nothing bad in principle (we can stipulate that elements are indexed in order of importance using some external source of information), we think that a more flexible approach decouples the problem of the ground truth from the problem of weighting. We thus define the *ranked-weight* product

$$\langle \mathbf{r}, \mathbf{s} \rangle_{\rho, w} := \sum_{i < j} \text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j) w(\rho(i), \rho(j)),$$

where $\rho : [n] \rightarrow [n] \cup \{ \infty \}$ is a ranking function associating with each index a *rank*, the highest rank being zero. We admit the possibility of rank ∞ , given that the weight function provides a meaningful value in such a case, to include also the case of *partial ground truths*. The definition of the ranked-weighted product induces, as in (1), a correlation index $\tau_{\rho, w}$, and the machinery we developed applies immediately, as $w(\rho(-), \rho(-))$ is just a different weight function.

What if there is no ground truth to rely on? Our best bet is to use the rankings induced by the vectors \mathbf{r} and \mathbf{s} . Let us denote by $\rho_{\mathbf{r}, \mathbf{s}}$ the ranking defined by ordering elements lexicographically with respect to \mathbf{r} and then \mathbf{s} in case of a tie (in descending order), and analogously for $\rho_{\mathbf{s}, \mathbf{r}}$ (if two elements are at a tie in both vectors, their can be placed in any order, as their rank does not influence the value of $\tau_{\rho, w}$). We define

$$\tau_{w, \bullet}(\mathbf{r}, \mathbf{s}) := \frac{\tau_{\rho_{\mathbf{r}, \mathbf{s}}, w}(\mathbf{r}, \mathbf{s}) + \tau_{\rho_{\mathbf{s}, \mathbf{r}}, w}(\mathbf{r}, \mathbf{s})}{2}. \quad (3)$$

The same approach has been used in [27] to make AP correlation symmetric. This is the definition used in the rest of the paper.

4.2 Choosing a weighting scheme

There are of course many ways to choose w . For computational reasons, we will see that it is a good idea to restrict to a class of weighting schemes in which w is obtained by combining additively or multiplicatively a one-argument weighting function $f : [n] \rightarrow \mathbf{R}_{\geq 0}$ applied to each element of a pair.

Shieh [23], for instance, combines weights multiplicatively, without giving a motivation. We have, however, two important motivations for *adding* weights. First and foremost, unless weights are scaled in some way that depends on n (which we would like to avoid), the largest weight will be some constant, and then weight will decrease monotonically with importance. As a result, an exchange between the first and the last element would be assigned an extremely low

weight. Second, adding weights paves the way to a natural measure for *top k correlation* [6] by assigning rank ∞ to elements after the first k . The definition of such a measure in the multiplicative case is quite contrived and ends up being case-by-case.

For what matters f , we are particularly interested in the *hyperbolic* weight function.

$$f(r) := \frac{1}{r+1}.$$

This function gives more importance to elements of high rank, and weights zero only pairs in which both index have infinite rank. Using a hyperbolic weight has a number of useful features. First, it reminds the well-motivated weight given to exchanges by AP correlation. Second, it guarantees that as n grows the mass of weight grows indefinitely. Using a function with quadratic decay, for instance, might end up in making the influence of low-rank element vanish too quickly, as it is summable. For the opposite reason, a *logarithmic* decay might fail to be enough discriminative to provide additional information with respect to the standard τ .

We try to make this intuition more concrete in Figure 1, where we display a number of scatter plots showing the correlation between Kendall's τ and the additive weighted τ defined by (3) under different weighting schemes. The left half of the plots correlates all permutations on 12 elements with the identity permutation. The right half correlates all score vectors made of 15 values with skewed distribution (there are $t+1$ elements with score $0 \leq t \leq 4$) with the same vector in descending order. A visual examination of the plots suggests, indeed, that logarithmic weighting restricts too much the possible divergence from Kendall's τ , whereas quadratic weighting ends up in providing answers that are too uncorrelated. We will return to these consideration in Section 6.

5 Computing $\tau_{\rho,w}$

Our motivations come from the study of web and social graphs. It is thus essential that our new correlation measure can be evaluated efficiently. We now describe a generalization of Knight's algorithm [14] that makes it is possible to compute $\tau_{\rho,w}$ in time $O(n \log n)$ under some assumptions on w . Our first observation is that, similarly to the unweighted case, each pair of indices i, j with $i < j$ belongs to one of five subsets; it can be

- a *joint tie*, if $r_i = r_j$ and $s_i = s_j$;
- a *left tie*, if $r_i = r_j$ and $s_i \neq s_j$;
- a *right tie*, if $r_i \neq r_j$ and $s_i = s_j$;
- a *concordance*, if $\text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j) = 1$;
- a *discordance*, if $\text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j) = -1$.

Let J, L, R, C and D be the overall weight of joint ties, left ties, right ties, concordances and discordances, respectively. Clearly,

$$J + L + R + C + D = \sum_{i < j} w(\rho(i), \rho(j)) = T.$$

The first requirement for our technique to work is that T can be computed easily. This is possible if weights are computed additively or multiplicatively from some single-argument function f . In the additive case,

$$T = \sum_{i < j} (f(\rho(i)) + f(\rho(j))) = (n-1) \sum_i f(\rho(i)). \quad (4)$$

Also the multiplicative case is easy, as

$$2T = 2 \sum_{i < j} f(\rho(i))f(\rho(j)) = \left(\sum_i f(\rho(i)) \right)^2 - \sum_i f(\rho(i))^2. \quad (5)$$

The same observation leads to a simple $O(n \log n)$ algorithm to compute L : sort the indices in $[n]$ by r , and for each block of consecutive $k > 1$ elements with the same score apply (4) or (5) restricting the indices to the subset. In the same way one can compute R and J .

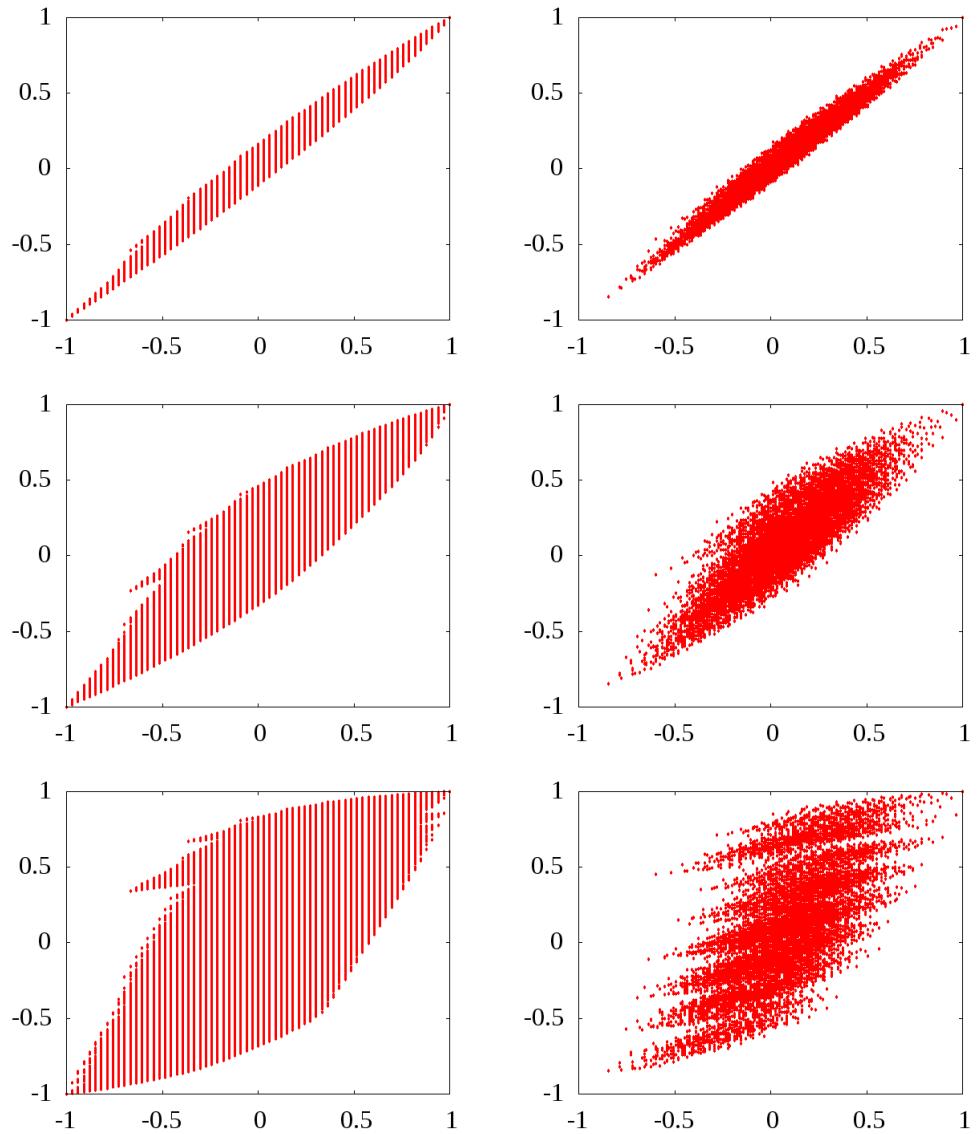


Figure 1: Scatter plots between Kendall's τ and the additive weighted τ . The rows, from top to bottom, represent logarithmic, hyperbolic and quadratic weighting. The plots are generated correlating a permutation of 12 elements versus the identity permutation (left), or a permuted set of scores with skewed distribution w.r.t. the same scores in descending order (right).

We now observe that, as in the unweighted case,

$$\langle \mathbf{r}, \mathbf{s} \rangle_{\rho, w} = C - D = T - (L + R - J) - 2D.$$

This can be easily seen from the fact that C is given by the total weight T , minus the weight of discordances D , minus the number of ties, joint or not, which is $L + R - J$ (we must avoid to count twice the weight of joint ties, hence the $-J$ term). In particular,

$$\langle \mathbf{r}, \mathbf{r} \rangle_{\rho, w} = T - L \quad \langle \mathbf{s}, \mathbf{s} \rangle_{\rho, w} = T - R,$$

as in this case there are just concordances and all ties are joint.

We are left with the computation of D . The core of Knight's algorithm is an *exchange counter*: an $O(n \log n)$ algorithm that given a list of elements and an order \preceq on the elements of the list computes the number of exchanges that are necessary to \preceq -sort the list. The algorithm is a modified MergeSort [15]⁴: during the merging phase, whenever an element is moved from the second list to the temporary result list the current number of elements of the first list is added to the number of exchanges. The number of discordances is then equal to the number of exchanges (as we evaluate whether there is a discordance on i and j only for $i < j$).

Our goal is to make this computation weighted: for this to happen, it must be possible to keep track incrementally of a *residual weight* r associated with the first list, and obtain in constant time the weight of the exchanges generated by the movement of an element from the second list.

If weights are computed multiplicatively or additively starting from a single-argument function f this is not difficult: it is sufficient to let r be the sum of the values of f applied to the elements currently in the first list. In the additive case, moving an element i from the second list increases the weight of exchanges by the residual r plus the weight $f(\rho(i))$ multiplied by the length of the first list. In the multiplicative case, we must instead use the weight $f(\rho(i))$ multiplied by the residual r . When we move an element from the first list we update the residual by subtracting its weight.

