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 X 2nd**

Period covered: **from 1.5.2012** **to 31.10.2013**

Name, title and organisation of the scientific representative of the project's coordinator[1]:

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 D1.1.

It is based on milestones M1, M2, M5 with deliverable publications

[1] P1.1 L.Ermann, A.D.Chepelianskii and D.L.Shepelyansky, **"Towards two-dimensional search engines"**, J. Phys. A: Math. Theor. v.45, p.275101 (2012) (arXiv:1106.6215[cs.IR])

[5] P1.5 K.M.Frahm and D.L. Shepelyansky **"Google matrix of Twitter"**, Eur. Phys. J. B v.85, p.355 (2012) (arXiv:1207.3414[cs.SI], 2012

[8] P1.8 Y.-H.Eom, K.M.Frahm, A.Benczur and D.L. Shepelyansky, **"Time evolution of Wikipedia network ranking"**, submitted Eur. Phys. J. B (2013) (arXiv:1304.6601 [physics.soc-ph], 2013)

[10] P1.10 Y.-H.Eom and D.L. Shepelyansky, **"Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles"**, PLoS ONE v.8(10), p.e74554 (2013) (arXiv:1306.6259 [cs.SI], 2013)

[13] P2.1 N. Litvak, and R. van der Hofstad, **"Uncovering disassortativity in large scale-free networks"**, Phys. Rev. E, 87(2), p. 022801 (2013) (arXiv:1204.0266v3[physics.soc-ph], 2012

[14] P2.2 N. Litvak, and R. van der Hofstad, **"Degree-degree correlations in random graphs with heavy-tailed degrees"**, to appear in Internet mathematics (2013) (arXiv:1202.3071[math.PR])

[15] P2.3 K.Avrachenkov, N. Litvak, M. Sokol, and D. Towsley, **"Quick detection of nodes with large degrees"**, In: 9th International Workshop on Algorithms and Models for the Web Graph, WAW 2012, 22-23 June 2012, Halifax, NS, Canada. pp.54-65. Lecture Notes in Computer Science 7323, Springer Verlag (2012). Extended version accepted in Internet Mathematics

[16] P2.4 K.Avrachenkov, N. Litvak, V.Medyanikov, M. Sokol, **"Alpha current flow betweenness centrality"**, Conference proceedings: 10th Workshop on Algorithms and Models for the Web Graph, WAW2013, 15-16 December, 2013, Harvard University (arXiv:1308.2591v1 [cs.SI], 2013)

[17] P2.5 L.Ostroumova, K.Avrachenkov, , N. Litvak, **"Quick detection of popular entities in large directed networks"**, submitted to Computer Science Conference, Oct 2013

[18] P2.6 P. van der Hoorn, N. Litvak, **"Degree-degree correlations in directed networks with heavy-tailed degrees"**, submitted Oct arXiv:1310.6528[math.PR] (2013)

[31] P4.7 P.Boldi and S.Vigna. **"In-core computation of geometric centralities with HyperBall: a hundred billion nodes and beyond"** to appear in the Proceedings of 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW 2013), (arXiv:1308.2144, 2013)

Toward two-dimensional search engines

2012 J. Phys. A: Math. Theor. 45 275101

(http://iopscience.iop.org/1751-8121/45/27/275101)

J. Phys. A: Math. Theor. **45** (2012) 275101 (20pp) doi:10.1088/1751-8113/45/27/275101

Toward two-dimensional search engines

L Ermann[1], **A D Chepelianskii**[2] **and D L Shepelyansky**[1]

[1] Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France
[2] Department of Physics, Cavendish Laboratory, University of Cambridge, CB3 0HE, UK

E-mail: dima@irsamc.ups-tlse.fr

Received 27 September 2011, in final form 17 May 2012
Published 18 June 2012
Online at stacks.iop.org/JPhysA/45/275101

Abstract
We study the statistical properties of various directed networks using ranking of their nodes based on the dominant vectors of the Google matrix known as PageRank and CheiRank. On average PageRank orders nodes proportionally to a number of ingoing links, while CheiRank orders nodes proportionally to a number of outgoing links. In this way, the ranking of nodes becomes two dimensional which paves the way for the development of two-dimensional search engines of a new type. Statistical properties of information flow on the PageRank–CheiRank plane are analyzed for networks of British, French and Italian universities, Wikipedia, Linux Kernel, gene regulation and other networks. A special emphasis is done for British universities networks using the large database publicly available in the UK. Methods of spam links control are also analyzed.

PACS numbers: 89.75.Fb, 89.75.Hc, 89.20.Hh

(Some figures may appear in colour only in the online journal)

1. Introduction

During the past decade, modern society has developed enormously large communication networks. The well-known example is the World Wide Web (WWW) which has started approaching 10^{11} webpages [1]. The sizes of social networks like Facebook [2] and VKONTAKTE [3] have also become enormously large, reaching 600 and 100 millions user pages, respectively. The information retrieval from such huge databases becomes the foundation and main challenge for search engines [4, 5]. The fundamental basis of the Google search engine is the PageRank algorithm [6]. This algorithm ranks all websites in a decreasing order of components of the PageRank vector (see e.g. detailed description at [7], historical surveys of PageRank are given at [8, 9]). This vector is a right eigenvector of the Google matrix at the unit eigenvalue, it is constructed on the basis of the adjacency matrix of the directed network, its components give a probability of finding a random surfer on a given node.

The Google matrix G of a directed network with N nodes is given by

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

where the matrix S is obtained by normalizing to unity all columns of the adjacency matrix $A_{i,j}$, and replacing columns with zero elements by $1/N$. An element A_{ij} of the adjacency matrix is equal to unity, if a node j points to node i and zero otherwise. The damping parameter α in the WWW context describes the probability $(1 - \alpha)$ for a random surfer to jump to any node. The value $\alpha = 0.85$ gives a good classification for the WWW [7] and thus we also use this value here. A few examples of Google matrix for various directed networks are shown in figure 1. The matrix G belongs to the class of Perron–Frobenius operators [7], its largest eigenvalue is $\lambda = 1$ and other eigenvalues have $|\lambda| \leqslant \alpha$. The right eigenvector at $\lambda = 1$ gives the probability $P(i)$ to find a random surfer at site i and is called the PageRank. Once the PageRank is found, all nodes can be sorted by decreasing probabilities $P(i)$. The node rank is then given by index $K(i)$ which reflects the relevance of the node i. The PageRank dependence on K is well described by a power law $P(K) \propto 1/K^{\beta_{\rm in}}$ with $\beta_{\rm in} \approx 0.9$. This is consistent with the relation $\beta_{\rm in} = 1/(\mu_{\rm in} - 1)$ corresponding to the average proportionality of PageRank probability $P(i)$ to its in-degree distribution $w_{in}(k) \propto 1/k^{\mu_{\rm in}}$, where $k(i)$ is a number of ingoing links for a node i [7, 10]. For the WWW, it is established that for the ingoing links $\mu_{\rm in} \approx 2.1$ (with $\beta_{\rm in} \approx 0.9$) while for the out-degree distribution $w_{\rm out}$ of outgoing links a power law has the exponent $\mu_{\rm out} \approx 2.7$ [11, 12]. Similar values of these exponents are found for the WWW British university networks [13], the procedure call network (PCN) of Linux Kernel software introduced in [14] and for Wikipedia hyperlink citation network of English articles (see e.g. [15]).

The PageRank gives at the top the most known and popular nodes. However, an example of the Linux PCN studied in [14] shows that in this case the PageRank puts at the top certain procedures which are not very important from the software view point (e.g. *printk*). As a result it was proposed [14] to use in addition another ranking taking the network with inverse link directions in the adjacency matrix corresponding to $A_{ij} \to A^T = A_{ji}$ and constructing from it an additional Google matrix G^* according to relation (1) at the same α. The eigenvector of G^* with eigenvalue $\lambda = 1$ then gives a new inverse PageRank $P^*(i)$ with ranking index $K^*(i)$. This ranking was named CheiRank [15] to mark that it allows us to *chercher l'information* in a new way (which in English means *search the information* in a new way). Indeed, for the Linux PCN the CheiRank gives at the top more interesting and important procedures compared to the PageRank [14] (e.g. *start_kernel*). While the PageRank ranks the network nodes in average proportionally to a number of ingoing links, the CheiRank ranks nodes in average proportionally to a number of outgoing links. The physical meaning of PageRank vector components is that they give the probability to find a random surfer on a given node when a surfer follows the given directions of network links. In a similar way, the CheiRank vector components give the probability to find a random surfer on a given node when a surfer follows the inverted directions of network links. The inversion of links is a mathematical way to give a weight to outgoing links. We note that each directed network has both outgoing and ingoing links, and thus it is important to characterize these two complementary properties of information flow on directed networks. Since each node belongs both to CheiRank and PageRank vectors, the ranking of information flow on a directed network becomes two dimensional. We note that there have been earlier studies of PageRank of the Google matrix with inverted directions of links [16, 17], but no systematic analysis of statistical properties of 2DRanking was presented there.

An example of variation of PageRank probability $P(K)$ with K and CheiRank probability $P^*(K^*)$ with K^* is shown in figure 2, for the WWW network of University of Cambridge in

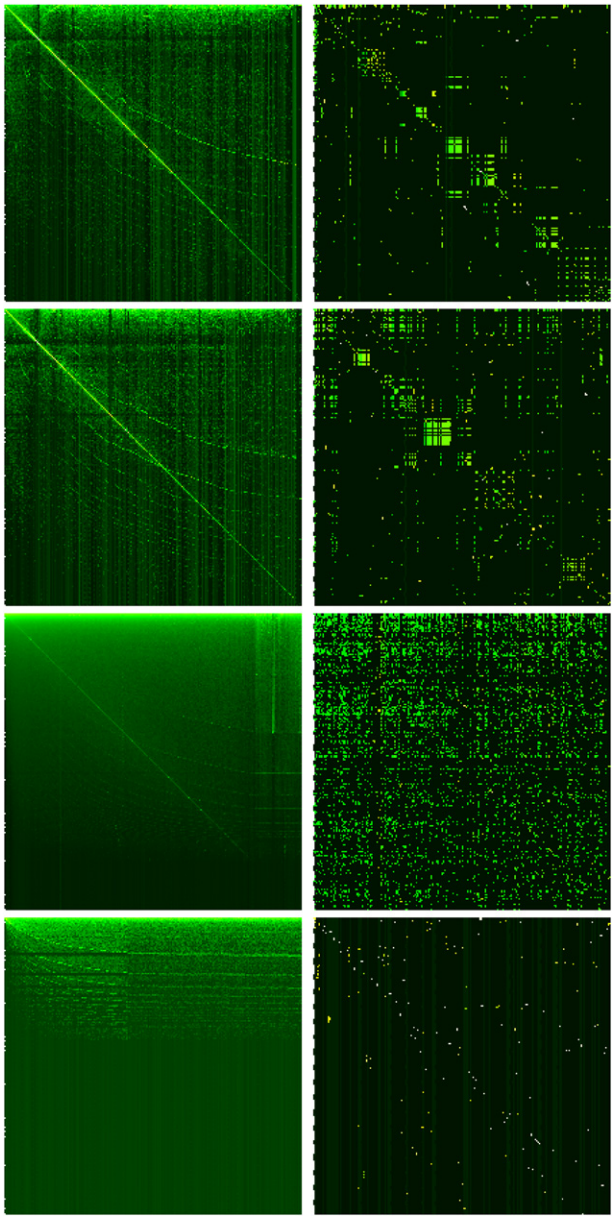


Figure 1. Google matrix gallery: all matrices are shown in the basis of PageRank index K (and K') of matrix $G_{KK'}$, which corresponds to x (and y) axis with $1 \leqslant K, K' \leqslant N$ (left column) and $1 \leqslant K, K' \leqslant 200$ (right column); all nodes are ordered by PageRank index K of matrix G and thus we have two matrix indexes K, K' for matrix elements in this basis. Left column: coarse-grained density of Google matrix elements $G_{K,K'}$ written in the PageRank basis $K(i)$ with indexes $j \to K(i)$ (in x-axis) and $i \to K'(i)$ (in a usual matrix representation with $K = K' = 1$ on the top-left corner); the coarse graining is done on 500×500 square cells for the networks of University of Cambridge 2006, University of Oxford 2006, Wikipedia English articles, PCN of Linux Kernel V2.6 (from top to bottom). Right column shows the first 200×200 matrix elements of G matrix at $\alpha = 0.85$ without coarse graining with the same order of panels as in the left column. Color shows the density of matrix elements changing from black for minimum value $((1 - \alpha)/N$ to white for maximum value via green and yellow (density is coarse grained in the left column).