The resulting recursive procedure (for the additive case) is Algorithm 1. The final layout of the computation of $\tau_{\rho, w}$ is thus as follows:

- Consider a list \mathcal{L} initially filled with the integers $[0 \dots n]$.
- Sort stably \mathcal{L} using \mathbf{r} as primary key and \mathbf{s} as secondary key.
- Compute T and L using \mathcal{L} to enumerate elements in the order defined by \mathbf{r} and \mathbf{s} .
- Apply Algorithm 1 to \mathcal{L} using \mathbf{s} to define the order \preceq , thus computing D and sorting \mathcal{L} by \mathbf{s} .
- Compute R using \mathcal{L} to enumerate elements in the order defined by \mathbf{s} .
- Compute T and put everything together.

The running time of the computation is dominated by the sorting phases, and it is thus $O(n \log n)$.

5.1 The asymmetric case and AP Correlation

It is easy to adapt Algorithm 1 for the case in which $w(i, j)$ is given by a combination of *two* different one-argument functions, one, f , for the left index and one, g , for the right index. The only modification of Algorithm 1 is the replacement of f with g at line 14, so that we combine the residual computed with f with a weight computed with g .

The formulae for computing T can be updated easily for the additive case:

$$T = \sum_{i < j} (f(\rho(i)) + g(\rho(j))) = \sum_{i \neq 0} i(f(\rho(n - 1 - i)) + g(\rho(i)))$$

and for the multiplicative case:

$$T = \sum_{i < j} f(\rho(i))g(\rho(j)) = \sum_i f(\rho(i)) \sum_{i < j} g(\rho(j)).$$

⁴In principle, any stable algorithm that sorts by comparison could be used. This is particularly interesting as entirely on-disk algorithms, such as *polyphase merge* [15], could be used to count exchanges using constant core memory.

Algorithm 1 A generalization of Knight's algorithm for weighing exchanges.

Input: A list \mathcal{L} , a comparison function \preceq for the elements of \mathcal{L} , a rank function ρ , and a single-argument weight function f that will be combined additively. e is a global variable initialized to 0 that will contain the weight of exchanges after the call $\text{weigh}(0, |\mathcal{L}|)$. The procedure works on a sublist specified by its starting index $0 \leq s < |\mathcal{L}|$ and its length ℓ . \mathcal{T} is a temporary list.

Output: the sum of $f(\rho(-))$ on the specified sublist.

```

0  function weigh( $s$  : integer,  $\ell$  : integer)
1    if  $\ell = 1$  then return  $f(\rho(\mathcal{L}[s]))$  fi
2     $\ell_0 \leftarrow \lfloor \ell/2 \rfloor$ 
3     $\ell_1 \leftarrow \ell - \ell_0$ 
4     $m \leftarrow s + \ell_0$ 
5     $r \leftarrow \text{weigh}(s, \ell_0)$ 
6     $w \leftarrow \text{weigh}(m, \ell_1) + r$ 
7     $i, j, k \leftarrow 0$ 
8    while  $j < \ell_0$  and  $k < \ell_1$  do
9      if  $\mathcal{L}[s + j] \preceq \mathcal{L}[m + k]$  then
10         $\mathcal{T}[i] = \mathcal{L}[s + j++]$ 
11         $r \leftarrow r - f(\rho(\mathcal{T}[i]))$ 
12      else
13         $\mathcal{T}[i] = \mathcal{L}[m + k++]$ 
14         $e \leftarrow e + f(\rho(\mathcal{T}[i])) \cdot (\ell_0 - j) + r$ 
15      fi
16       $i++$ 
17    od
18    for  $k = \ell_0 - j - 1, \dots, 0$  do
19       $\mathcal{L}[s + i + k] \leftarrow \mathcal{L}[s + j + k]$ 
20    od
21    for  $k = 0, \dots, i - 1$  do  $\mathcal{L}[s + k] \leftarrow \mathcal{T}[k]$  od
22    return  $w$ 
23  end

```

Both formulae can be computed in linear time using a suitable loop.

Given this setup, it is easy to compute AP correlation: as it can be easily checked from the very definition [27], the AP correlation of \mathbf{r} w.r.t. \mathbf{s} , where both vectors have no ties, is simply $\tau_{w, \rho_s}(\mathbf{r}, \mathbf{s})$, where ρ_s is the ranking induced by \mathbf{s} and the weight function w is computed additively from two weight functions $f(r) = 0$, $g(r) = 1/r$. In this case, $T = n - 1$, $J = L = R = 0$ (we are under the assumption that there are no ties) and Algorithm 1 can be considerably simplified, as the residual r is always zero.⁵

Algorithm 2 makes explicit the change to the selection statement of Algorithm 1 that is sufficient to compute AP correlation. Since keeping track of the residual is no longer necessary, the recursive function can be further simplified to a recursive procedure that does not return a value. The value e computed by the modified algorithm is all we need to compute AP correlation using the formula $(T - 2e)/T$.

⁵Of course, it is possible to forget that we are computing AP correlation and use the weight matrix just described combined with the machinery of Section 4 to define an ‘‘AP correlation with ties’’. In this case, J , L and R should be computed using the formulae for the asymmetric case, and the probabilistic interpretation would be lost. Such an index would probably give a notion of correlation very similar to τ_h , but we find more natural and more in line with Kendall’s original definition to introduce the weighted τ as a symmetric index in which both ends of an exchange are relevant in computing the exchange weight.

Algorithm 2 The replacement for lines 9–15 of Algorithm 1 to compute AP correlation.

```

9  if  $\mathcal{L}[s + j] \preceq \mathcal{L}[m + k]$  then
10    $\mathcal{T}[i] = \mathcal{L}[s + j++]$ 
11  else
12    $\mathcal{T}[i] = \mathcal{L}[m + k++]$ 
13    $e \leftarrow e + (\ell_0 - j)/\rho(\mathcal{T}[i])$ 
14 fi

```

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.95	0.98	0.90	0.27
PageRank	0.95	1	0.96	0.92	0.65
Katz	0.98	0.96	1	0.93	0.26
Harmonic	0.90	0.92	0.93	1	0.28
Closeness	0.27	0.65	0.26	0.28	1

Table 6: τ_h on Wikipedia.

6 Experiments

We now return to our main motivation—understanding the correlation between centralities on large graph. In this section, we gather the results of a number of computational experiment that help to corroborate our intuition that τ_h , the *additive hyperbolic weighted τ* , works as expected. We will find also an interesting surprise along the way.

Note that judging whether a new measure is useful for such a purpose is a difficult task: to be interesting, a new measure must highlight features that were previously undetectable or badly evaluated, but those are exactly those features on which a systematic assessment is problematic.

Table 6 reports the value of τ_h on the Wikipedia graph. We finally see data corresponding to the empirical evidence discussed in Section 3: indegree, Katz and PageRank are almost identical, harmonic centrality is highly correlated but definitely less than the previous triple, which matches our empirical observations. Closeness is not close to any ranking (and in particular, not to harmonic centrality) due to its pathological behavior.

There is of course a value that immediately stands out: the suspiciously high correlation (0.65) between closeness and PageRank. We reserve discussing this value for later.

In Table 7 we show the same data for logarithmic and quadratic weights. The intuition we gathered from Figure 1 is fully confirmed: logarithmic weights provides results almost indistinguishable from Kendall’s τ (compare with Table 3), and quadratic weights make the influence of the tail so low that all non-pathological scores collapse.

To gather a better understanding of the behavior of τ_h we extended our experiments to two very different datasets: the *Hollywood co-starship graph*, an undirected graph (2 million nodes, 229 million edges) with an edge between two persons appearing in the Internet Movie Data Base if they ever worked together, and a *host graph* (100 million nodes, 2 billion arcs) obtained from a large-scale crawl gathered by the Common Crawl Foundation⁶ in the first half of 2012.⁷ As (unavoidably anecdotal) empirical evidence we report the top 20 nodes for both graphs.

Table 8 should be compared with Table 10. PageRank and harmonic centrality turns to be less correlated to indegree than Katz in Table 8, and indeed we find many quirky choices in the very top PageRank actors (Ron Jeremy is a famous porn star; Lloyd Kaufman is an independent horror/splatter filmmaker and Debbie Rochon an actress working with him). Harmonic centrality provides unique names such as Malcolm McDowell, Robert De Niro, Anthony Hopkins and Sylvester Stallone, and drops all USA presidents altogether. Kendall’s τ values, instead, suggest that PageRank and harmonic centrality are entirely uncorrelated (whereas we find several common items), and that harmonic and closeness centrality should be extremely similar.

We see analogous results comparing Table 9 with Table 11. Here τ_h separates in a very strong way harmonic centrality from the first three, and indeed we see a significant difference in the lists, with numerous sites that have a high indegree and appear in at least two of the three lists because of technical or political reasons (gmpg.org,

⁶<http://commoncrawl.org/>

⁷The crawl contains 3.53 billion web documents; we are using the associated host graph, which has a node for each host and an arc between two hosts x and y if some page in x points to some page in y . More information about the graph can be found in [18], and the complete host ranking can be accessed at <http://wwwranking.webdatacommons.org/>.

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.76	0.90	0.63	0.55
PageRank	0.76	1	0.76	0.62	0.56
Katz	0.90	0.76	1	0.70	0.62
Harmonic	0.63	0.62	0.70	1	0.91
Closeness	0.55	0.56	0.62	0.91	1

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	1.00	1.00	1.00	0.22
PageRank	1.00	1	1.00	1.00	0.85
Katz	1.00	1.00	1	1.00	0.18
Harmonic	1.00	1.00	1.00	1	0.07
Closeness	0.22	0.85	0.18	0.07	1

Table 7: The logarithmic (top) and quadratic (bottom) additive τ on Wikipedia.

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.42	0.93	0.55	0.43
PageRank	0.42	1	0.36	0.10	0.18
Katz	0.93	0.36	1	0.61	0.49
Harmonic	0.55	0.10	0.61	1	0.86
Closeness	0.43	0.18	0.49	0.86	1

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.90	0.98	0.91	0.10
PageRank	0.90	1	0.88	0.81	0.64
Katz	0.98	0.88	1	0.92	0.11
Harmonic	0.91	0.81	0.92	1	0.18
Closeness	0.10	0.64	0.11	0.18	1

Table 8: Kendall’s τ (top) and τ_h (bottom) on the Hollywood co-starship graph.

rtalabel.org, staff.tumblr.com, miibeian.gov.cn, phpbb.com) disappearing altogether in favor of sites such as apple.com, amazon.com, myspace.com, microsoft.com, bbc.co.uk, nytimes.com and guardian.co.uk, which do not appear in any other list. If we look at Kendall’s τ , we should expect PageRank and Katz to give very different rankings, whereas more than half of their top 20 elements are in common.

6.1 PageRank and closeness

It is now time to examine the mysteriously high τ_h between PageRank and closeness we found in all our graphs. When we first computed our correlation tables, we were puzzled by its value. The phenomenon is interesting for three reasons: first, it has never been reported—using standard, unweighted indices this correlation is simply undetectable; second, it was known for techniques based on singular vectors [17]; third, we know *exactly* the cause of this correlation, because the only real difference between harmonic and closeness centrality is the score assigned to nodes unreachable from the giant component. We thus expect to discover an unsuspected correlation between the way PageRank and closeness rank these nodes.

To have a visual understanding of what is happening, we created Figure 2, 3 and 4 in the following way: first, we isolated the nodes that are unreachable from the giant component (in the case of Hollywood, which is undirected, these nodes form separate components), omitting nodes which have indegree zero, modulo loops (as all measures give the lowest score to such nodes); then, we sorted the nodes in order of decreasing closeness rank, and plotted for each node

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.71	0.89	0.61	0.54
PageRank	0.71	1	0.66	0.50	0.50
Katz	0.89	0.66	1	0.69	0.59
Harmonic	0.61	0.50	0.69	1	0.86
Closeness	0.54	0.50	0.59	0.86	1

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.91	0.96	0.72	0.20
PageRank	0.91	1	0.90	0.81	0.69
Katz	0.96	0.90	1	0.78	0.15
Harmonic	0.72	0.81	0.78	1	0.35
Closeness	0.20	0.69	0.15	0.35	1

Table 9: Kendall's τ (top) and τ_h (bottom) on the on the Common Crawl host graph.

Indegree	PageRank	Katz	Harmonic	Closeness
Shatner, William	Jeremy, Ron	Shatner, William	Sheen, Martin	Östlund, Claes Göran
Flowers, Bess	Hitler, Adolf	Sheen, Martin	Clooney, George	Östlund, Catarina
Sheen, Martin	Kaufman, Lloyd	Hanks, Tom	Jackson, Samuel L.	von Preußen, Oskar Prinz
Reagan, Ronald (I)	Bush, George W.	Williams, Robin (I)	Hopper, Dennis	von Preußen, Georg Friedrich
Clooney, George	Reagan, Ronald (I)	Clooney, George	Hanks, Tom	von Mannstein, Robert Grund
Jackson, Samuel L.	Clinton, Bill (I)	Reagan, Ronald (I)	Stone, Sharon (I)	von Mannstein, Concha
Williams, Robin (I)	Sheen, Martin	Willis, Bruce	Brosnan, Pierce	von der Busken, Mart
Hanks, Tom	Rochon, Debbie	Jackson, Samuel L.	Hitler, Adolf	van der Putten, Thea
Jeremy, Ron	Kennedy, John F.	Stone, Sharon (I)	McDowell, Malcolm	de la Bruheze, Joel Albert
Hitler, Adolf	Hopper, Dennis	Freeman, Morgan (I)	Williams, Robin (I)	de la Bruheze, Emile
Willis, Bruce	Nixon, Richard	Flowers, Bess	De Niro, Robert	te Riele, Marloes
Clinton, Bill (I)	Estevez, Joe	Brosnan, Pierce	Willis, Bruce	de Reijer, Eric
Freeman, Morgan (I)	Shatner, William	Douglas, Michael (I)	Hopkins, Anthony	des Bourvieu, Jan
Hopper, Dennis	Jackson, Samuel L.	Madonna (I)	Madonna (I)	de Klijn, Judith
Stone, Sharon (I)	Stewart, Jon (I)	Travolta, John	Lee, Christopher (I)	de Freitas, Luís (II)
Madonna (I)	Carradine, David (I)	Hopper, Dennis	Douglas, Michael (I)	de Freitas, Luís (I)
Bush, George W.	Asner, Edward	Ford, Harrison (I)	Sutherland, Donald (I)	Zuu, Winnie Otondi
Harris, Sam (II)	Zirnkilton, Steven	Asner, Edward	Freeman, Morgan (I)	Zuu, Emmanuel Dahngbay
Brosnan, Pierce	Colbert, Stephen	MacLaine, Shirley	Stallone, Sylvester	Zilbersmith, Carla
Travolta, John	Madsen, Michael (I)	Clinton, Bill (I)	Ford, Harrison (I)	Zilber, Mac

Table 10: Top 20 pages of the Hollywood co-starship graph.

its rank following the other measures (we average ranks on block of nodes so to contain the number of points in the plots). A point of high abscissa in the figures implies a high rank.

All three pictures show clearly that *PageRank assigns a preposterously high rank to nodes belonging to components that are unreachable from the giant component*. This behavior is actually related to PageRank's *insensitivity to size*: for instance, in a graph made of two components, one of which is a 3-clique and the other a k -clique, the PageRank score of all nodes is $1/(3+k)$, independently of k . This explains why small dense components end up being so highly ranked. The same phenomenon is at work when the community around Lloyd Kaufman's production company (very small and very dense) is attributed such a great importance that its elements make their way to the very top ranks (even if Kaufman himself has indegree rank 219 and Debbie Rochon 1790).

We remark that the gap in rank is lower in the case of Wikipedia, but this is fully in concordance with the higher baseline value of Kendall's τ .

7 Conclusions

In this paper, motivated by the need to understand similarity between score vectors, such as those generated by centrality measures on large graphs, we have defined a weighted version of Kendall's τ starting from its 1945 definition for scores with ties. We have developed the mathematical properties of our generalization following a mathematical similarity with internal products, and showing that for a wide range of weighting schemes our new measure behaves as expected, providing a correlation index between -1 and 1, and hitting boundaries only for opposite or equivalent scores.

Indegree	PageRank	Katz	Harmonic	Closeness
wordpress.org	gmpg.org	wordpress.org	youtube.com	0-p.com
youtube.com	wordpress.org	youtube.com	en.wikipedia.org	0-0-0-0-0-0.indahiphop.ru
gmpg.org	youtube.com	gmpg.org	twitter.com	0-0-1.i.tiexue.net
en.wikipedia.org	livejournal.com	en.wikipedia.org	google.com	0-00cigarettes.info
tumblr.com	tumblr.com	tumblr.com	wordpress.org	0-Omos00.hi5.com
twitter.com	en.wikipedia.org	twitter.com	flickr.com	0-0new0-0.hi5.com
google.com	twitter.com	google.com	facebook.com	0-0sunny0-0.hi5.com
flickr.com	networkadvertising.org	flickr.com	apple.com	0-1.i.tiexue.net
rtalabel.org	promodj.com	rtalabel.org	vimeo.com	0-1.sxsy.co
wordpress.com	skriptmail.de	wordpress.com	creativecommons.org	0-2.paparazziwannabe.com
mp3shake.com	parallels.com	mp3shake.com	amazon.com	0-311.cn
w3schools.com	tistory.com	w3schools.com	adobe.com	0-360.rukazan.ru
domains.lycos.com	google.com	creativecommons.org	myspace.com	0-5days.com
staff.tumblr.com	miibeian.gov.cn	staff.tumblr.com	w3.org	0-5days.net
club.tripod.com	phpbb.com	domains.lycos.com	bbc.co.uk	0-5kalibr.pdj.ru
creativecommons.org	blog.fc2.com	club.tripod.com	nytimes.com	0-9-0-4-4-9.promoradio.ru
vimeo.com	tw.yahoo.com	vimeo.com	yahoo.com	0-9-0-9.dbass.ru
miibeian.gov.cn	w3schools.com	miibeian.gov.cn	microsoft.com	0-9-0-9.promodj.ru
facebook.com	wordpress.com	facebook.com	guardian.co.uk	0-9-1125.i.tiexue.net
phpbb.com	domains.lycos.com	phpbb.com	imdb.com	0-9-7-16.software.informer.com

Table 11: Top 20 hosts of the Common Crawl host graph.

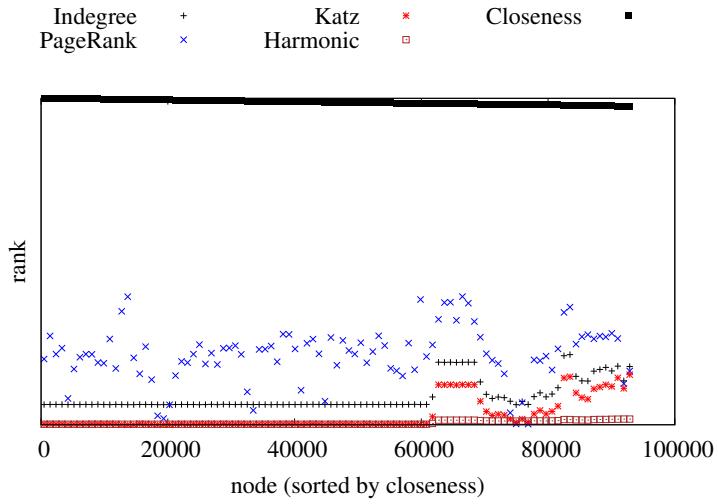


Figure 2: Ranks of components unreachable from the giant component of the Wikipedia graph.

We have then proposed families of weighting schemes that are intuitively appealing, and showed that they can be computed in time $O(n \log n)$ using a generalization of Knight’s algorithm, which makes them suitable for large-scale applications. The fact that the main cost of the algorithm is a modified stable sort makes it possible to apply standard techniques to run the algorithm exploiting multicore parallelism, or in distributed environment such as MapReduce [5]. The algorithm can be also used to compute AP correlation [27].

In search for a confirmation of our mathematical intuition, we have then applied our measure of choice τ_h (which uses additive hyperbolic weights) to diverse graph such as Wikipedia, the Hollywood co-starship graph and a large host graph, finding that, contrarily to Kendall’s τ , τ_h provides results that are consistent with an anecdotal examination of lists of top elements.

Our measure was also able to discover a previously unnoticed correlation between PageRank and closeness on small components that are unreachable from the giant component, providing a quantifiable account of the strong bias of PageRank towards small-sized dense communities. This bias might well be the cause of the repeatedly assessed better performance of indegree w.r.t. PageRank in ranking documents [20, 3], as in all our experiments the τ_h between PageRank and indegree is above 0.9.

A generalization similar to the one described in this paper can be also applied to *Goodman–Kruskal’s γ* , which in the notation of Section 5 is just $(C - D)/(C + D)$. The problem with γ is that the ranking of ties is only implicit (they are simply not counted). Thus, the value of w on tied pairs does not appear at all in the above formula. This “forgetful” behavior can lead to unnatural results, and suggests the Kendall’s τ is a better candidate for this approach.

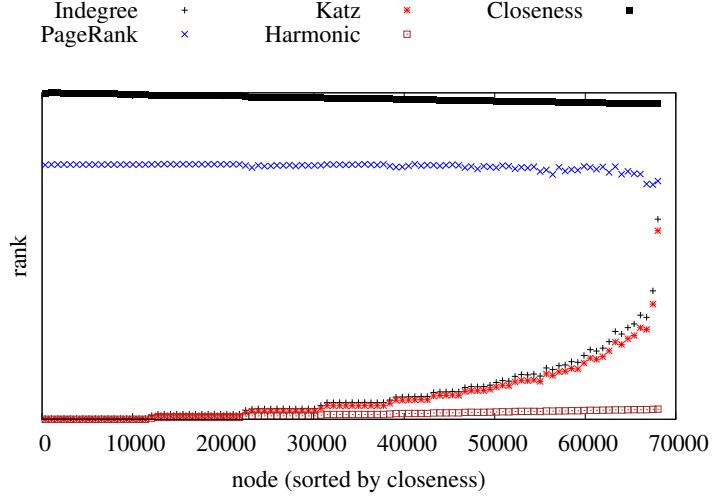


Figure 3: Ranks of components unreachable from the giant component of the Hollywood.