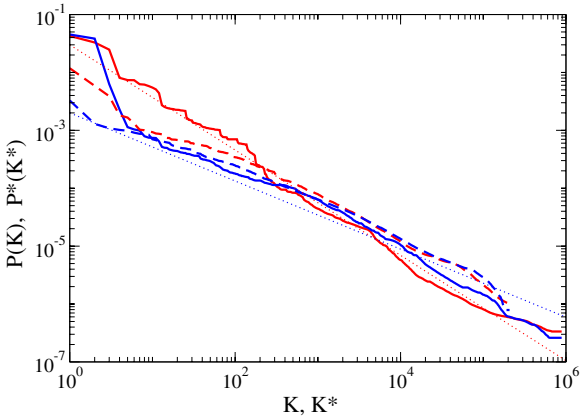


Figure 2. Dependence of probabilities of PageRank $P(K)$ (red/gray curve) and CheiRank $P^*(K^*)$ (blue/black curve) on corresponding ranks K and K^* for the network of University of Cambridge in 2006 (dashed curve) and in 2011 (full curve). The power-law dependences with the exponents $\beta \approx 0.91, 0.59$, corresponding to the relation $\beta = 1/(\mu - 1)$ with $\mu = 2.1, 2.7$, respectively, are shown by dotted straight lines.

years 2006 and 2011. Other examples for PCN Linux Kernel and Wikipedia can be found in [14, 15]. Detailed parameters of networks which we analyze in this paper and their sources are given in the appendix.

A detailed comparative analysis of PageRank and CheiRank two-dimensional classification was done in [15] for the example of the Wikipedia hyperlink citation network of English articles. It was shown that CheiRank highlights communicative property of nodes leading to a new way of two-dimensional ranking. While according to PageRank the top three countries are (1) USA, (2) UK and (3) France, CheiRank gives (1) India, (2) Singapore and (3) Pakistan as the most communicative Wikipedia country articles. The top 100 personalities of PageRank have the following percents in five main category activities: 58 (politics), 10 (religion), 17 (arts), 15 (science) and 0 (sport) [15]. Clearly, the significance of politicians is overestimated (many of them are USA presidents not broadly known to public). In contrast, CheiRank gives a more balanced distribution over these categories with 15, 1, 52, 16 and 16, respectively. It allows us to classify information in a new way finding composers, architects, botanists and astronomers who are not well known but who, for example, discovered a lot of Australian butterflies (*George Lyell*) or many asteroids (*Nikolai Chernykh*). These two people appear in the large listings of Australian butterflies and in the listing of asteroids (since they discovered many of them) and due to that they gain high CheiRank values. In a similar way, popular singers and musicians have long listings of their songs and music which increase their outgoing links and CheiRank. This shows that the information retrieval, which uses both PageRank and CheiRank, allows us to rank nodes not only by an amount of their popularity (how known is a given node) but also by an amount of their communicative property (how communicative is a given node). This 2DRanking was also applied to the brain model of the neuronal network [18] and the business process management network [19], and it was shown that it gives a new useful way of information treatment in these networks. The 2DRanking in the PageRank–CheiRank plane also naturally appears for the world trade network corresponding to import and export trade flows [20]. Thus, the 2DRanking based on PageRank and CheiRank paves the way for the development of 2D search engines which can become more intelligent than the present Google search based on the 1D PageRank algorithm.

In this work, we study the statistical properties of such a 2DRanking using examples of various real directed networks, including the WWW of British, French and Italian university networks [21], Wikipedia network [15], Linux Kernel networks [14, 22], gene regulation networks [23, 24] and other networks. The rest of the paper is organized as follows: in section 2, we study the properties of node density in the plane of PageRank and CheiRank; in section 3, the correlator properties between PageRank and CheiRank vectors are analyzed for various networks; information flow on the plane of PageRank and CheiRank is analyzed in section 4; the methods of control of SPAM outgoing links are discussed in section 5; 2DRanking applications for the gene regulation networks are considered in section 6 and the discussion of results is presented in section 7. The parameters of the networks and references on their sources are given in the appendix.

2. Node density of 2DRanking

A few examples of the Google matrix for four directed networks are shown in figure 1. There is a significant similarity in the global structure of G for the Universities of Cambridge and Oxford with well-visible hyperbolic curves (left column) even if at small scales the matrix elements are rather different (right column) in these two networks (see figure 1). Such hyperbolic curves are also visible in the Google matrix of Wikipedia (left column) even if here they are less pronounced due to much larger averaging inside the cells which contain about 15 times larger number of nodes (see network parameters in the appendix). We make a conjecture that the appearance of such curves is related to the existence of certain natural categories existing in the network, e.g., departments for universities or countries, cities, personalities etc for Wikipedia. We expect that there are relatively more links inside a given category compared to links between categories. However, this is only a statistical property, since on small scales at small K values the hyperbolic curves are not visible (right column in figure 1). Hence, more detailed studies are required to verify this conjecture. At small scale, the G matrix of Wikipedia is much more dense compared to the cases of Cambridge and Oxford (right column). We attribute such an increase of density of significant matrix elements to a stronger connectivity between nodes with large K in Wikipedia compared to the case of universities where the links have a more hierarchical structure. Partially this increase of density can be attributed to a larger number of links per node in the case of Wikipedia, but this increase by a factor 2.1 is not so strong and cannot explain all the differences of densities at small K scale. For Wikipedia, there are about 20% of nodes at the bottom of the matrix where there are almost no links. For PCN of Linux Kernel, this fraction becomes significantly larger with about 60% of nodes. The hyperbolic curves are still well visible for Linux PCN inside the remaining 40% of nodes. On a small scale, the density of matrix elements for Linux is rather small compared to the three previous cases. We attribute this to a much smaller number of links per node which is by factor 5 smaller for Linux compared to the university networks of figure 1 (see data in appendix).

The distributions of density of nodes $W(K, K^*) = dN_i/dKdK^*$ in the plane of PageRank and CheiRank in the logscale are shown for four networks of British universities in figure 3. Here, dN_i is a number of nodes in a cell of size $dKdK^*$ (see the detailed description in [15]). Even if the coarse-grained G matrices for Cambridge and Oxford look rather similar the density distributions in the (K, K^*) plane are rather different, at least at moderate values of K, K^*. The density distributions for all four universities clearly show 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 such a line should correspond to significant correlations between $P(K(i))$ and

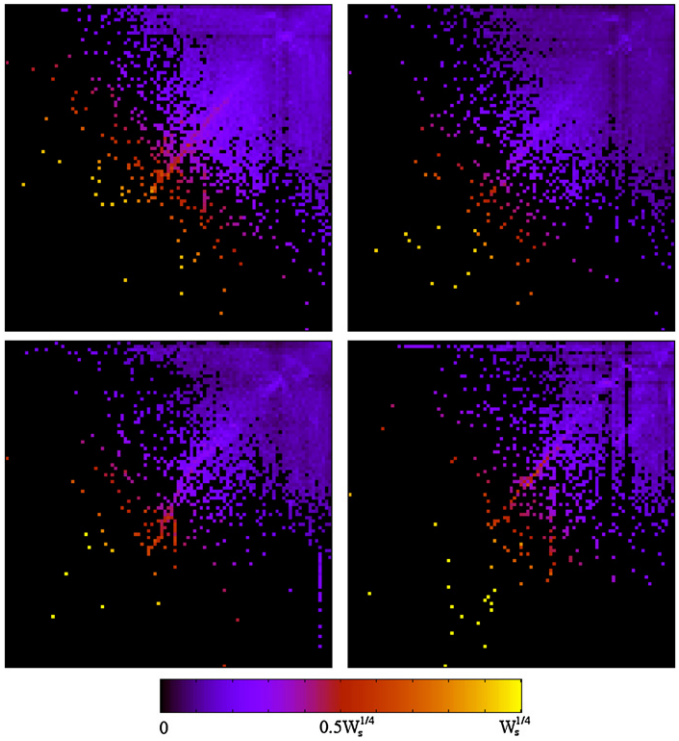


Figure 3. Density distribution $W(K, K^*) = dN_i/dKdK^*$ for networks of four British universities in the plane of PageRank K and CheiRank K^* indexes in the logscale $(\log_N K, \log_N K^*)$. The density is shown for 100×100 equidistant grid in $\log_N K$, $\log_N K^* \in [0, 1]$, the density is averaged 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 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 $0.5W_M^{1/4} \leqslant W^{1/4} \leqslant W_M^{1/4}$ to show low densities in a better way. The panels show networks of University of Cambridge (2006) with $N = 212\,710$ (top left); University of Oxford with $N = 200\,823$ (top right); University of Bath with $N = 73\,491$ (bottom left); University of East Anglia with $N = 33\,623$ (bottom right). The axes show $\log_N K$ in the x-axis and $\log_N K^*$ in the y-axis, in both axes the variation range is $(0, 1)$.

$P^*(K^*(i))$ vectors that will be discussed in more detail in the next section. 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 $K - K^* = \text{const}$. This is similar to the properties of density distribution for the Wikipedia network discussed in [15] (see also the bottom-right panel in figure 13).

The density of nodes for Linux networks is shown in figure 4. In these networks, the density is homogeneous along lines $K + K^* = \text{const}$ that correspond to absence of correlations between $P(K(i))$ and $P^*(K^*(i))$. Indeed, in the absence of such correlations the distribution of nodes in the K, K^* plane is given by the product of independent probabilities. In the log-scale format used in figure 4, this leads to a homogeneous density of nodes in the top-right corner of the $(\log_N K, \log_N K^*)$ plane as it was discussed in [15, see right panel in figure 4]. Indeed, the distributions in figure 4 are very homogeneous inside the top-right triangle. We note that, a part from fluctuations, the distributions remain rather stable even if the size of the network is changed by factor 20 from the V2.0 to V2.6 version. The physical reasons for the absence of correlations between $P(i)$ and $P^*(i)$ have been explained in [14] on the basis of the concept of 'separation of concerns' used in software architecture. As discussed in [14], a good code should

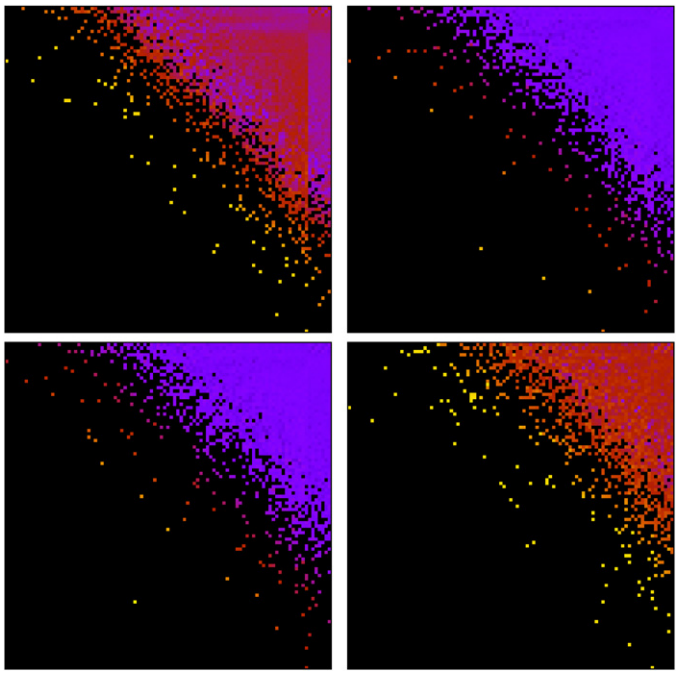


Figure 4. Density distribution $W(K, K^*) = dN_i/dKdK^*$ of four Linux Kernel networks shown in the same frame as in figure 3. The panels show networks for Linux versions V2.0 with $N = 14\,080$ (top left); V2.3 with $N = 41\,117$ (top right); V2.4 with $N = 85\,757$ (bottom left); V2.6 with $N = 285\,510$ (bottom right). Color panel is the same as in figure 3 with a saturation value of $W_s^{1/4} = 0.2W_M^{1/4}$ so that the same color is fixed for $0.2W_M^{1/4} \leqslant W^{1/4} \leqslant W_M^{1/4}$ to show low densities in a better way. The axes show $\log_N K$ in the x-axis and $\log_N K^*$ in the y-axis, in both axes the variation range is $(0, 1)$.

decrease a number of procedures that have high values of both PageRank and CheiRank; 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. These 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 fulfil 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. We discuss the properties of PageRank–CheiRank correlations in the next section.

3. Correlations between PageRank and CheiRank

The correlations between PageRank and CheiRank can be quantitatively characterized by the correlator

$$\kappa(\tau) = N \sum_{i=1}^{N} P(K(i) + \tau)P^*(K^*(i)) - 1. \tag{2}$$

Such a correlator was introduced in [14] for $\tau = 0$ and we will use the same notation $\kappa = \kappa(\tau = 0)$. This correlator at $\tau = 0$ shows if there are correlations and dependences between PageRank and CheiRank vectors. Indeed, for homogeneous vectors $P(K) = P^*(K^*) = 1/N$ we have $\kappa = 0$ corresponding to absence of correlations. We will see below that the values

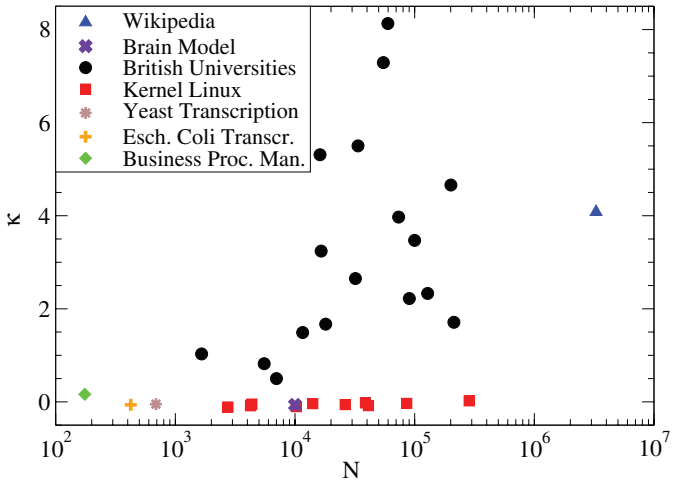


Figure 5. Correlator κ as a function of the number of nodes N for different networks: Wikipedia network, 17 British universities, ten versions of Kernel Linux Kernel PCN, *Escherichia Coli* and Yeast transcription gene networks, brain model network and business process management network. The parameters of networks are given in the appendix.

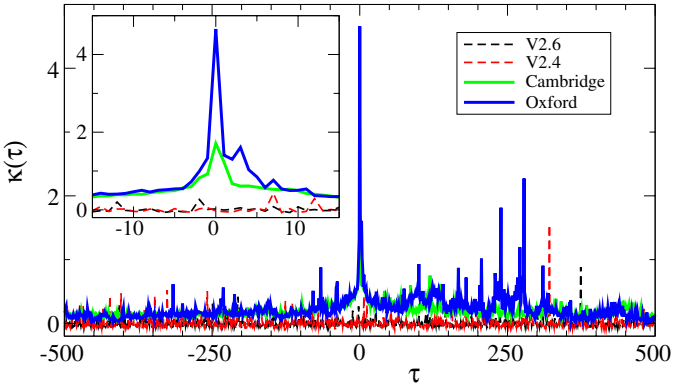


Figure 6. Correlator $\kappa(\tau)$ for two different long and short range of τ in the main and inset panel, respectively. The Kernel Linux PCN V2.6 and V2.4 are shown by dashed curves while universities networks of Cambridge and Oxford are shown by full curves.

of κ are very different for various directed networks. Hence, this new characteristic is able to distinguish various types of networks even if they have rather similar algebraic decay of PageRank and CheiRank vectors.

The values of κ for networks of various size N are shown in figure 5. The two types of networks are well visible according to these data. The human created university and Wikipedia networks have typical values of κ in the range $1 < \kappa < 8$. Other networks like Linux PCN, gene transcription networks, brain model and business process management networks have $\kappa \approx 0$.

The dependence of $\kappa(\tau)$ on the correlation 'time' τ is shown in figure 6. For the PCN of Linux there are no correlations at any τ, while for the university networks we find that the correlator drops to small values with increase of $|\tau|$ (e.g. $|\tau| > 5$) even if at certain rather large values of $|\tau|$ significant values of correlator κ can reappear.

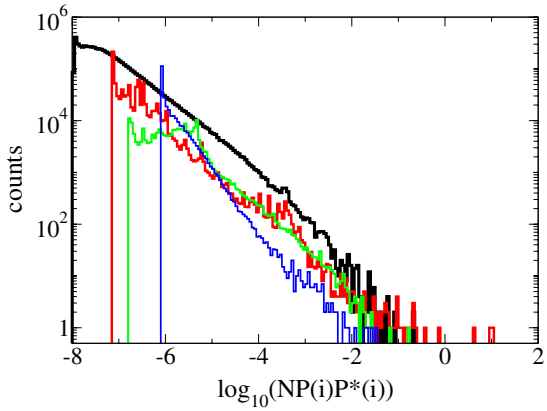


Figure 7. Histogram of frequency appearance of correlator components $\kappa_i = NP(K(i))P^*(K^*(i))$ for networks of Wikipedia (black), University of Cambridge in 2006 (green) and in 2011 (red), and PCN of Linux Kernel V2.6 (blue). For the histogram the whole interval $10^{-8} \leqslant \kappa_i \leqslant 10^2$ is divided into 200 cells of equal size in the logarithmic scale. Curve colors are black, red, green and blue from left to right at the bottom of the vertical axis.

It is interesting to see what are typical values $\kappa_i = NP(K(i))P^*(K^*(i))$ of contributions in the correlator sum (2) at $\tau = 0$. The distribution of κ_i values for a few networks are shown in figure 7. All of them follow a power law with an exponent a =1.23 for PCN Linux, 0.70 for Wikipedia and 0.76 (2006) and 0.66 (2011) for University of Cambridge. We note that further studies are required to analytically obtain the values of the exponent a. In the latter two cases the exponent and the distribution shape remains stable in time; however, in 2011 there appear few nodes with very large κ_i values which give a significant increase of the correlator from $\kappa = 1.71$ (in 2006) up to $\kappa = 30.0$ (in 2011). It is possible that such a situation can appear if it is imposed that practically any page points to the main university page, which may have a rather high CheiRank due to many outgoing links to other departments and university divisions. We suppose that these are also the reasons why we have the appearance of large values of $\kappa(\tau)$ in university networks. At the same time more detailed studies are required to clarify the correlation properties on directed networks of a deeper level. We will return in section 7 to a discussion of university networks collected in 2011.

Another way to analyze the correlations between PageRank and CheiRank is simply to count the number of nodes $\Delta(n)$ inside a square $1 \leqslant K(i), K^*(i) \leqslant n$. For a totally correlated distribution with $K(i) = K^*(i)$ we have $\Delta(n)/N = n/N$, while in absence of correlations we should have points homogeneously distributed inside a square $n \times n$ that gives $\Delta(n)/N = (n/N)^2$. The dependence of such point-count correlator $\Delta(n)$ on size n is displayed in figure 8 for various networks. These data clearly show that the Linux PCN is uncorrelated being close to the limiting uncorrelated dependence, while Wikipedia and British university networks show intermediate strength of correlations being between the two limiting functions of $\Delta(n)$.

4. Information flow of 2DRanking

According to 2DRanking, all network nodes are distributed on a two-dimensional plane (K, K^*). The directed links of the network create an information flow in this plane. To visualize this flow, we use the following procedure:

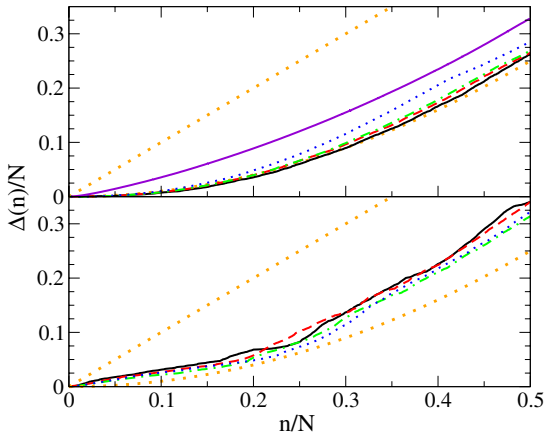


Figure 8. Dependence of the point-count correlation function $\Delta(n)/N$ on n/N for networks of Wikipedia, British universities and Kernel Linux PCN. The curves in the top panel show the cases of Wikipedia (solid violet/gray) and four versions of PCN of Linux Kernel with V2.0 (solid black), V2.3 (dashed red), V2.4 (dot-dashed green) and V2.6 (dotted blue). The curves in the bottom panel show the cases of British universities with East Anglia (solid black), Bath (dashed red), Oxford (dot-dashed green) and Cambridge 2006 (dotted blue). Dotted orange curves represent the totally correlated case with $\Delta(n)/N = n/N$ and the totally uncorrelated one with $\Delta(n)/N = (n/N)^2$.

(a) each node is represented by one point in the (K, K^*) plane;
(b) the whole space is divided into equal size cells with indexes (i, i^*) with the number of nodes inside each cell being n_{i,i^*}, in figure 9 we use cells of equal size in usual (left column) and logarithmic (right column) scales;
(c) for each node inside the cell (i, i^*), pointing to any other cell $(i', i^{*\prime})$, we compute the vector $(i' - i, i^{*\prime} - i^*)$ and average it over all nodes n_{i,i^*} inside the cell (the weight of links is not taken into account);
(d) we put an arrow centered at (i, i^*) with the modulus and direction given by the average vector computed in (c).

Examples of such average flows for the networks of figure 1 are shown in figure 9. All flows have a fixed point attractor. The fixed point is located at rather large values $K, K^* \sim N/4$, that is, due to the fact that in average nodes with maximal values $K, K^* \sim N$ point to lower values. At the same time nodes with very small $K, K^* \sim 1$ still point to some nodes which have larger values of K, K^* that places the fixed point at certain intermediate K, K^* values. We note that the analyzed directed networks have dangling nodes which have no outgoing links, the fraction of such nodes is especially large for the Linux network. Due to the absence of outgoing links, we obtain an empty white region in the information flow shown in figure 9. A more detailed analysis of statistical properties of information flows on the PageRank–CheiRank plane requires further study.