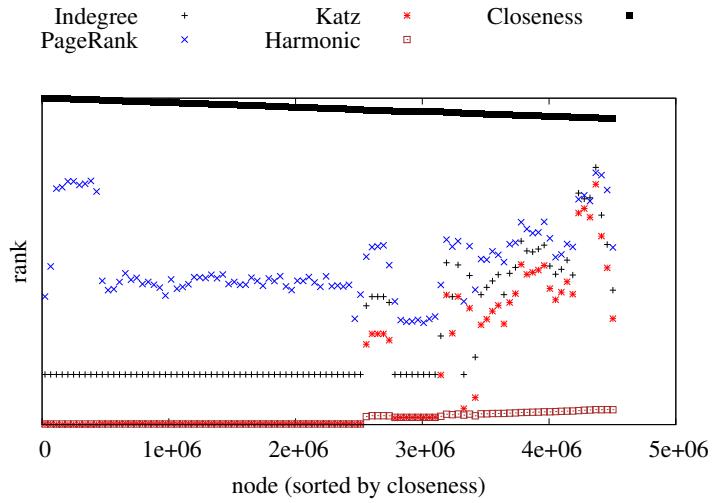


Figure 4: Ranks of components unreachable from the giant component of the Common Crawl host graph.

We remark that an interesting application of additive hyperbolic weighting is that of measuring the correlation between top k lists. By assuming that the rank function ρ returns ∞ after rank k , we obtain a correlation index that weighs zero pairs outside the top k , weights only “by one side” pairs with just one element outside the top k , and weights fully pairs whose elements are within the top k . Formula (3) could provide then in principle a finer assessment than, for instance, the modified Kendall’s τ proposed in [6], as the position of each element inside the list, beside the fact that it appears in the top k or not, would be a source of weight. We leave the analysis of such a correlation measure for future work.

References

- [1] Alex Bavelas. Communication patterns in task-oriented groups. *J. Acoust. Soc. Am.*, 22(6):725–730, 1950.
- [2] Paolo Boldi and Sebastiano Vigna. Axioms for centrality. *CoRR*, abs/1308.2140, 2013. To appear in *Internet Mathematics*.

- [3] Nick Craswell, David Hawking, and Trystan Upstill. Predicting fame and fortune: PageRank or indegree? In *In Proceedings of the Australasian Document Computing Symposium, ADCS2003*, pages 31–40, 2003.
- [4] Henry E. Daniels. The relation between measures of correlation in the universe of sample permutations. *Biometrika*, 33(2):129–135, 1943.
- [5] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. In *OSDI '04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150, 2004.
- [6] Ronald Fagin, Ravi Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.
- [7] Farzad Farnoud and Olgica Milenkovic. Aggregating rankings with positional constraints. In *Proc. IEEE Information Theory Workshop (ITW)*, Seville, Spain, 2013.
- [8] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61, 2013.
- [9] Ronald L. Iman and W. J. Conover. A measure of top-down correlation. *Technometrics*, 29(3):351–357, 1987.
- [10] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.
- [11] John G. Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959.
- [12] Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
- [13] Maurice G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.
- [14] William R. Knight. A computer method for calculating Kendall’s tau with ungrouped data. *Journal of the American Statistical Association*, 61(314):436–439, June 1966.
- [15] Donald E. Knuth. *Sorting and Searching*, volume 3 of *The Art of Computer Programming*. Addison-Wesley, second edition, 1997.
- [16] Ravi Kumar and Sergei Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web*, pages 571–580. ACM, 2010.
- [17] Ronny Lempel and Shlomo Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks*, 33(1):387–401, 2000.
- [18] Robert Meusel, Sebastiano Vigna, Oliver Lehmburg, and Christian Bizer. Graph structure in the web — Revisited, or a trick of the heavy tail. In *WWW’14 Companion*, pages 427–432. International World Wide Web Conferences Steering Committee, 2014.
- [19] Carl D. Meyer. *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics, 2000.
- [20] Marc Najork, Hugo Zaragoza, and Michael J. Taylor. HITS on the web: how does it compare? In Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 471–478. ACM, 2007.
- [21] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library Technologies Project, Stanford University, 1998.
- [22] I. Richard Savage. Contributions to the theory of rank order statistics—the two-sample case. *The Annals of Mathematical Statistics*, 27(3):590–615, 1956.
- [23] Grace S. Shieh. A weighted kendall’s tau statistic. *Statistics & Probability Letters*, 39(1):17–24, 1998.
- [24] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

- [25] Ellen M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82. ACM, 2001.
- [26] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38, 2010.
- [27] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594. ACM, 2008.

Local Ranking Problem on the BrowseGraph

Michele Trevisiol*†
trevisiol@acm.org

Luca Maria Aiello[†]
alucca@yahoo-inc.com

Paolo Boldi[§]
boldi@di.unimi.it

Roi Blanco[†]
roi@yahoo-inc.com

[†]Yahoo Labs
Barcelona, Spain

*Web Research Group
Universitat Pompeu Fabra
Barcelona, Spain

§ Univ. degli Studi di Milano
Milano, Italy

ABSTRACT

The “Local Ranking Problem” (LRP) is related to the computation of a centrality-like rank on a *local* graph, where the scores of the nodes could significantly differ from the ones computed on the *global* graph. Previous work has studied LRP on the hyperlink graph but never on the *BrowseGraph*, namely a graph where nodes are webpages and edges are browsing transitions. Recently, this graph has received more and more attention in many different tasks such as ranking, prediction and recommendation. However, a web-server has only the browsing traffic performed on its pages (*local BrowseGraph*) and, as a consequence, the local computation can lead to estimation errors, which hinders the increasing number of applications in the state of the art. Also, although the divergence between the local and global ranks has been measured, the possibility of *estimating* such divergence using only local knowledge has been mainly overlooked. These aspects are of great interest for online service providers who want to gauge their ability to correctly assess the importance of their resources only based on their local knowledge, and by taking into account real user browsing fluxes that better capture the actual user interest than the static hyperlink network. We study the LRP problem on a *BrowseGraph* from a large news provider, considering as subgraphs the aggregations of browsing traces of users coming from different domains. We show that the distance between rankings can be accurately predicted based only on structural information of the local graph, being able to achieve an average rank correlation as high as 0.8.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous; E.1 [Data Structures]: Graphs and Networks

Keywords

Local Ranking Problem, BrowseGraph, PageRank, Centrality Algorithms, Domain-specific Browsing Graphs

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 20XX ACM X-XXXXXX-XX-X/XX/XX ...\$10.00.

1. INTRODUCTION

The ability to identify the online resources that are perceived as important by the users of a website is crucial for online service providers. Metrics to estimate the importance of the page from the structure of online links between them are widely used: algorithms that compute the *centrality* of the nodes in a network, such as PageRank [24], HITS [17] and SALSA [19], have been employed extensively in the last two decades in a vast variety of applications. Born and spread in conjunction with the growth of the Web, they can determine a value of importance of a page from the complex network of links that surrounds it. More recently, centrality metrics have been applied to *browsing graphs*, (also referred to as *BrowseGraphs* [22, 28, 27]) where nodes are webpages and edges represent the transitions made by the users who navigate the links between them. Differently from the hyperlink networks, this data source provides the analyst a way of studying directly the dynamics of the navigational patterns of users who consume online content. Also, unlike hyperlinks, browsing traces account for the variation of consumption patterns in time, for instance in the case of online news where articles tend to become rapidly stale. Comparative studies have shown that centrality-based algorithms applied over *BrowseGraphs* provide higher-quality rankings compared to standard hyperlink graphs [23, 22].

Most centrality measures aim at estimating the importance of a node, using information coming from the *global* knowledge of the graph topology—potentially the addition of new nodes and edges, can have a cascade effect on the centrality values of all other nodes in the network. This fact entails high computational and storage cost for big networks. More critically, there are some situations in which a global computation on the entire graph is unfeasible, for example when the information about the entire network is unavailable or if only an estimation for specific web pages is required. This is an important limitation in many real-world scenarios, where the graphs at hand are often very large (Web scale) and, most importantly, their topology is not fully known. This practical issue raises the problem of how well one can estimate the actual centrality value of a node by knowing only a local portion of the graph. This is known as the *Local Ranking Problem* (LRP) [10].

One of the questions behind LRP is whether it is possible to estimate efficiently the PageRank score of a web page using only a small subgraph of the entire Web [9]. In other words, if one starts from a small graph around a page of interest and extends it with external nodes and arcs (*i.e.*, those belonging to the whole graph), how fast will one ob-

serve the computed scores converging to the real values of PageRank?

We extend this line of work in the context of browsing graphs. For the first time we study the LRP on the *BrowseGraph* and shed some light on the bias that PageRank incurs, when estimating the centrality score of nodes in a *BrowseGraph*, when only partial information about the graph is available. To achieve that, we monitor the browsing traffic of the news portal and we extract different browsing subgraphs induced by the browsing traces of users coming from different *domains*, such as search engines (*e.g.*, Google, Yahoo, Bing) and social networks (*e.g.*, Facebook, Twitter, Reddit). In this setting, the local *BrowseGraphs* are the subgraphs induced by the different domains, and the global *BrowseGraph* is the one built using indistinctly all the navigation logs of the news portal. We describe and evaluate models that tell apart a subgraph from the others just by looking at the behavior of a random surfer that navigates through their links. The results show how it is possible to recognize the graph using only the very first few nodes visited by the users, because the graphs are very different among them (even if they are extracted from the same big log of the news portal). The implication of this experiment is two-fold: first it highlights how navigation patterns of the users differ among these subgraphs. Second, we learn that it is possible to infer the user domain of origin from the very first browsing steps. This capability enables several types of services, including user profiling [12], web site optimization [31], user engagement estimation [18], and cold-start recommendation [27], even when the referrer URL is not available (*e.g.* when the user comes from mobile social media applications or URL shortening services).

Once we show that the subgraphs are different enough, we proceed to perform more involved experiments that we call “Growing Balls”. We examine the behavior of the PageRank computed on the local and the global graphs. In order to study how the local PageRank converges to the global one, we apply some strategies of incremental addition (“growing”) of external nodes to these subgraphs (“balls”).

Finally, we build on these findings by setting up a prediction experiment that, for the first time, tackles the task of estimating the reliability of the PageRank computed locally. We measure *how much* the local PageRank diverges from the global one using only structural features of the local graph, usually available to the local service provider.

To sum up, the main contributions of this work are the following:

- We study the *LRP* on a large-scale *BrowseGraph* built from a very popular news website. To the best of our knowledge we are the first to tackle this problem on the increasingly popular *BrowseGraph* [27, 28, 12, 22]. We present an analysis of the convergence of the PageRank on the local graph to the global one, by incrementally expanding the local graph in a snowball fashion.
- We tackle the problem of discovering the referrer domain of a user session, when this information is missing or hidden. We show that this is possible using a random surfer model, that is able to tell the referrer domain with high accuracy, just after the very first browsing transitions.
- We show that an accurate estimation of the distance between the local and global PageRank can be obtained

looking at the structural properties of the local graph, such as degree distribution or assortativity.

The remainder of the paper is organized as follows. In §2, we overview relevant prior work in the area and in §3 we describe our dataset and the extraction of the browsing graphs. In §4 we analyze the (sub-)graphs and we highlight their differences. In §5 we study the LRP problem on the *BrowseGraph* and compare the approximation accuracy of different graph expansion strategies. In §6 we present the prediction experiment of the PageRank errors of the local graph. Last, in §7 we wrap up and highlight possible extensions to the work.

2. RELATED WORK

This work encompasses two main different research areas that we introduce shortly. Our focus is the *Local Ranking Problem* but our contribution relates also to previous work on browsing log data, especially the ones that investigate or make use of centrality-based algorithms.