5. Control of spam links

For many networks, ingoing and outgoing links have their own importance and thus should be treated on equal grounds by PageRank and CheiRank as described above. However, for the WWW it is more easy to manipulate outgoing links which are handled by an owner of a

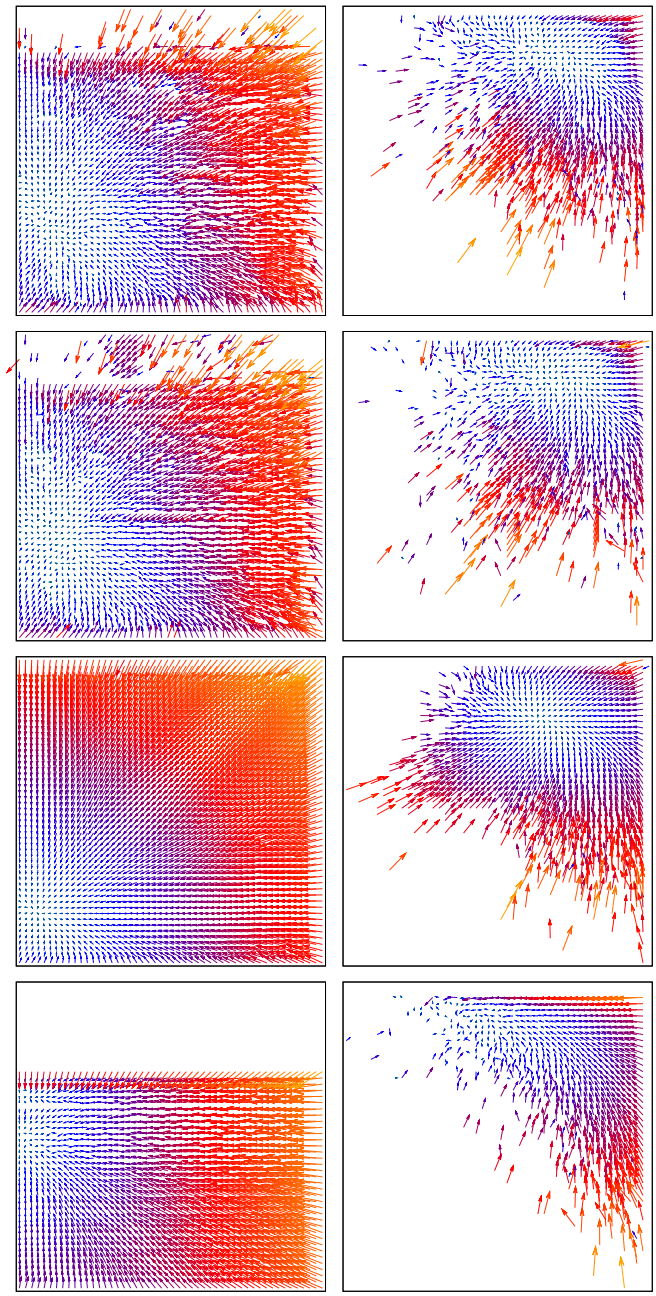


Figure 9. Information flow on the PageRank–CheiRank plane (K, K^*) generated by directed links of the networks of figure 1. Outgoing links flow is shown in the linear scale (K, K^*) with $K, K^* \in [1, N]$ on left panels, and in the logarithmic scale $(\log_N K, \log_N K^*)$ for $\log_N K, \log_N K^* \in [0, 1]$ on right panels. The flow is shown by arrows whose size is proportional to the vector amplitude, which is also indicated by color [from yellow for large to blue for small amplitudes]. The rows corresponds to University of Cambridge (2006); University of Oxford (2006), Wikipedia English articles, PCN of Linux Kernel V2.6 (from top to bottom). The axes show: on the left column K/N in the x-axis, K^*/N in the y-axis; on the right column $\log_N K$ in the x-axis, $\log_N K^*$ in the y-axis; in all axes the variation range is $(0, 1)$.

given webpage, while ingoing links are handled by other users. This requires the introduction of some level of control on the outgoing links which should be taken into account for the ratings. Since it is very easy to create links to highly popular sites, we will call 'spam links' links for which the destination site is much more popular than the source. A quantitative measure of popularity can be provided by the PageRank of the sites. We do not think that spam links are frequent in networks such as procedure calls in the Linux kernel, Wikipedia and gene regulation. Even for university networks we think that there is not much reason to put spam links inside the university domain. However, for a large-scale WWW an excessive number of such spam links can become harmful for the network performance. However, for WWW networks spam links are probably more widespread. Some websites may try to improve their rating by carefully choosing their outgoing links. Also it is a common policy to have links back to a website's root pages to facilitate navigation. Naturally, a good rating should not be sensitive to the presence of such links. Thus it is important to treat spam links appropriately in order to construct a two-dimensional web-search engine. Below we propose a method for spam links control and test it on an example of the Wikipedia network which has the largest size among networks analyzed in this paper. We stress that this is done as a test example and not because we think that there are spam links between Wikipedia articles.

With this aim, we propose the following filter procedure for computation of CheiRank. The standard procedure described above is to invert the directions of all links of the network and then to compute the CheiRank. The filter procedure inverts a link from j to i only if $\eta P(K(j)) > P(K(i))$, where η is some positive filter parameter. After such an inversion of certain links, while other links remain unchanged, the matrix S^* and G^* are computed and the CheiRank vector $P^*(K^*(i))$ of G^* is determined in a usual way. From the definition it is clear that for $\eta = 0$ there are no inverted links, and thus after filtering P^* is the same as the PageRank vector P. In the opposite limit $\eta = \infty$ all links are inverted and P^* is then the usual CheiRank discussed in previous sections. Thus intermediate values of η allow us to handle the properties of CheiRank depending on a wanted strength of filtering. We note that the proposed filtering procedure is rather generic and can be applied to various types of directed networks.

The dependence of the fraction f of inverted links (defined as a ratio between the number of inverted links to the total number of links) on the filter parameter η is shown for various networks in figure 10. There is a significant jump of f at $\eta \approx 1$ for British university networks. In fact the condition $\eta \approx 1$ corresponds approximately to the border relation $P(K) \approx P(K')$ with $K \approx K'$ that marks the diagonal of the G matrix shown in figure 1, which has a significant density of matrix elements. As a result for $\eta > 1$, we have a significant increase of inversion of links leading to a jump of f present in figure 10. The diagonal density is most pronounced for university networks so that for them the jump of f is mostly sharp.

It is also convenient to consider another condition for link inversion defined not for $P(K_i)$ but directly in the plane (K, K') defined by the condition: links are inverted only if $K(j) < \eta_K K(i)$ (where node j points to node i, $j \to i$). In a first approximation, we can assume that the links are homogeneously distributed in the plane of transitions from K to K'. This density is similar to the density distribution of Google matrix elements $G_{K'K}$ shown in figure 1. For the homogeneous distribution, the fraction f of inverted links is given by an area $\eta_K/2$ of a triangle, whose height is 1 and the basis is η_K, for $\eta_K \leqslant 1$. In a similar way, we have $f = 1 - 1/2\eta_K$ for $\eta_K \geqslant 1$. We can generalize this distribution assuming that there are only links with $1 \leqslant K' \leqslant aN$, that is, approximately the case for Linux network where $a = 0.4$ (see figure 1 bottom row), and that inside this interval the density of links decreases as $1/(K')^{\nu}$. Then after computing the area we obtain the expression for the fraction of inverted links valid

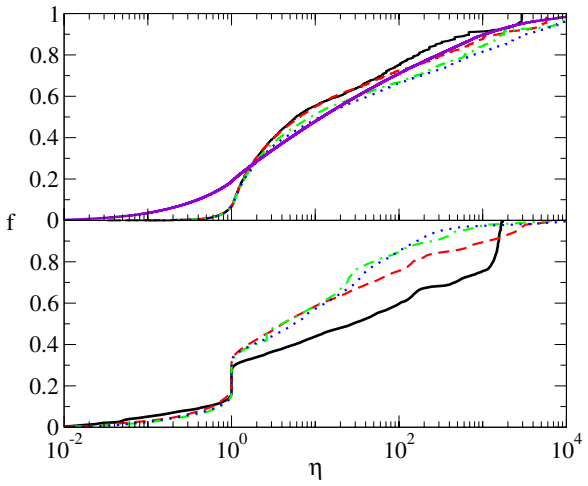


Figure 10. Fraction f of inverted links as a function of filter parameter η for various studied networks. Top panel: Wikipedia (violet/gray curve) and four versions of Kernel Linux PCN with V2.0 (solid black curve), V2.3 (dashed red curve), V2.4 (dot-dashed green curve) V2.6 (dotted blue curve). Bottom panel shows data for British university networks with East Anglia (solid black curve), Bath (dashed red curve), Oxford (dot-dashed green curve) and Cambridge 2006 (dotted blue curve).

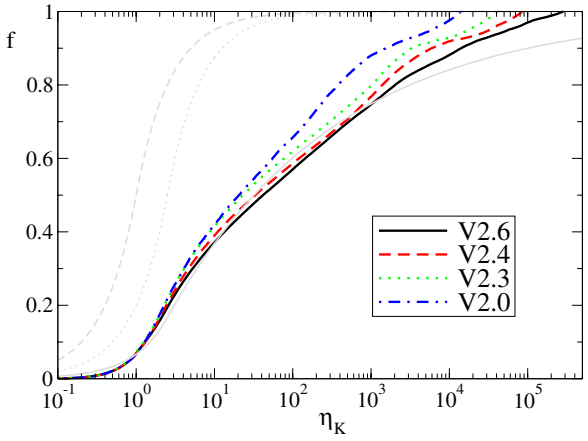


Figure 11. Fraction f of inverted links in the (K, K') plane with the condition $K(j) < \eta_K K(i)$ shown as a function of filter parameter η_K for Linux networks versions shown by different curves. Gray curves from left to right are the theory curves with $a = 1$, $\nu = 0$ (dashed); $a = 0.4$, $\nu = 0$ (dotted) and $a = 0.4$, $\nu = 0.8$ (full) (see text).

for $0 \leqslant \nu < 1$:

$$f(\eta_K) = \begin{cases} \dfrac{1-\nu}{2-\nu}(a\eta_K) & \eta_K \leqslant 1/a \\ 1 + \left(\dfrac{1-\nu}{2-\nu} - 1\right)(a\eta_K)^{\nu-1} & \eta_K > 1/a \end{cases} . \tag{3}$$

The comparison of this theoretical expression with the numerical data for Linux PCN is shown in figure 11. It shows that the data for Linux are well described by the theory (3) with $a = 0.4$

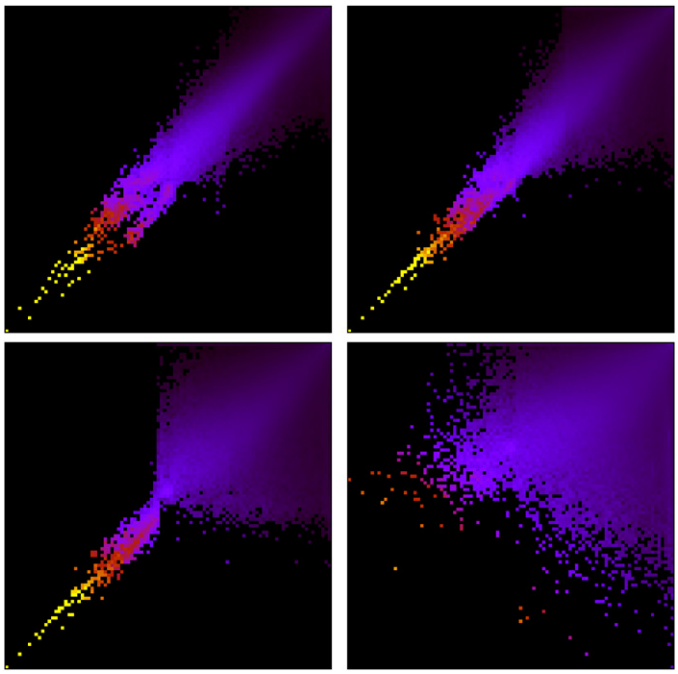


Figure 12. Density distribution $W(K, K^*) = dN_i/dKdK^*$ for Wikipedia in the plane of PageRank and filtered CheiRank indexes, $(\log_N K, \log_N K^*)$, in a equidistant 100×100 lattice with $\log_N K, \log_N K^* \in [0, 1]$. The filter parameter is $\eta = 10$ (left-top panel), 100 (right-top panel), 1000 (left-bottom panel), 10^5 where all links are inverted (right-bottom panel). The color panel is the same as in figure 3 with the saturation value $W_s^{1/4} = 0.5W_M^{1/4}$. The axes show: $\log_N K$ in the x-axis, $\log_N K^*$ in the y-axis, in both axes the variation range is $(0, 1)$.

and $\nu = 0.8$. The last value takes into account the fact that the density of links decreases with PageRank index K' as it is well visible in figure 1.