Local Ranking Problem

The *Local Ranking Problem* (LRP) was first introduced by Chen *et al.* [10] in 2004, who addressed the problem to approximate/update the PageRank of individual nodes, without performing a large-scale computation on the entire graph. They proposed an approach that can tackle this problem by including a moderate number of nodes in the local neighborhood of the original nodes. Furthermore, Davis and Dhillon [14] estimated the global PageRank values of a local network using a method that scales linearly with the size of the local domain. Their goal was to rank webpages in order to optimize their crawling order, something similar to what was done by Cho *et al.* [13] who instead selected the top-ranked pages first. However, this latter strategy results to be in contrast with Boldi *et al.* [6], as they found that crawling first the pages with highest global PageRank actually perform worse, if the purpose is fast convergence to the real (global) rank values. In this work, we partial expand the local graph with the neighboring nodes with highest (local) PageRank showing an initial improvement on the convergence speed. In 2008 the problem was reconsidered by Bar-Yossef and Mashiach [3], where they simplified the problem calculating a local *Reverse PageRank* proving that it is more feasible and computationally cheaper, as the reverse natural graphs tend to have low in-degree maintaining a fast PageRank convergence. Bressan and Pretto [9] proved that, in the general case, an efficient local ranking algorithm does not exist, and in order to compute a *correct* ranking it is necessary to visit at least a number of nodes linear in the size of the input graph. They also raised some of the research questions tackled in our paper that we discuss in Section 6.1. They reinforce their findings in later work [8], where they summarized two key factors necessary for efficient local PageRank computations: *exploring the graph non-locally* and *accepting a small probability error*. These two constraints are also considered in this paper in order to perform our experiments on the browsing graphs. When one wants to estimate PageRank in a local graph, the problem of the missing information is tackled in various ways. In [3, 9] for example, the authors make use of a model called *link server* (also known as *remote connectivity server* [5]), that responds to any query about a given node with all the in-coming and out-going edges and

relative nodes. This approach, with the knowledge about the LRP, allows to estimate the PageRank ranking, or even the score, with the relative costs. A similar problem was studied by Andersen *et al.* [2], where their goal was to compute the PageRank contributions in a local graph motivated by the problem of detecting link-spam: given a page, its PageRank contributors are the pages that contribute most to its rank; contributors are used for spam detection since you can quickly identify the set of pages that contribute significantly to the PageRank of a suspicious page.

The problem we consider here is different and largely unexplored, because we are studying the PageRank of the different subgraphs based on user browsing patterns.

BrowseGraph

In recent years a large number of studies of user browsing traces have been conducted. Specifically, in the last years there was a surge of interest in the *BrowseGraph*, a graph where the nodes are web pages and the edges represent the transitions from one page to another made by the navigation of the users. Characterizing the browsing behavior of users is a valuable source of information for a number of different tasks, ranging from understanding how people's search behaviors differ [32], ranking webpages through search trails [1, 33] or recommending content items using past history [29]. A comparison between the standard hyperlink graph, based on the structure of the network, with the browse graph built by the users' navigation patterns, has been made by Liu *et al.* [22, 23]. They compared centrality-based algorithms like PageRank [24], TrustRank [15], and BrowseRank [22], on both types of graphs. The results agree on the higher quality of ranking based on the browse graph, because it is a more reliable source; they also tried out a combination of the two graphs with very interesting outcomes. The user browsing graph and related PageRank-like algorithms have been exploited to rank different types of items including images [28, 12], photostreams [11], and predicting users demographic [16] or optimizing web crawling [21]. Trevisiol *et al.* [28] made a comparison between different ranking techniques applied to the Flickr *BrowseGraph*. Chiarandini *et al.* [12] found strong correlations between the type of user's navigation and the type of external Referrer URL. Hu *et al.* [16] have shown that demographic information of the users (*e.g.*, age and gender) can be identified from their browsing traces with good accuracy. The *BrowseGraph* has been used also for recommending sequences of photos that users often like to navigate in sequence, following a collaborative filtering approach [11]. In order to implement an efficient news recommender the user's taste have to be considered as they might change over time. Indeed, studying the users browsing patterns, Liu *et al.* [20] showed that more recent clicks have a considerably higher value to predict future actions than the historical browsing record. Finally, Trevisiol *et al.* [27] exploited the *BrowseGraph* in order to build some user models in the news domain, and recommend the next article the user is going to visit. They introduced the concept of *ReferrerGraph*, that is a *BrowseGraph* built with sessions that are generated by the same referrer domain. Even if the purposes of our work are very different, we construct the *ReferrerGraphs* in the same way in order to be in-line with their investigation.

To the best of our knowledge there is no work in the state of the art that tackles the *Local Ranking Problem* on a

browsing graphs with the prediction task that we perform and describe in this paper.

3. DATASET

For the purpose of this study, we took a sample of Yahoo News network's¹ user-anonymized log data collected in 2013. In this section we summarize how we built the dataset and the graphs, but the reader may refer to the aforementioned paper for further details. The data is comprised by a large number of pageviews, which are represented as plain text files that contain a line for each HTTP request satisfied by the Web server. For each pageview in the dataset, we gathered the following fields:

$\langle BCookie, Time, ReferrerURL, CurrentURL, UserAgent \rangle$

The *BCookie* is an anonymized identifier computed from the browser cookie. This information allowed us to re-construct the navigation session of the different users. *CurrentURL* and *ReferrerURL* represent, respectively, the current page the user is visiting and the page the user visited before arriving at the destination page. Note that the *Referrer URL* could belong to any domain, *e.g.*, it may be external to the Yahoo News network. The *User-Agent* identifies the user's browser, an information that we used to filter out Web crawlers, and *Timestamp* indicates when the page was visited. All the data were anonymized and aggregated prior to building the browsing graphs. After applying the filtering steps described above, our sample contains approximately 3.8 million unique pageviews and 1.88 billion user transitions.

3.1 Session Identification and Characteristics

The *BrowseGraph* is a graph whose nodes are web pages, and whose edges are the browsing transitions made by the users. To build it we extract the transitions of users from page to page, and in order to preserve the user behavior (that could vary over time), we group pageviews into *sessions*. We split the activity of a single user, taking the *BCookie* as an identifier, into different sessions when either of these two conditions holds:

- **Timeout:** the inactivity between two pageviews is longer than 25 minutes.
- **External URL:** if a user leaves the news platform and returns from an external domain, the current session ends even if previous visits are within the 25 minute threshold.

Moreover, each news article of the dataset is annotated with a high-level *category* manually assigned by the editors.

3.2 Subgraphs Based on Session Referrer URL

We aim to compare the PageRank scores of the nodes between the full *BrowseGraph*, computed with all the Yahoo News logs, and a subgraph that represents the local graph. This is a way to simulate a real-world scenario in which a service provider knows only the users navigation logs inside its network (subgraph) while the external navigations are unknown (full *BrowseGraph*). Since it is not possible to use the full Web browsing log, we perform a simulation

¹We considered a number of different subdomains like *Yahoo news, finance, sports, movies, travel, celebrity, etc.*

Subgraphs	Nodes	Edges	Density	%GCC
Google	142,646	779,185	$3.8 \cdot 10^{-5}$	0.93
Yahoo	101,116	404,378	$3.9 \cdot 10^{-5}$	0.95
Bing	61,531	255,464	$6.7 \cdot 10^{-5}$	0.91
Homepage	60,287	335,836	$9.2 \cdot 10^{-5}$	0.99
Facebook	21,060	70,266	$1.5 \cdot 10^{-4}$	0.95
Twitter	4,206	7,080	$4.0 \cdot 10^{-4}$	0.87
Reddit	2,445	4,868	$8.1 \cdot 10^{-4}$	0.95

Table 1: Size of the extracted subgraphs. Note that there is not a strict relation between the size of the subgraph and the amount of browsing traffic generated in it.

using different subgraphs extracted from the same *BrowseGraph* that represent the local graphs of different providers. In order to do that, we extract from the *BrowseGraph* of the Yahoo News dataset various subgraphs built with sessions of users generated by the same Referrer URL. It has been shown [27] that a *BrowseGraphs* constructed in this way contain very different users sessions in terms of content consumed (nodes visited). In particular we consider users accessing the news portal directly from the homepage, that is the main entry point for regular news consumption, and in addition, from a number of domains that fall outside the Yahoo News network: *search engines* (Google, Yahoo, Bing), and *social networks* (Facebook, Twitter, Reddit). For each source domain we extract a subgraph from the overall *BrowseGraph*, by considering only the browsing sessions whose initial Referrer URL matches that domain. For example, if a user clicks on a link referring to our network that has been posted on Twitter, her Referrer URL will be the Twitter page where she found the link. Next, we consider all the following pageviews belonging to the same session of the user, as being a part of the *twitter-subgraph*, given that all of them have been reached through Twitter. We applied the same procedure for all the sources defined before, and finally, we obtained a weighted graph for each different external URL, where the *Weight* accounts for the number of times a user has navigated from the source page to the destination page. On Table 1 a summary with the size of the graphs (in terms of number of nodes and edges) and with their structure is shown. It is interesting to see that all the graphs, even presenting very different size, are very well connected (%GCC between 0.87 and 0.99).

4. REFERER GRAPHS ANALYSIS

In this section we describe some analysis on these *RefererGraphs*, proving that they are consistently different not only in term of nodes and content but also in term of navigation patterns of the users. We also propose an experiment to understand how much the graphs are distinguishable.

4.1 Subgraphs comparison

We consider the seven subgraphs extracted from the main news portal graph with the procedure discussed in §3. Browsing patterns generated by different types of audiences, can lead to different pieces of news pages to emerge as the most central ones in the *BrowseGraph*. To check that, we ran the PageRank algorithm on each of the (weighted) subgraphs, and for every pair of subgraphs we compared the scores ob-

tained on their common nodes, using Kendall's τ distance. The intersection between the node sets of the networks is always large enough to allow us to compute the τ on the intersection only (> 1000 nodes in the case with less overlap). Kendall's τ will provide a clear measure of how much the importance of the same set of nodes varies among different subgraphs. When the ranking between two subgraphs differs greatly (*i.e.*, low Kendall's τ), it is an indication that they either show different content (*i.e.*, webpages) or that the collective browsing behaviour in the two graphs privileged different sets of pages.

Table 2 reports on the cross-distance among the subgraphs and also with respect to the full graph using Kendall's τ . Interestingly, most of the similarity values tend to be very low (< 0.3), confirming the hypothesis that the user's interests are tightly related to the domain where they come from. Some of these similarities, however, are considerably higher, remarkably the ones between the three subgraphs that are originated from search engines traffic, *i.e.*, Bing, Google and Yahoo, which yield the most similar rankings of pages (> 0.5). However, for the purpose of this work we expect to find a difference among the subgraphs in order to use them as local *BrowseGraph* and study the LRP with the full graph (*i.e.*, *BrowseGraph* made with the entire news log).

4.2 Random Surfer

In §4.1 we showed how users coming from different sources (*i.e.*, referrer domains) behave differently in terms of content discovery and, as a consequence, the importance of the news articles vary significantly among the different *BrowseGraphs*. It has been shown how the referrer domain might be extremely useful to characterize user sessions [12], to estimate user engagement [18] or to perform cold-start recommendation [27]. However, the user's referrer URL is not always visible and, in many cases, it is hidden or masked by services or clients. For instance, any Twitter or mail client (*i.e.*, third-party application) shows an empty referrer URL in the web logs. A similar situation happens with the widespread URL-shortening services (*e.g.*, Bitly.com), that mask the original Web page the user is coming from. Nonetheless, in all these cases, a provider could make use of her knowledge of the user's trail, to identify automatically the source where the user started her navigation in the local graph. As we have shown, the referrer URL might be useful to characterize the interest of the users, especially in the case where the users are unknown (*i.e.*, the user profile is not available). Thus, being able to identify the referrer URL when it is not available, is an advantage for the content provider. In this section we want to understand if it is feasible to detect the referrer URL of a user while he browses and how many browsing steps are required to be able to do so accurately. Moreover, if we find that the user sessions are easily distinguishable, it means that the subgraphs are different enough to be considered, in our experiment, as *local BrowseGraphs* of different service providers.

Therefore, we consider the following scenario: a content provider is observing a user surfing the pages of its web service, but it is unaware of the user's referrer URL. In terms of our experimental dataset, this scenario maps into the problem of observing a browsing trace left by a random surfer on one of the referrer-based subgraphs, and having to identify which graph it is. Intuitively, the larger the number of page visits (or *steps*) the surfer will make, the more distinc-

	Full	Facebook	Google	Bing	Yahoo	Reddit	Homepage	Twitter
Full	1.0000	0.1791	0.3931	0.3278	0.3548	0.0656	0.2797	0.0764
Facebook	0.1791	1.0000	0.3146	0.4111	0.3430	0.2616	0.4070	0.3026
Google	0.3931	0.3146	1.0000	0.5815	0.5860	0.1088	0.4217	0.1297
Bing	0.3278	0.4111	0.5815	1.0000	0.6624	0.1469	0.5238	0.1688
Yahoo	0.3548	0.3430	0.5860	0.6624	1.0000	0.1245	0.4632	0.1386
Reddit	0.0656	0.2616	0.1088	0.1469	0.1245	1.0000	0.1534	0.2309
Homepage	0.2797	0.4070	0.4217	0.5238	0.4632	0.1534	1.0000	0.1523
Twitter	0.0764	0.3026	0.1297	0.1688	0.1386	0.2309	0.1523	1.0000

Table 2: Kendall’s τ correlations between PageRank values ($\alpha = 0.85$) between the common nodes of the subgraphs.

Algorithm 1: `RandomSurfer(k, α , steps, G)`

```

logPr ← initialize vector with size  $G_k.length()$ ;
n ← total number of nodes;
 $x_j$  ← choose (random) starting node  $\in G_k$ ;
/* For each step, compute a random walk in  $G_k$ , and
compare the probability to be in all the other  $G$  */
for  $s \leftarrow 1$  to steps do
    /* Pick the next node of  $G_k$  with random walk */
     $x_k$  = next_node(  $G_k$ ,  $x_j$  );
    for  $i \leftarrow 0$  to  $G.length()$  do
         $\langle k_{out} \rangle$  ← get_outdegree( $n_p$ );
        if  $\langle k_{out} \rangle == 0$  then
            | logPr[ i ] ← logPr[ i ] + log(1/n);
        else
            |  $p_i(x) = (1 - \alpha)/n$ ;
            |  $Pd_{x_j}$  ← get_prob_distribution( $G_i$ ,  $x_j$ );
            |  $S_{x_j}$  ← get_successors( $G_i$ ,  $x_j$ );
            | if  $x_k \in S_{x_j}$  then
            |     |  $p_i(x) \leftarrow p_i(x) + \alpha * Pd_{x_j}(x_k)$ ;
            | logPr[ i ] ← logPr[ i ] + log( $p_i(x)$ );
    return logPr

```

tive its trace will be, and the easier the identification of the graph. Algorithm 1 shows the pseudocode that describes the process to compute the random surfer experiment.

Formally, observing the sequence of the surfer’s visited nodes $\mathbf{x} = (x_1, x_2, \dots, x_s)$ and computing the probability $p_i(\mathbf{x})$ that the surfer has gone through them given that it is surfing G_i , we need to deduce what is G_i (e.g., by maximum log-likelihood). With this goal in mind, we sort the indices of the subgraphs i_1, i_2, \dots so that $p_{i_1}(\mathbf{x}) \geq p_{i_2}(\mathbf{x}) \geq \dots$ and stop as soon as the gap between $\log p_{i_1}(\mathbf{x})$ and $\log p_{i_2}(\mathbf{x})$ is large enough (e.g., $\log p_{i_1}(\mathbf{x}) - \log p_{i_2}(\mathbf{x}) \geq \log 2$), with a maximum of 20 steps that we consider as a representation of a long user session.

In this set of experiments, we considered the seven URL-referral subgraphs G_1, \dots, G_7 , one at a time. For each subgraph G_i , we simulated a random surfer moving around in G_i (i.e., calling the function `RandomSurfer(i, α , steps, G)`), computing at each step (i.e., page visited) the probability of the surfer to navigate in each subgraph G_1, \dots, G_7 : we expect that the probability corresponding to G_i will increase at each step, and will eventually dominate all the others.

To estimate the number of steps required to identify cor-

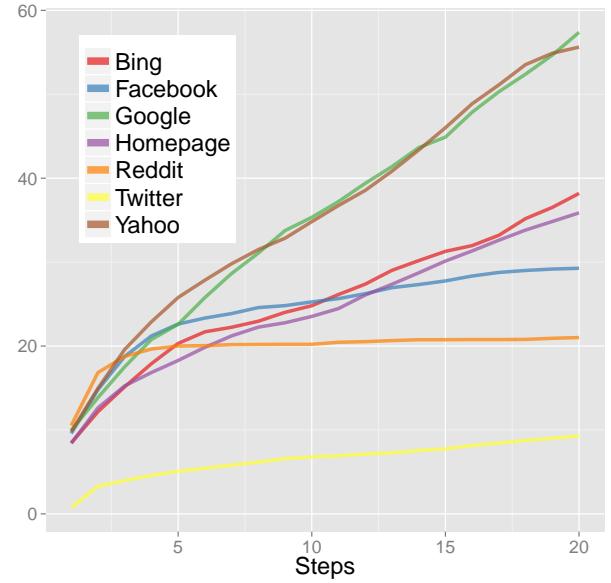


Figure 1: Random Surfer Experiment. On the y-axis: log-ratio of the probabilities (as explained in the text). X-axis: number of browsing steps performed by the surfer.

rectly the graph that the surfer is browsing, we measure the difference between log-probabilities for the correct graph G_i and for the graph with the largest log-probability among the other ones. As with PageRank we introduced a certain damping factor ($\alpha = 0.85$); this is necessary to avoid being stuck in terminal components of the graph. Recall that α is the balancing parameter that determines the probability of following in the random walk, instead of teleporting. The results are shown in Figure 1, averaged over 100 executions. The values on the y-axis represent the difference between the log-probabilities (i.e., the logarithm of their ratio): in general, we can observe that the very first steps are enough to understand correctly (and with a huge margin) in which graph the surfer is moving. The inset of Figure 1 displays the first 20 steps and the relative probability to identify the correct graph. Almost all the referrer domains are recognizable at the first step. This translates into a strong advantage for the service provider as it can identify from where the users are coming from, even if they use clients or services that masquerade it. With this information the service provider can personalize the content of the web pages for any users with respect to the referrer.

Interestingly, the plot reveals that some surfers are easier to single out than others; we read this as yet another confirmation that the subgraphs have a distinguished structural difference, or (if you prefer) that users have a markedly different behavior depending on where they come from. This experiment does not only show that it is possible to detect from which referrer domain the surfer is coming from, but that the graphs are quite different and that they can be used for our study.

5. PAGERANK ON THE BROWSEGGRAPH

Next, we study the convergence of the PageRank ranking between the *local BrowseGraphs* (*ReferrerGraphs*) and the full *BrowseGraph*. We want to understand how different are the ranking computed using less or more knowledge about the full graph. We present an experiment, called “Growing Balls”, that compute the distance between the rankings expanding at each step the known nodes (and edges) with the neighbors of the subgraphs.

5.1 “Growing Balls” Experiment

We first focus on the study of the *Local Ranking Problem* on browsing graphs. An important question related to this problem is how much the PageRank node values vary, when new nodes and edges are added to the local graph. A natural way to determine this is to expand incrementally the graph by adding new nodes and edges in a Breadth-First Search (BFS) fashion, and comparing the PageRank computed on the expanded graph with the one on the global graph.

More formally, given a graph H which is a subgraph of the full graph G , we simulate a growth sequence $H_0, H_1 \dots H_n$ in the following way:

- $H_0 \leftarrow H$;
- $V_{H_{k+1}} \leftarrow \{\Gamma_{out}(V_{H_k}) \cup V_{H_k}\}$, with V_x being the set of vertices of a graph, and Γ being the vertex neighborhood function;
- $E_{H_{k+1}} \leftarrow \{(v_1, v_2) | v_1 \in V_{H_{k+1}} \wedge v_2 \in V_{H_{k+1}}\}$, with E_x being the set of edges of a graph.

Using the standard graph terminology, we refer to the various steps of this expansion as “*balls*”, where the ball H_0 is the initial subgraph and subsequent balls are obtained by adding all the outgoing arcs that depart from the nodes in the current ball and end in nodes that are not in the ball. Observe that, depending on how it is built, H_0 may not be an induced subgraph of G , but H_1, \dots, H_n are always induced subgraphs, by definition of the expansion algorithm.

Using the Kendall’s τ function, we measure the difference between the local PageRank computed for each ball H_i , and the global PageRank computed on G . The main objective is to understand how much the ranking gets close the global one at each consecutive step, and whether the ranking values are able to converge even if we just consider a piece of the information contained in the whole graph.

To check the dependency of results from the initial graph selected, we consider three different sets of initial subgraphs, that we will study separately. We describe them next.

- **Referrer-based (RB).** The seven browsing subgraphs built by referrer URL: Facebook, Twitter, Reddit, Homepage, Yahoo, Google and Bing;

- **Same size referrer-based (SRB).** To measure how much the different sizes of the graphs impact on the observed behavior, we fix a number of nodes and extract a portion of each subgraph in order to obtain exactly the same size for all networks. The selection is performed with several attempts of BFS expansion, starting from a random node in each graph, until the resulting graphs have very similar size ($\pm 9.4\%$): other ways of selecting subgraphs would end up with disconnected samples, which of course would void the purpose of this experiment. With this procedure instead, we are able to compare the graphs on equal grounds and at the same time control for the effect of size (about $3K$ nodes and $20K$ edges).

- **Random (R).** To check whether the observed behavior has to do with the user behavior underlying the graph under examination (*e.g.*, the particular structure of the graph determined by the sessions of users coming from Twitter), we take a set of seven *random* graphs each of them reflecting the size of each of the referrer-based subgraphs. Thus, we can explore the behavior of browsing graphs, that preserve the size of the graphs originated by specific types of users, but that are “artificial” in the sense that destroy any connection with the behavior connected to a particular user class. To make sure that the size is the same, we start from a BFS exploration and we prune the last level to match exactly the size we need.