The variation of nodes density in the plane of PageRank and filtered CheiRank (K, K^*) for the Wikipedia network is shown in figure 12 with the filtering by η for $P(K)$ and $P(K')$ values. At moderate values $\eta = 10$ the density is concentrated near the diagonal, with further increase of $\eta = 100, 1000$ a broader density distribution appears at large K values which goes to smaller and smaller K until the limiting distribution without filtering is established at very large η. The top 100 Wikipedia articles obtained with filtered CheiRank at the above values of η are given at [25]. We also give there top articles in 2DRank which gives articles in order of their appearance on the borders of a square of increasing size in (K, K^*) plane (see the detailed description in [15]). These data clearly show that filtering eliminates articles with many outgoing links and gives a significant modification of top CheiRank articles. Thus the described method can be efficiently used for control of spam links present in the WWW.

6. 2DRanking of gene regulation networks

The method of 2DRanking described above is rather generic and can be applied to various types of directed networks. Here, we apply it to gene regulation networks of *Escherichia Coli* and Yeast with the network links taken from [24]. Such transcription regulation networks control the expression of genes and have important biological functions [23].

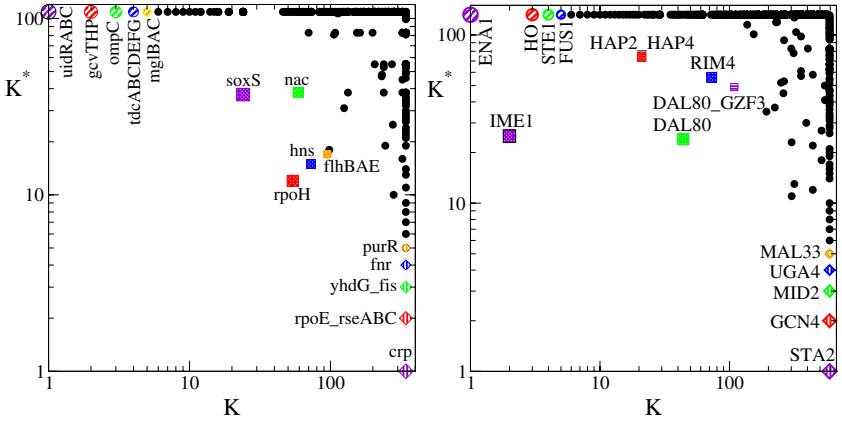


Figure 13. Distribution of nodes in the plane of PageRank K and CheiRank K^* for *Escherichia Coli* and Yeast transcription networks on left and right panels, respectively (network data are taken from [24]). The nodes with the five top probability values of PageRank, CheiRank and 2DRank are labeled by their corresponding node names; they correspond to five lowest index values.

The distribution of nodes in PageRank–CheiRank plane is shown in figure 13. The top five nodes in CheiRank probability value (lowest CheiRank indexes) are those which send many outgoing orders, the top five in PageRank probability value are those which obtain many incoming signals and the top five indexes in 2DRank (with five lowest 2DRank index values) combine these two functions. For these networks the correlator κ is close to zero (even slightly negative), which indicates the statistical independence between outgoing and ingoing links 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. We use these networks here to show that the general methods proposed above can be applied to these directed networks as well. Whether the obtained ratings can bring deep insights into the functioning of gene regulation can only be assessed by experts in the field. However, we hope that such an analysis will prove to be useful for a better understanding of gene regulation networks.

7. Discussion

Above we presented extensive studies of statistical properties of 2DRanking based on PageRank and CheiRank for various types of directed networks. All studied networks are of a free-scale type with an algebraic distribution of ingoing and outgoing links with a usual value of exponents. In spite of that their statistical characteristics related to PageRank and CheiRank are rather different. Some networks have high correlators between PageRank and CheiRank (e.g. Wikipedia, British universities), while others have practically zero correlators (PCN of Linux Kernel, gene regulation networks). The distribution of nodes in PageRank–CheiRank plane also varies significantly between different types of networks. Thus 2DRanking discussed here gives more detailed classification of information flows on directed networks.

We think that 2DRanking gives 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 (see appendix and figure 2). At present, web robots start automatically generating new webpages. These features can be responsible for the appearance of gaps in the density distribution in the (K, K^*) plane at large $K, K^* \sim N$ values visible for large-scale university networks of Cambridge and ENS Paris in 2011 (see

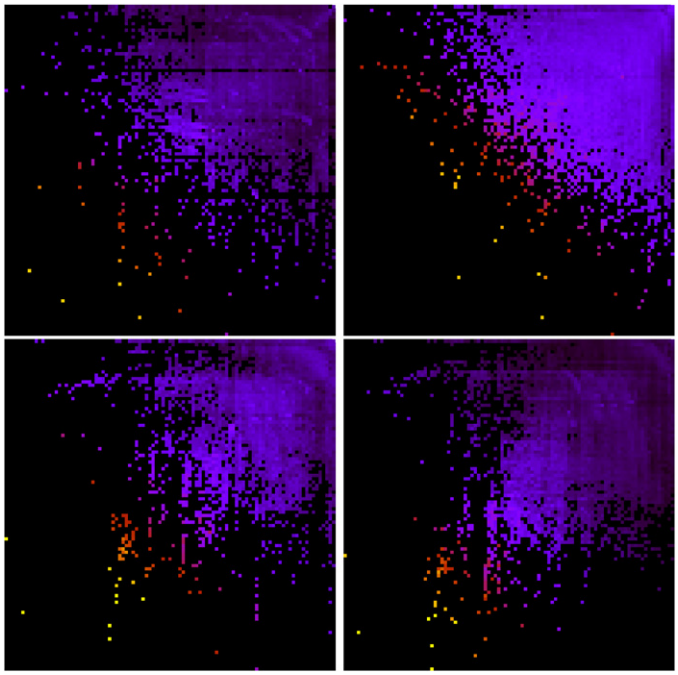


Figure 14. Density distribution $W(K, K^*) = dN_i/dKdK^*$ shown in the same frame as in figure 3 for networks collected in 2011: University of Cambridge (top left), University of Bologna (top right), ENS Paris for crawling level 5 (bottom left) and 7 (bottom right). The color panel is the same as in figure 3 with the saturation value $W_s^{1/4} = 0.5W_M^{1/4}$. The axes show: $\log_N K$ in the x-axis, $\log_N K^*$ in the y-axis, is both axes the variation range is (0, 1).

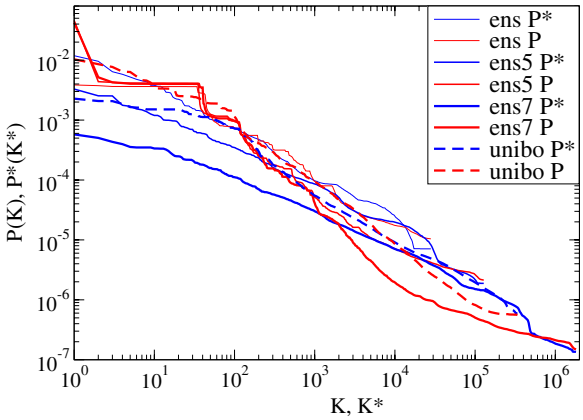


Figure 15. Dependence of probabilities of PageRank $P(K)$ (red/gray curve) and CheiRank $P^*(K^*)$ (blue/black curve) on corresponding ranks K and K^* for the networks of ENS Paris (crawling levels 3,5,7) and the University of Bologna.

figure 14). Such an automatic generation of links can change the scale-free properties of networks. Indeed, for ENS Paris we observe the appearance of a large step in the PageRank distribution $P(K)$ shown in figure 15. This step for $P(K)$ remains not sensitive to the deepness of crawling which goes on a level of 3, 5 and 7 links. However, the CheiRank distribution changes with the deepness level becoming more and more flat (see figure 15). Such a tendency

in a modification of network statistical properties is visible in 2011 for large-size university networks, while networks of moderate size, like the University of Bologna 2011 (see data in figures 14 and 15), are not yet affected. A sign of ongoing changes is a significant growth of the correlator value κ which increases up to a very large value (30 for Cambridge 2011 and 63 for ENS Paris). There is a danger that automatic generation of links can lead to a delocalization transition of PageRank that can destroy efficiency of information retrieval from the WWW. We note that it is known that PageRank delocalization can appear in certain models of Markov chains and Ulam networks [26, 27] (see e.g. in [26] figure 1 (right-top panel) and figure 6 directly showing the delocalization of PageRank vector). Such a delocalization of PageRank would make the ranking of nodes inefficient due to high sensitivity of ranking to fluctuations that would create a very dangerous situation for the WWW information retrieval and ranking. We also note that the spectrum of the Google matrix of British universities networks has been recently analyzed in [28]. The spectrum and eigenstates analysis can be a sensitive tool for location of precursors of a delocalization transition.

Our studies of 2DRanking pave the way to the development of two-dimensional search engines which will use the advantages of both PageRank and CheiRank. Indeed, the Google search engine uses as the fundamental mathematical basis the one-dimension ranking related to PageRank [7]. Of course, there are various other important elements used by the Google search which remain the company secret, and not only PageRank order matters for the Google ranking. However, the mathematical aspects of these additional elements are not really known (e.g. they are not described in [7]). At the same time, the size of databases generated by the modern society continues its enormous growth. Due to that, the information retrieval and ordering of such datasets becomes of primary importance and new mathematical tools should be developed to operate and characterize efficiently their information flows and ranking. Here we proposed and analyzed the properties of the new two-dimensional search engine, which we call Dvvadi from Russian 'dva (two)' and 'dimension' that will use the complementary ranking abilities of both PageRank and CheiRank. Now the procedure of ordering of all network nodes uses not one but two vectors of the Google matrix of a network. The computational efforts are twice as expensive but for that we obtain a new quality, since now the nodes are ranked in the 2D plane not only by their degree of popularity but also by their degree of communicability. Thus for the Wikipedia network the top three articles in PageRank probability are three countries (most popular), while the top three articles in CheiRank probability are three listings of knowledge, state leaders and geographical places (most communicative). Hence, we can rank the nodes of the network in a new two-dimensional manner which highlight complementary properties of node popularity and communicability. Thus, the Dvvadi search can present nodes not in a line but on a 2D plane characterizing these two complementary properties of nodes. Examples of such 2D representation of nodes selected from Wikipedia articles by a specific subject are shown in figure 16: we determine global K and K^* indexes of all articles, select a specific subject (e.g. *countries*) and then represent countries in the local index K and K^* corresponding to their appearance in the global order via PageRank and CheiRank. For countries, we see a clear tendency that the countries on the top of PageRank probability (low K) have relatively high CheiRank index (high K^*) (e.g. US, UK, France) while small countries in the region $K \approx 50$, $K^* \approx 10$ have another tendency (e.g. Singapore). We attribute this to specific routes of cultural and industrial development of the world: e.g. Singapore was a colony of UK and became a strong trade country and due to that has historically many links pointing to the UK and other developed countries. For universities we also see that those at the top of PageRank (Harvard, Oxford and Cambridge) are not very communicative having high K^* values, while Columbia and Berkeley are more balanced, and Florida and FSU are very communicative probably due to the initial location of the Wikimedia Foundation

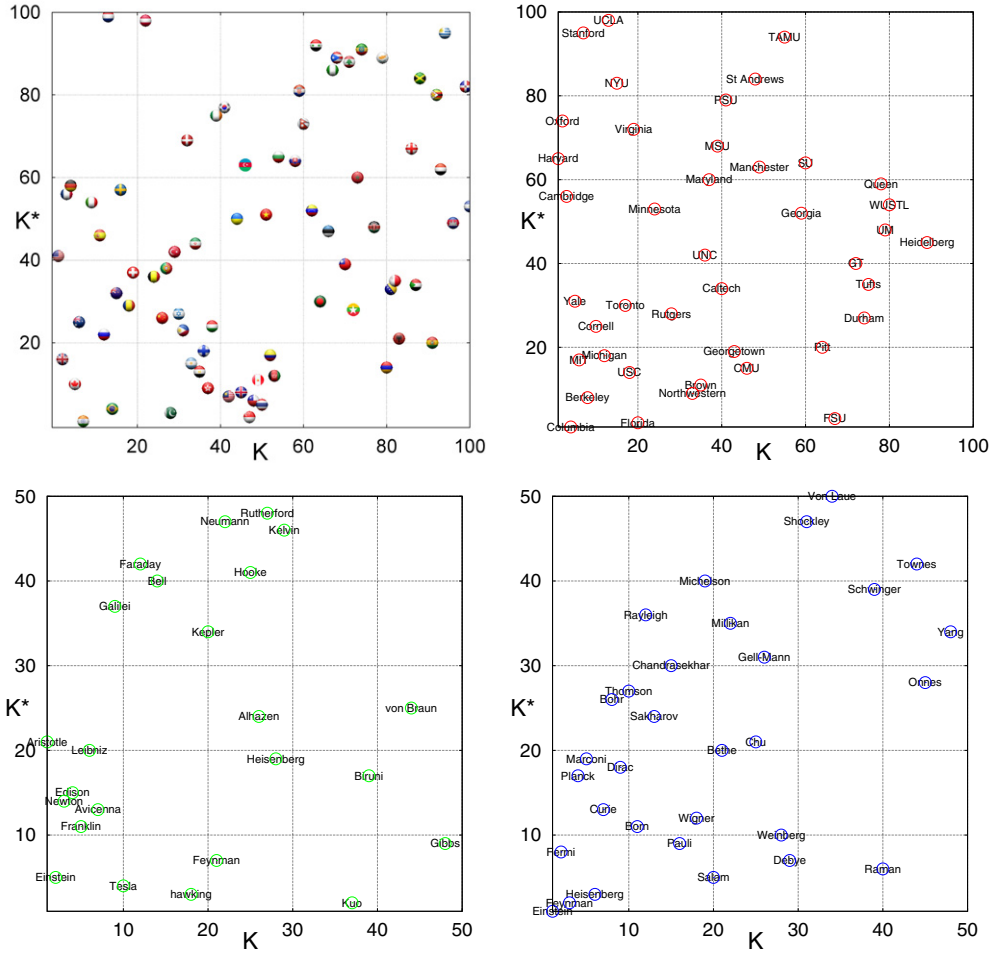


Figure 16. Examples of Dvvadi search analysis of Wikipedia articles shown on the 2D plane of PageRank K and CheiRank K^* local indexes for specific subjects (articles): countries marked by their flag (top left), universities (top right), physicists (bottom left), Nobel laureates in physics (bottom right), circles mark the node location; high resolution figures and listings of names with local (K, K^*) values in 100×100 square are available at [25] (listings with global ranking are available at [15]).

at Florida. For physicists, we see that links to many scientific fields (like Shen Kuo) or popularization of science (like Hawking and Feynman) place those people at the top positions of CheiRank. In a similar way, for the Nobel laureates in physics we see that CheiRank stresses the communicative aspects: e.g. Feynman, due to his popularization of physics; Salam, due to the institute with his name at Trieste, with a broad international activity; Raman, due to the Raman effect.

On the basis of the above results, we think that PageRank–CheiRank classification of network nodes on 2D plane will allow us to analyze the information flows on directed networks in a better way. It is also important to note that 2DRanking is very natural for financial and trade networks. Indeed, the world trade usually uses the import and export ranking which is analogous to PageRank and CheiRank, as it is shown in [20]. We think that such Dvvadi

Table A1. Linux Kernel network parameters

Version	N	N_{links}	κ
V1.0	2 752	5 933	$\kappa = -0.11$
V1.1	4 247	9 710	$\kappa = -0.083$
V1.2	4 359	10 215	$\kappa = -0.048$
V1.3	10 233	24 343	$\kappa = -0.102$
V2.0	14 080	34 551	$\kappa = -0.037$
V2.1	26 268	59 230	$\kappa = -0.058$
V2.2	38 767	87 480	$\kappa = -0.022$
V2.3	41 117	89 355	$\kappa = -0.081$
V2.4	85 757	195 106	$\kappa = -0.034$
V2.6	285 510	588 861	$\kappa = 0.022$

Table A2. British universities network parameters

University	N	N_{links}	κ
RGU (Abardeen)	1 658	15 295	$\kappa = 1.03$
Uwic (Wales)	5 524	111 733	$\kappa = 0.82$
NTU (Nottingham)	6 999	143 358	$\kappa = 0.50$
Liverpool	11 590	141 447	$\kappa = 1.49$
Hull	16 176	236 525	$\kappa = 5.31$
Keele	16 530	117 944	$\kappa = 3.24$
UCE (Birmingham)	18 055	351 227	$\kappa = 1.67$
Kent	31 972	277 044	$\kappa = 2.65$
East Anglia	33 623	325 967	$\kappa = 5.50$
Sussex	54 759	804 246	$\kappa = 7.29$
York	59 689	414 200	$\kappa = 8.13$
Bath	73 491	541 351	$\kappa = 3.97$
Glasgow	90 218	544 774	$\kappa = 2.22$
Manchester	99 930	1254 939	$\kappa = 3.47$
UCL (London)	128 450	1397 261	$\kappa = 2.33$
Oxford	200 823	1831 542	$\kappa = 4.66$
Cambridge (2006)	212 710	2015 265	$\kappa = 1.71$

engine/motor [25] will find useful applications for the treatment of enormously large databases created by modern society.

Acknowledgments

We thank K M Frahm and B Georgeot for useful discussions of properties of British university networks. This work is done in the frame of the EC FET Open project 'New tools and algorithms for directed network analysis' (NADINE No 288 956).

Appendix

We list below the directed networks used in this work giving for them number of nodes N, number of links N_{links} and correlator between PageRank and CheiRank κ. Additional data can be find at [25].

Linux Kernel PCNs are taken from [14] (see also [22]) with the parameters for various kernel versions shown in table A1.

Web networks of British universities dated by year 2006 are taken from [21] and are shown in table A2.

We also developed a special code with which we performed crawling of university web networks in January—March 2011 with the parameters given below: University of Cambridge (2011) with $N = 898\,262$, $N_{\rm links} = 15\,027\,630$, $\kappa = 30.0$; École Normale Supérieure, Paris (ENS 2011) with $N = 28\,144$, $N_{\rm links} = 971\,856$, $\kappa = 1.67$ (crawling deepness level of three links), $N = 129\,910$, $N_{\rm links} = 2111\,944$, $\kappa = 16.2$ (crawling deepness level of five links), $N = 1820\,015$, $N_{\rm links} = 25\,706\,373$, $\kappa = 63.6$ (crawling deepness level of seven links); University of Bologna with $N = 339\,872$, $N_{\rm links} = 16\,345\,488$, $\kappa = 2.63$.

The data for the hyperlink network of Wikipedia English articles (2009) are taken from [15] with $N = 3282\,257$, $N_{\rm links} = 71\,012\,307$, $\kappa = 4.08$.

Transcription gene networks are taken from [24]. We have for them: *Escherichia Coli* with $N = 423$, $N_{\rm links} = 519$, $\kappa = -0.0645$; Yeast with $N = 690$, $N_{\rm links} = 1079$, $\kappa = -0.0497$; for all links the weight is taken to be the same.

Business process management network is taken from [19] with $N = 175$, $N_{\rm links} = 240$, $\kappa = 0.164$.

Brain model network is taken from [18] with $N = 10\,000$, $N_{\rm links} = 1960\,108$, $\kappa = -0.054$ (unweighted), $\kappa = -0.065$ (weighted).

References

[1] See e.g. The Size of the World Wide Web www.worldwidewebsize.com/
[2] Wikipedia (The Free Encyclopedia) 2011 Facebook http://en.wikipedia.org/wiki/Facebook
[3] Wikipedia (The Free Encyclopedia) 2011 Vkontakte http://en.wikipedia.org/wiki/VK_(social_network)
[4] Wikipedia (The Free Encyclopedia) 2011 Web search engine http://en.wikipedia.org/wiki/Web_search_engine
[5] Büttcher S, Clarke C L A and Cormack G V 2010 *Information Retrieval: Implementing and Evaluating Search Engines* (Cambridge: MA: MIT Press)
[6] Brin S and Page L 1998 *Comput. Netw. ISDN Syst.* **30** 107
[7] Langville A M and Meyer C D 2006 *Google's PageRank and Beyond: The Science of Search Engine Rankings* (Princeton, NJ: Princeton University Press)
[8] Franceschet M 2010 PageRank: Standing on the shoulders of giants arXiv:1002.2858v3 [cs.IR]
[9] Vigna S 2011 Spectral Ranking (available at http://vigna.dsi.unimi.it/ftp/papers/SpectralRanking.pdf)
[10] Litvak N, Scheinhardt W R W and Volkovich Y 2008 *Lecture Notes Comput. Sci.* **4936** 72
[11] Donato D, Laura L, Leonardi S and Millozzi S 2004 *Eur. Phys. J.* B **38** 239
[12] Pandurangan G, Raghavan P and Upfal E 2005 *Internet Math.* **3** 1
[13] Georgeot B, Giraud O and Shepelyansky D L 2010 *Phys. Rev.* E **81** 056109
[14] Chepelianskii A D 2010 arXiv:1003.5455 [cs.SE] (www.quantware.ups-tlse.fr/QWLIB/linuxnetwork/)
[15] Zhirov A O, Zhirov O V and Shepelyansky D L 2010 *Eur. Phys. J.* B **77** 523 (www.quantware.ups-tlse.fr/QWLIB/2drankwikipedia/)
[16] Fogaras D 2003 Where to start browsing the Web? (*Lecture Notes in Computer Science* vol 2877) p 65
[17] Hrisitidis V, Hwang H and Papakonstantinou Y 2008 *ACM Trans. Database Syst.* **33** 1
[18] Shepelyansky D L and Zhirov O V 2010 *Phys. Lett.* A **374** 3206
[19] Abel M and Shepelyansky D L 2011 *Eur. Phys. J.* B **84** 493 (www.quantware.ups-tlse.fr/QWLIB/cheirankbusiness/)
[20] Ermann L and Shepelyansky D L 2011 *Acta Phys. Pol.* **120** A158 (arXiv:1103.5027 [q-fin.GN]) (www.quantware.ups-tlse.fr/QWLIB/tradecheirank/)
[21] Academic Web Link Database Project http://cybermetrics.wlv.ac.uk/database/
[22] Ermann L, Chepelianskii A D and Shepelyansky D L 2011 *Eur. Phys. J.* B **79** 115
[23] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U 2002 *Science* **298** 824
[24] Uri Alon Lab www.weizmann.ac.il/mcb/UriAlon/ (Complex networks section)
[25] Ermann L, Chepelianskii A D and Shepelyansky D L 2012 arXiv:1106.6215 [cs.IR] (www.quantware.ups-tlse.fr/QWLIB/dvvadi/)
[26] Shepelyansky D L and Zhirov O V 2010 *Phys. Rev.* E **81** 036213
[27] Ermann L and Shepelyansky D L 2010 *Phys. Rev.* E **81** 036221
[28] Frahm K M, Georgeot B and Shepelyansky D L 2011 *J. Phys. A: Math. Theor.* **44** 465101