The results related to the **RB** case are shown in Figure 2 (left). The convergence happens relatively quickly, as the value τ approaches 1 in the first 3 iterations. The curves related to different subgraphs are shifted with respect to each other, apparently mainly due to their different size, the biggest networks starting from higher τ values and converging faster than the smaller ones. To determine the dependency on the graph size, we repeat the same experiment for the **SRB** case. The results for this case are shown in Figure 2 (center). Even if the curves resulted to be more flattened (confirming that the initial size has indeed a role in the convergence), we still observe noticeable differences between the curves for the first two expansion levels. This means that different subgraphs are substantially different from one another in terms of their structure: even after forcing them to have the same size, the convergence rates observed on the different graphs varies. At the first iteration, for instance, all the subgraphs in **SRB** have Kendall’s τ between 0.3 and 0.5, whereas the ones in **RB** are between 0.4 and 0.6. Moreover in **SRB** the biggest networks starting from higher τ values are not converging faster. This intuition is confirmed by repeating the experiment on graphs selected with the **R** strategy. Results, displayed in Figure 2 (right), show that convergence in this case is much slower and the difference between the curves is less prominent.

Summarizing, with the previous experiment, we show that the Growing Balls on random subgraphs behave differently, especially when considering the number of iterations required in order to converge.

5.2 Growing Balls with Selection of Nodes

Besides the selection of the initial graph, the rank convergence depends also on the way the growing balls are built at each iteration. How does the expansion influence conver-

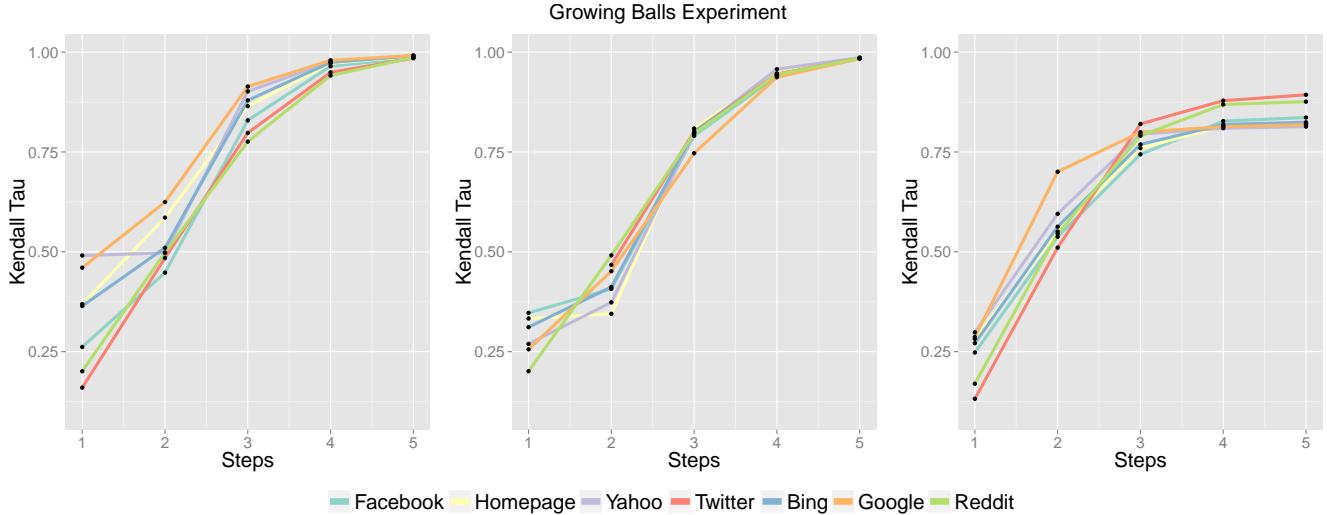


Figure 2: Growing Balls experiment on: (left) original subgraphs built based on the referrer URL, (center) seven subsubgraphs with very similar size, (right) eight subgraphs random selected from the full graph, where each of them has the same size of one of the original.

gence if only few more representative nodes are selected? To what extent a higher *volume* of selected nodes helps a quicker convergence or adds more *noise*? At a first glance, one may argue that using all the nodes is equivalent to injecting all the available information, so the convergence to the values of PageRank computed on the full graph G should be faster. On the other hand, instead, one may observe that we are introducing a huge number of nodes in each iteration (as the growth is at each step larger), adding also the ones that are less important and this can induce an incorrect PageRank for some time, until all the graph becomes known. In order to shed light on this aspect, we introduce a variant in the growing-balls expansion algorithm, and we select only the nodes with highest PageRank.

More formally, considering H_k as the subgraph at iteration k and V_{H_k} its set of nodes, we select all the external nodes in $Y = \{V_G \setminus V_{H_k}\}$, that are connected through outgoing arcs from the nodes in V_{H_k} . We then compute the PageRank values on the subgraph H_k extended with the nodes Y , and obtain a ranked list of nodes. Among all the nodes in Y we select the top $n\%$ with largest PageRank value, and only those ones will be added to H_k in order to build H_{k+1} and advance to the next iteration.

We conducted experiments with this partial expansion at different percentages: 5%, 10%, 30%, 50%, and 100%, and then we computed the average Kendall's τ value for each one of the percentages. The results are shown in Figure 3. Remarkably, the figure highlights how expanding the graph by adding fewer nodes, although the most representative ones, leads to PageRank values that are closer to the *global* ones in the first iterations. Since we are expanding the local graph with a small (highly-central) number of nodes, we could argue that they initially help to boost the local PageRank scores. However, given that we keep on expanding using a few nodes at each iteration, the nodes that have not been added before exclude a large number of nodes among which there might also be highly central ones. This might explain why in the first iteration(s) the convergence rate is

high, but on the limit the final convergence values result in a low Kendall's τ . Contrarily, in the long run, expansions that include the highest number of nodes present convergence values closer to 1. This is somehow expected, given that at each iteration any subgraph H closer in size to the full graph G will include almost every node and arc.

Nonetheless, the main significant outcome of this experiment is that it is possible to obtain a yet satisfactory PageRank convergence, with few but very representative nodes. For situations in which including additional pieces of information, in terms of node/arc insertions, implies a non-negligible cost, requesting just a little amount of well-selected information allows to obtain good approximations while minimizing the costs.

6. PAGERANK PREDICTION

In the previous section we have shown how the approximation to the global PageRank varies with the expansion of the initial subgraph. The ranking of the nodes converges quite fast on all the subgraphs: they differ in terms of their content, although they are similar in terms of structure in that all of them are built based on users' navigational patterns. Building upon the findings about how local and global PageRank computed on the *BrowseGraphs* relate to each other, we designed an experiment to assess how well a learned model could perform in predicting this relationship.

We address the problem of predicting the Kendall's τ between the local and the global PageRank, only considering information available on the local graph such as topological features. This is an extremely common situation given that, in general, the information pertaining the local graph is the only one that is readily available, and usually of a limited size. Computing this distance accurately has a high value for service providers, since it translates directly into an estimation of the reliability of the PageRank scores computed on their local subgraphs. As a direct consequence one can apply, with different levels of confidence, methods for optimizing web sites [31], studying user en-

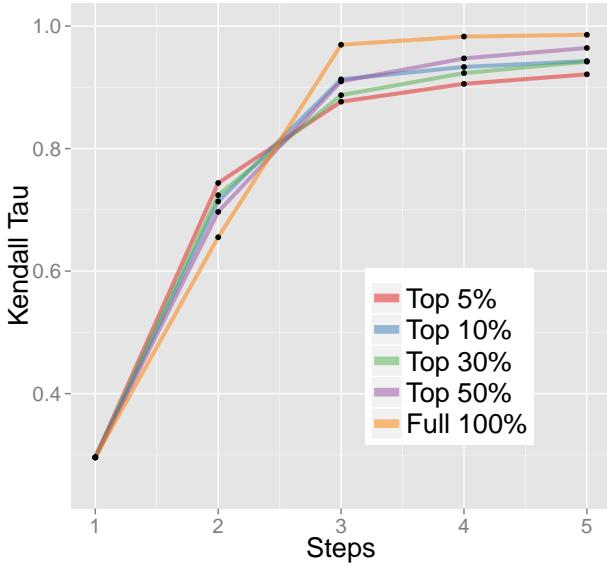


Figure 3: Growing Balls using only the nodes with highest PageRank. The plot shows the average values of the Kendall- τ at each step computed for all the subgraph.

gagement [18], characterizing user’s session [12] or content recommendation [27].

6.1 Prediction of Kendall τ Distance

We have seen that the deviation of the local PageRank with respect to the global one can be relevant, depending on factors such as the size of the local graph and the different behavior of the users who browse it (see §5.1 and particularly Figure 2). Recall that we compute the distance comparing the rankings with Kendall’s τ , since we are interested in obtaining a ranking as close as possible to the one computed on the entire graph. Although we have previously shown how to expand the view on the local graphs with nodes residing at the border, this practice might not always be possible in a real-world scenario, since service providers often can have access only to the browsing data *within* their domain.

Previous work on local ranking on graphs raised several questions related to this scenario, highlighting practical applications of the local rank estimation not only for web pages but also in social networks [9]. Critically, so far it is not clear whether there are some topological properties of the local graph that make the local ranking problem easier or harder, and if these properties can be exploited by local algorithms to improve the quality of the local ranking. We explore this research direction by studying a fundamental aspect that is at the base of the open questions in this area, namely the possibility of estimating the deviation of the local PageRank from the global one, using the structural information of the local network. The intuition is that, some structural properties of the graph could be good proxies for the τ value difference, computed between local and global ranks. Being able to estimate the Kendall’s τ distance between the subgraph available to the service provider and the global graph, implies the ability to estimate the reliability of the current ranking using only information of the local subgraph.

To verify this hypothesis we resort to regression analysis. Starting from the seven subgraphs in the dataset, we build a training set using the jackknife approach, by removing nodes in bulks (1%, 5%, 10%, 20%) and computing the τ value between the full subgraph and their reduced versions. Then, for each instance in the training set, we compute 62 structural graph metrics [30, 4] belonging to the following categories:

- **Size and connectivity (S).** Statistics on the size and basic wiring properties, such as number of nodes and edges, graph density, reciprocity, number of connected components, relative size of the biggest component.
- **Assortativity (A).** The tendency of node with a certain degree, to be linked with nodes with similar degree. We computed different combinations that take into account the in/out/full degree of the target node vs. the in/out/full degree of the nodes that are connected with it.
- **Degree (D).** Statistics (average, median, standard deviation, *etc.*) on the degree distribution of nodes.
- **Weighted degree (W).** Same as **degree**, but considering the weight on edges, that usually referred as node strength. As the edges are the transitions made by the users during the navigation, the weight stand for the number of times the users have navigated the transition.
- **Local Pagerank (P).** Statistics on the distribution of the PageRank values computed on the local graph.
- **Closeness centralization (C).** Statistics on the distances (number of hops), that separate a node to the others in the graph, in the spirit of the closeness centralization [30].

We employed different regression algorithms, although we report the performance using random forests [7], which performed better in this scenario than other approaches like support vector regression [25]. We computed the mean square error (MSE) across all examples in all sampled subgraphs. The random forest regression has been computed over a five-fold cross validation averaged over 10 iterations. The mean square residuals that we obtained is very low, around $2.4 \cdot 10^{-6}$. Results, computed for the full set of features and for each category separately, are given in Table 3. The most predictive feature category is the *weighted degree*, which yields a performance that is better (or comparable) than the model using all the features, whereas the *assortativity* features seem to be the ones that have the less predictive power on their own. This might be due to the fact the model with 62 features is too complex for the amount of training data available. On the other hand, the *weighted degree* that is the best performing class of features, contains the statistics of the degree distribution on the weighted edges. In Figure 4 the features included in *weighted degree* are ranked by their discriminative power in predicting the Kendall τ distance using the permutation test proposed by Strobl *et al.* [26]. These features, which are based on the distribution of the out- and in-degree of the nodes, are straightforward to compute from the local graph—a very affordable task for service providers.