EPJ B
Condensed Matter and Complex Systems
EPJ.org
your physics journal

Eur. Phys. J. B (2012) 85: 355 DOI: 10.1140/epjb/e2012-30599-6

Google matrix of Twitter

K.M. Frahm and D.L. Shepelyansky

edp sciences
Società Italiana di Fisica
Springer

Eur. Phys. J. B (2012) 85: 355
DOI: 10.1140/epjb/e2012-30599-6

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Google matrix of Twitter

K.M. Frahm and D.L. Shepelyansky[a]

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

Received 14 July 2012 / Received in final form 14 July 2012
Published online 24 October 2012 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2012

Abstract. We construct the Google matrix of the entire Twitter network, dated by July 2009, and analyze its spectrum and eigenstate properties including the PageRank and CheiRank vectors and 2DRanking of all nodes. Our studies show much stronger inter-connectivity between top PageRank nodes for the Twitter network compared to the networks of Wikipedia and British Universities studied previously. Our analysis allows to locate the top Twitter users which control the information flow on the network. We argue that this small fraction of the whole number of users, which can be viewed as the social network elite, plays the dominant role in the process of opinion formation on the network.

1 Introduction

Twitter is an online directed social network that enables its users to exchange short communications of up to 140 characters [1]. In March 2012 this network had around 140 million active users [1]. Being founded in 2006, the size of this network demonstrates an enormously fast growth with 41 million users in July 2009 [2], only three years after its creation. The crawling and statistical analysis of the entire Twitter network, collected in July 2009, was done by the KAIST group [2] with additional statistical characteristics available at LAW DSI of Milano University[1]. This network has scale-free properties with an average power law distribution of ingoing and outgoing links[1] [2] being typical for the World Wide Web (WWW), Wikipedia and other social networks (see e.g [3–5]). In this work we use this Twitter dataset to construct the Google matrix [6,7] of this directed network and we analyze the spectral properties of its eigenvalues and eigenvectors. Even if the entire size of Twitter 2009 is very large the powerful Arnoldi method (see e.g. [8–11]) allows to obtain the spectrum and eigenstates for the largest eigenvalues.

A special analysis is performed for the PageRank vector, used in the Google search engine [6,7], and the CheiRank vector studied for the Linux Kernel network [12,13], Wikipedia articles network [5], world trade network [14] and other directed networks [15]. While the components of the PageRank vector are on average proportional to a number of ingoing links [16], the components of the CheiRank vector are on average proportional to a number of outgoing links [5,12] that leads to a two-dimensional ranking of all network nodes [15]. Thus our studies allow to analyze the spectral properties of the entire Twitter network of an enormously large size which is by one-two orders of magnitude larger compared to previous studies [5,11,13,15].

The paper is organized as follows: the construction of the Google matrix and its global structure are described in Section 2; the properties of spectrum and eigenvectors of the Google matrix of Twitter are presented in Section 3; properties of 2DRanking of Twitter network are analyzed in Section 4 and the discussion of the results is given in Section 5. Detailed data and results of our statistical analysis of the Twitter matrix are presented at the web page[2].

2 Google matrix construction

The Google matrix of the Twitter network is constructed following the standard rules described in [6,7]: we consider the elements A_{ij} of the adjacency matrix being equal to unity if a user (or node) j points to user i and zero otherwise. Then the Google matrix of the network with N users is given by

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

where the matrix S is obtained by normalizing to unity all columns of the adjacency matrix $A_{i,j}$ with at least one non-zero element, and replacing columns with only zero elements, corresponding to the dangling nodes, by $1/N$. The damping factor α in the WWW context describes the probability $(1-\alpha)$ to jump to any node for a random surfer. The value $\alpha = 0.85$ gives a good classification for WWW [7] and thus we also use this value here. The matrix G belongs to the class of Perron-Frobenius

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

[1] Twitter web data of [2] are downloaded from the web site maintained by S. Vigna, http://law.dsi.unimi.it/webdata/twitter-2010.

[2] http://www.quantware.ups-tlse.fr/QWLIB/twittermatrix/.

operators [7], its largest eigenvalue is $\lambda = 1$ and other eigenvalues have $|\lambda| \leq \alpha$. The right eigenvector at $\lambda = 1$ gives the probability $P(i)$ to find a random surfer at site i and is called the PageRank. Once the PageRank is found, all nodes can be sorted by decreasing probabilities $P(i)$. The node rank is then given by index $K(i)$ which reflects the relevance of the node i. The top PageRank nodes are located at small values of $K(i) = 1, 2, \ldots$

The PageRank dependence on K is well described by a power law $P(K) \propto 1/K^{\beta_{in}}$ with $\beta_{in} \approx 0.9$. This is consistent with the relation $\beta_{in} = 1/(\mu_{in} - 1)$ corresponding to the average proportionality of PageRank probability $P(i)$ to its in-degree distribution $w_{in}(k) \propto 1/k^{\mu_{in}}$ where $k(i)$ is a number of ingoing links for a node i [7,16]. For the WWW it is established that for the ingoing links $\mu_{in} \approx 2.1$ (with $\beta_{in} \approx 0.9$) while for the out-degree distribution w_{out} of outgoing links the power law has the exponent $\mu_{out} \approx 2.7$ [3,4]. Similar values of these exponents are found for the WWW British university networks [11], the procedure call network of Linux Kernel software introduced in [12] and for Wikipedia hyperlink citation network of English articles (see e.g. [5]).

In addition to the Google matrix G we also analyze the properties of matrix G^* constructed from the network with inverted directions of links, with the adjacency matrix $A_{i,j} \to A_{j,i}$. After the inversion of links the Google matrix G^* is constructed via the procedure (1) described above. The right eigenvector at unit eigenvalue of the matrix G^* is called the CheiRank [5,12]. In analogy with the PageRank the probability values of CheiRank are proportional to number of outgoing links, due to links inversion. All nodes of the network can be ordered in a decreasing order with the CheiRank index $K^*(i)$ with $P^* \propto 1/K^{*\beta_{out}}$ with $\beta_{out} = 1/(\mu_{out} - 1)$. Since each node i of the network is characterized both by PageRank $K(i)$ and CheiRank $K^*(i)$ indexes the ranking of nodes becomes two-dimensional. While PageRank highlights well-know popular nodes, CheiRank highlights communicative nodes. As discussed in [5,12,15], such 2DRanking allows to characterize an information flow on networks in a more efficient and rich manner. It is convenient to characterize the interdependence between PageRank and CheiRank vectors by the correlator

$$\kappa = N \sum_{i=1}^{N} P(K(i)) P^*(K^*(i)) - 1. \qquad (2)$$

As it is shown in [12,15], we have $\kappa \approx 0$ for Linux Kernel network, transcription gene networks and $\kappa \approx 2-4$ for University and Wikipedia networks.

In this work we apply the Google matrix analysis developed in [5,11–15] to the Twitter 2009 network available at[1] [2]. The total size of the Google matrix is $N = 41\,652\,230$ and the number of links is $N_\ell = 1\,468\,365\,182$. This matrix size is by one-two orders of magnitude larger than those studied in [11,13,15]. The number of links per node is $\xi_\ell = N_\ell/N \approx 35$ being by a factor 1.5−3.5 larger than for Wikipedia network or Cambridge University 2006 network [15]. The matrix elements of G and G^* are shown in Figure 1 on a scale of top 200 (top panels) and 400 (middle panels) values of K (for G) and K^* (for G^*) and in a coarse grained image for the whole matrix size scale (bottom panels).

It is interesting to note that the coarse-grained image has well visible hyperbolic onion curves of high density which are similar to those found in [15] for Wikipedia and University networks. In [15] the appearance of such curves was attributed to existence of specific categories. We assume that for the Twitter network such curves are a result of enhanced links between various categories of users (e.g. actors, journalists, etc.) but a detailed origin is still to be established.

In the following sections we also compare the properties of the Twitter network with those of the Wikipedia articles network from [5]. Some spectral properties of the Wikipedia network with $N = 3\,282\,257$ nodes and $N_\ell = 71\,012\,307$ links are analyzed in [11,15]. We also compare certain parameters with the networks of Cambridge and Oxford Universities of 2006 with $N = 212\,710$ and $N = 200\,823$ nodes and with $N_\ell = 2\,015\,265$ and $N_\ell = 1\,831\,542$ links respectively. The properties of these networks are discussed in [11,15]. The gallery of the Google matrix G images for these networks, as well as for the Linux Kernel network, are presented in [15]. The comparison with the data shown in Figure 1 here shows that for the Twitter network we have much stronger interconnection matrix at moderate K values. We return to this point in Sections 4 and 5.