Feature Class	No. Features	MSE
weighted degree	15	$2.2 \cdot 10^{-6}$
degree	15	$2.9 \cdot 10^{-6}$
local PageRank	10	$3.3 \cdot 10^{-6}$
size and connectivity	9	$3.4 \cdot 10^{-6}$
closeness	5	$4.1 \cdot 10^{-6}$
assortativity	8	$9.3 \cdot 10^{-6}$
ALL features	62	$2.4 \cdot 10^{-6}$

Table 3: MSE of cross validation. Average differences are statistically significant with respect to *weighted degree* and *ALL features* (t-test, $p < 0.01$).

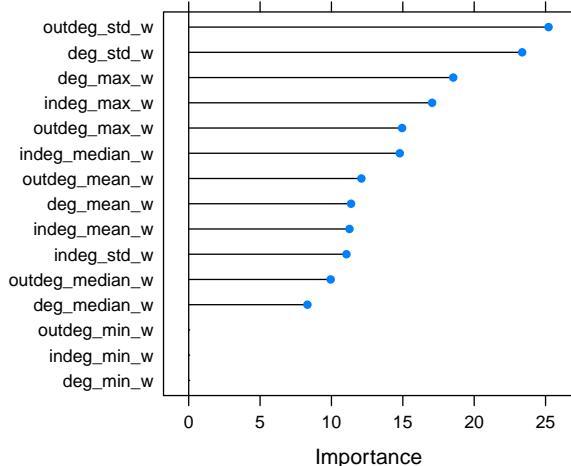


Figure 4: The 15 features of *weighted degree*, the most predictive class, sorted by importance. Note that some of them do not have any contribution to the Kendall- τ prediction, therefore just few features are necessary in order to estimate the distance.

We then use the learned model to predict the τ values of the seven subgraphs. When we applied the predictive models learned in the subsamples to regressing the full graphs, the MSE, is less than 0.026 on average, which, even if relatively low, it is higher than the cross-validated performance in the sub-samples. However, the model was able to rank the seven different subgraphs by their Kendall's τ almost perfectly. When using all the features the Spearman's correlation coefficient between the true order and the predicted one is 0.85 (high correlation), and when we used the most predictive features (weighted degree) the correlation was as high as 0.80 (moderate high correlation). Overall, the final rankings are just one swap away (Kendall's τ is over 0.70 in this case). This kind of information can be very helpful when comparing different local sub-domains to determine which one has pages that better estimate the global PageRank.

7. CONCLUSION

In this paper we tackled the *Local Ranking Problem*, i.e., how to estimate the PageRank values of nodes when a portion of the graph is not available, which arises commonly in

real use cases of PageRank. We investigated this problem for a novel environment, namely estimating PageRank on a large user-generated browsing graph from a large news provider. The peculiar characteristic of this graph is that it is built from user's navigation patterns, where nodes represent web pages and edges are the transitions made by the users themselves. Moreover, the information about the domain of origin of the users (namely the referrer URL of their sessions), is also available.

We built a set of *ReferrerGraphs* including the browsing subgraphs based on different referrer URLs, and then we studied their difference in terms of user navigation patterns. We found that all of the browsing patterns initiated from different domains exhibit remarkable differences in terms of which pages users visited next. The referrer URL (or domain) has been found to be extremely useful for characterizing the user behavior [12] or for recommendation of content [27]. With this observation in mind and motivated by the cases where the domain from where the user is coming is not available, such as Facebook and Twitter clients or URL shortening services, we performed a series of experiments with the aim of predicting from which referrer URL the user joined the network, i.e., if a model can predict reliably where the user is entering our network. In general, just a few steps (i.e., visited pages) suffice to recognize the referrer URL correctly because the surfing behavior is very distinctive of the domain the user is coming from.

Then, using the *ReferrerGraphs*, we performed several experiments using a very large network of sites (with almost two billions of user transitions) to assess to what extent the browsing patterns information can be generalized, if one is only provided with information from smaller subgraphs. First, we computed the PageRank of the subgraphs and on their step-by-step BFS expansion, measuring the distance in terms of Kendall's τ with the PageRank computed on the full graph. To control for the subgraph size and type, and to study the impact of the expansion strategy on the PageRank convergence, we used two flavors of BFS and three different sets of initial subgraphs. We found that expanding the local graph with few nodes of largest value of PageRank leads to a faster (74% at the first expansion step), although less accurate convergence in the long run. On the other hand, adding more nodes lead to a slower converge rate in the first steps (65%). Therefore, in all the cases where a strong convergence with the values of the global PageRank is not required, selecting few specific nodes is enough to significantly improve the PageRank values of the local nodes, without having to request and process a larger amount of data.

Finally, we considered the case of a service provider that wants to estimate the reliability of the scores of PageRank computed on its local *BrowseGraph*, with respect to the ones computed on the global graph. Therefore, we performed another experiment trying to predict the value of the Kendall's τ between the local and the global PageRank, only considering information available on the local graph. We explored six different sets of topological and structural features of the browse graph, namely size and connectivity, assortativity, degree, weighted degree, local PageRank and closeness. Then we computed those features on a training set that we obtained by applying a jackknife sampling of our subgraphs, and we ran a regression on the Kendall's τ of the PageRank of the full subgraph and the various samplings.

We found that a random forest ensemble built on *weighted degree*, outperforms all the other in terms of mean square error. When applying the regression to the task of predicting the τ value of the global graph with the eight subgraphs at hand, we were able to reproduce quite well the ranking of their estimated τ values with their actual ranking, up to a Spearman's coefficient of 0.8.

Future Work. We envision different routes worth being taken into consideration for future work. One line of research we plan to investigate deals with the problem of user browsing prediction. In other words, what extent it may be possible to identify what are the most common patterns of topical behavior in the network and also, to build per-user browsing models to predict what would be the page to be visited next. Further, motivated by real use case scenarios, we considered subgraphs determined by the referrer URL of user sessions; we believe that interesting analytical results could be found, when considering other types of subgraphs, such as networks induced by nodes that belong to the same topical area.

8. ACKNOWLEDGMENTS

This work was partially funded by Grant TIN2009-14560-C03-01 of the Ministry of Science and Innovation of Spain, by the EU-FET grant NADINE (GA 288956) and by a Yahoo Faculty Research Engagement Program.

9. REFERENCES

- [1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR*, pages 19–26, New York, NY, USA, 2006. ACM.
- [2] R. Andersen, C. Borgs, J. Chayes, J. Hopcroft, V. S. Mirrokni, and S.-H. Teng. Local computation of pagerank contributions. In *WAW*, pages 150–165, San Diego, CA, USA, 2007. Springer-Verlag.
- [3] Z. Bar-Yossef and L.-T. Mashiach. Local approximation of pagerank and reverse pagerank. In *CIKM*, pages 279–288, Napa Valley, California, USA, 2008. ACM Press.
- [4] A. Barrat, M. Barthélémy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, New York, NY, USA, 2008.
- [5] K. Bharat, A. Broder, M. Henzinger, P. Kumar, and S. Venkatasubramanian. The connectivity server: fast access to linkage information on the web. In *WWW*, volume 30, pages 469–477, Brisbane, Australia, 4 1998. Elsevier Science Publishers B. V.
- [6] P. Boldi, M. Santini, and S. Vigna. Do your worst to make the best : Paradoxical effects in pagerank incremental computations. In *WAW*, pages 168–180. Springer, 2004.
- [7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, oct 2001.
- [8] M. Bressan, E. Peserico, U. Padova, and L. Pretto. The power of local information in pagerank. In *WWW Companion*, pages 179–180, Rio de Janeiro, Brazil, 2013.
- [9] M. Bressan and L. Pretto. Local computation of pagerank: the ranking side. In *CIKM*, pages 631–640. ACM, 2011.
- [10] Y.-Y. Chen, Q. Gan, and T. Suel. Local methods for estimating pagerank values. In *CIKM*, pages 381–389, New York, NY, USA, 2004. ACM.
- [11] L. Chiarandini, P. Grabowicz, M. Trevisiol, and A. Jaimes. Leveraging browsing patterns for topic discovery and photostream recommendation. In *ICWSM*, Cambridge, MA, USA, 2013. AAAI.
- [12] L. Chiarandini, M. Trevisiol, and A. Jaimes. Discovering social photo navigation patterns. In *ICME*, pages 31–36. IEEE, 2012.
- [13] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through url ordering. In *WWW*, volume 30, pages 161–172, Brisbane, Australia, 4 1998. Elsevier Science Publishers B. V.
- [14] J. V. Davis and I. S. Dhillon. Large scale analysis of web revisit patterns. In *KDD*, volume 08, pages 116–125, Philadelphia, PA, USA, 2006. ACM Press.
- [15] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *VLDB*, pages 576–587, Toronto, ON, Canada, 2004.
- [16] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *WWW*, pages 151–160, New York, NY, USA, 2007. ACM.
- [17] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
- [18] J. Lehmann, M. Lalmas, and R. Baeza-Yates. Measuring inter-site engagement. In *Big Data, 2013 IEEE International Conference on*, pages 228–236. IEEE, 2014.
- [19] R. Lempel and S. Moran. Salsa : The stochastic approach for link- structure analysis. *Challenge*, 19(2):131–160, 2001.
- [20] J. Liu, P. Dolan, and E. R. Pedersen. Personalized news recommendation based on click behavior. In *IUI*, pages 31–40, New York, NY, USA, 2010. ACM.
- [21] M. Liu, R. Cai, M. Zhang, and L. Zhang. User browsing behavior-driven web crawling. In *CIKM*, pages 87–92, New York, NY, USA, 2011. ACM.
- [22] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li. Browserank: letting web users vote for page importance. *SIGIR*, 31:451–458, 2008.
- [23] Y. Liu, T.-Y. Liu, B. Gao, Z. Ma, and H. Li. A framework to compute page importance based on user behaviors. *Information Retrieval*, 13(1):22–45, 6 2009.
- [24] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *World Wide Web Internet And Web Information Systems*, 54(2):1–17, 1998.
- [25] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. Technical report, Statistics and Computing, 2003.
- [26] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, 2008.
- [27] M. Trevisiol, L. M. Aiello, R. Schifanella, and A. Jaimes. Cold-start news recommendation with domain-dependent browse graph. In *RecSys*, Foster City, CA, 2014. ACM.
- [28] M. Trevisiol, L. Chiarandini, L. M. Aiello, and A. Jaimes. Image ranking based on user browsing behavior. In *SIGIR*, pages 445–454, New York, NY, USA, 2012. ACM.
- [29] M. Tsagkias and R. Blanco. Language intent models for inferring user browsing behavior. In *SIGIR*, pages 335–344, New York, NY, USA, 2012. ACM.
- [30] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
- [31] B. Weischedel and E. K. R. E. Huizingh. Website optimization with web metrics: A case study. In *ICEC*, pages 463–470, New York, NY, USA, 2006. ACM.
- [32] R. W. White. Investigating behavioral variability in web search. In *In Proc. WWW*, pages 21–30, 2007.
- [33] R. W. White and J. Huang. Assessing the scenic route: measuring the value of search trails in web logs. In *SIGIR*, pages 587–594, New York, USA, 2010. ACM.