3 Spectrum and eigenstates of Twitter

To obtain the spectrum of the Google matrix of Twitter we use the Arnoldi method [8–10]. However, at first, following the approach developed in [11], we determine the invariant subspaces of the Twitter network. For that for each node we find iteratively the set of nodes that can be reached by a chain of non-zero matrix elements of S. Usually, there are several such invariant isolated subsets and the size of such subsets is smaller than the whole matrix size. These subsets are invariant with respect to applications of matrix S. We merge all subspaces with common members, and obtain a sequence of disjoint subspaces V_j of dimension d_j invariant by applications of S. The remaining part of nodes forms the wholly connected *core space*. Such a classification scheme can be efficiently implemented in a computer program, it provides a subdivision of network nodes in N_c core space nodes (typically 70−80% of N for British University networks [11]) and N_s subspace nodes belonging to at least one of the invariant subspaces V_j inducing the block triangular structure,

$$S = \begin{pmatrix} S_{ss} & S_{sc} \\ 0 & S_{cc} \end{pmatrix}. \qquad (3)$$

Here the subspace-subspace block S_{ss} is actually composed of many diagonal blocks for each of the invariant subspaces. Each of these blocks corresponds to a column sum normalized matrix of the same type as G and has

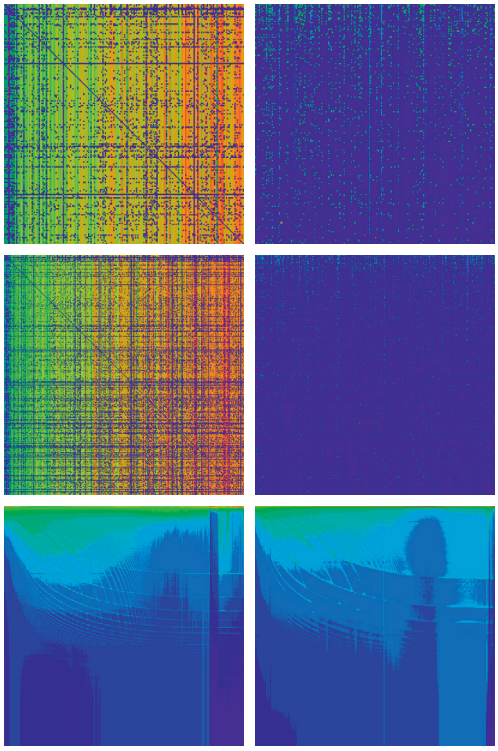


Fig. 1. (Color online) Google matrix of Twitter: matrix elements of G (left column) and G^* (right column) are shown in the basis of PageRank index K (and K') of matrix $G_{KK'}$ (left column panels) and in the basis of CheiRank index K^* (and $K^{*'}$) of matrix $G^*{}_{K^*K^{*'}}$ (right column panels). Here, x (and y) axis shows K (and $K^{'}$) (left column) (and respectively K^* and $K^{*'}$ on right column) with the range $1 \leq K, K' \leq 200$ (top panels); $1 \leq K, K' \leq 400$ (middle panels); $1 \leq K, K' \leq N$ (bottom panels). All nodes are ordered by PageRank index K of the matrix G and thus we have two matrix indexes K, K' for matrix elements in this basis (left column) and respectively $K^*, K^{*'}$ for matrix G^* (right column). Bottom panels show the coarse-grained density of matrix elements $G_{K,K'}$ and $G^*_{K^*K^{*'}}$; the coarse graining is done on 500×500 square cells for the entire Twitter network. We use a standard matrix representation with $K = K' = 1$ on top left panel corner (left column) and respectively $K^* = K^{*'} = 1$ (right column). Color shows the amplitude of matrix elements in top and middle panels or their density in the bottom panels changing from blue for minimum zero value to red at maximum value. Here the PageRank index K (and CheiRank index K^*) has been calculated for the damping factor $\alpha = 0.85$. However, the matrix elements G are shown for the damping factor $\alpha = 1$ since a value $\alpha < 1$ only adds a uniform background value and modifies the overall scale in the density plots.

therefore at least one unit eigenvalue thus explaining the high degeneracy. Its eigenvalues and eigenvectors are easily accessible by numerical diagonalization (for full matrices) thus allowing to count the number of unit eigenvalues.

We find for the G matrix of Twitter 2009 that there are $N_s = 40\,307$ subset sites with a maximal subspace dimension of 44 (most subspaces are of dimension 2 or 3). For the matrix G^* we find $N_s = 180\,414$ also with a lot of subspaces of dimension 2 or 3 and a maximal subspace dimension of 2959. The remaining eigenvalues of S can be obtained from the projected core block S_{cc} which is not column sum normalized (due to non-zero matrix elements in the block S_{sc}) and has therefore eigenvalues strictly inside the unit circle $|\lambda_j^{(\mathrm{core})}| < 1$. We have applied the Arnoldi method (AM) [8–10] with Arnoldi dimension $n_A = 640$ to determine the largest eigenvalues of S_{cc} which required a machine with 250 GB of physical RAM memory to store the non-zero matrix elements of S and the 640 vectors of the Krylov space.

In general the Arnoldi method provides numerically accurate values for the largest eigenvalues (in modulus) but their number depends crucially on the Arnoldi dimension. In our case there is a considerable density of real eigenvalues close to the points 1 and -1 where convergence is rather difficult. Comparing the results for different values of n_A, we find that for the matrix S (S^*) the first 200 (150) eigenvalues are correct within a relative error below 0.3% while the marjority of the remaining eigenvalues with $|\lambda_j| \geq 0.5$ ($|\lambda_j| \geq 0.6$) have a relative error of 10%. However, the well isolated complex eigenvalues, well visible in Figure 2, converge much better and are numerically accurate (with an error $\sim 10^{-14}$). The first three core space eigenvalues of S (S^*) are also numerically acurrate with an error of $\sim 10^{-14}$ ($\sim 10^{-8}$).

The composed spectrum of subspaces and core space eigenvalues obtained by the Arnoldi method is shown in Figure 2 for G and G^*. The obtained results show that the fraction of invariant subspaces with $\lambda = 1$ ($g_1 = N_s/N \approx 10^{-3}$) is by orders of magnitude smaller than the one found for British Universities ($g_1 \approx 0.2$ at $N \approx 2 \times 10^5$) [11]. We note that the cross and triple-star structures are visible for Twitter spectrum in Figure 2 but they are significantly less pronounced as compared to the case of Cambridge and Oxford network spectrum (see Fig. 2 in [11]). It is interesting that such a triplet and cross structures naturally appear in the spectra of random unistochastic matrices of size $N = 3$ and 4 which have been analyzed analytically and numerically in [17]. A similar star-structure spectrum appears also in sparse regular graphs with loops studied recently in [18] even if in the later case the spectrum goes outside of unit circle. This shows that even in large size networks the loop structure between 3 or 4 dominant types of nodes is well visible for University networks. For Twitter network it is less pronounced probably due to a larger number ξ_ℓ of links per node. At the same time a circle structure in the spectrum remains well visible both for Twitter and University networks. The integrated number of eigenvalues as a function of $|\lambda|$ is shown in the bottom panels of Figure 2. Further detailed analysis is required for a better understanding of the origin of such spectral structures.

It is interesting to note that a circular structure, formed by eigenvalues λ_i with $|\lambda_i|$ being close to unity (see red and blue point in top left and right panels of Fig. 3), is rather similar to those appearing in the Ulam networks of intermittency maps (see Fig. 4 in [19]). Following an analogy with the dynamics of these one-dimensional maps

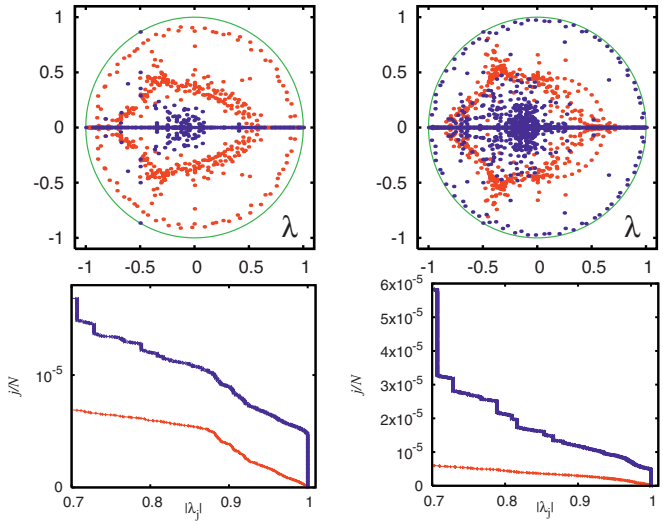


Fig. 2. (Color online) Spectrum of the Twitter matrix S (S^* with inverted direction of links) for the Twitter network shown on left panels (right panels). Top panel: subspace eigenvalues (blue dots) and core space eigenvalues (red dots) in λ-plane (green curve shows unit circle); there are 17 504 (66 316) invariant subspaces, with maximal dimension 44 (2959) and the sum of all subspace dimensions is $N_s = 40\,307$ (180 414). The core space eigenvalues are obtained from the Arnoldi method applied to the core space subblock S_{cc} of S with Arnoldi dimension 640 as explained in reference [11]. Bottom panels: fraction j/N of eigenvalues with $|\lambda| > |\lambda_j|$ for the core space eigenvalues (red bottom curve) and all eigenvalues (blue top curve) from raw data of top panels. The number of eigenvalues with $|\lambda_j| = 1$ is 34135 (129 185) of which 17505 (66 357) 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 the bottom panels 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 number of those eigenvalues corresponds to $j/N = 8.195 \times 10^{-4}$ (3.102×10^{-3}) which is strongly outside the vertical panel scale.

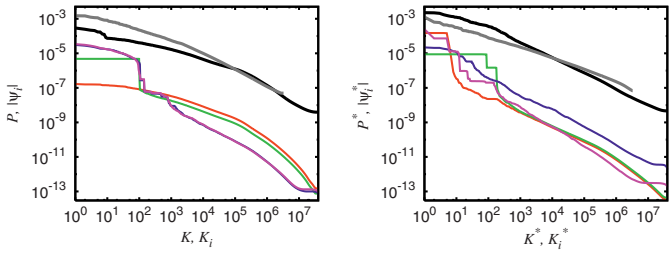


Fig. 3. (Color online) The left (right) panel shows the PageRank P (CheiRank P^*) versus the corresponding rank index K (K^*) for the Google matrix of Twitter at the damping parameter $\alpha = 0.85$ (thick black curve); for comparison the PageRank (CheiRank) of the Google matrix of Wikipedia network [5] is shown by the gray curve at same α. The colored thin curves (shifted down by factor 1000 for clarity) show the modulus of four core space eigenvectors $|\psi_i|$ ($|\psi_i^*|$) of S (S^*) versus their own ranking indexes K_i (K_i^*). Red and green lines are the eigenvectors corresponding to the two largest core space eigenvalues (in modulus) $\lambda_1 = 0.99997358$, $\lambda_2 = 0.99932634$ ($\lambda_1 = 0.99997002$, $\lambda_2 = 0.99994658$); blue and pink lines are the eigenvectors corresponding to the two complex eigenvalues $\lambda_{151} = 0.09032572 + i\,0.90000530$, $\lambda_{161} = -0.47504961 + i\,0.76576321$ ($\lambda_{457} = 0.38070896 + i\,0.39207668$, $\lambda_{105} = -0.45794117 + i\,0.80825210$). Eigenvalues and eigenvectors are obtained by the Arnoldi method with Arnoldi dimension 640 as for the data in Figure 2.

we may say that the eigenstates related to such a circular structure corresponds to quasi-isolated communities, being similar to orbits in a vicinity of intermittency region, where the information circulates mainly inside the community with only a very little flow outside of it.

The eigenstates of G and G^* with $|\lambda|$ being unity or close to unity are shown in Figure 3. For the PageRank P (CheiRank P^*) we compare its dependence on the corresponding index K (K^*) with the PageRank (CheiRank) of the Wikipedia network analyzed in [5,11,15] which size N (number of links N_ℓ) is by a factor of 10 (20) smaller. Surprisingly we find that the PageRank $P(K)$ of Twitter, approximated by the algebraic decay $P(K) = a/K^\beta$, has a slower drop as compared to Wikipedia case. Indeed, we have $\beta = 0.540 \pm 0.004$ ($a = 0.00054 \pm 0.00002$) for the PageRank of Twitter in the range $1 \leq \log_{10} K \leq 6$ (similar value as in [20] for the range $\log_{10} K \leq 5.5$) while we have $\beta = 0.767 \pm 0.0005$ ($a = 0.0086 \pm 0.00035$) for the same range of PageRank of Wikipedia network. Also we have a sharper drop of CheiRank with $\beta = 0.857 \pm 0.003$ ($a = 0.0148 \pm 0.0004$) compared to those of PageRank of Twitter while for CheiRank of Wikipedia network we find an opposite tendency ($\beta = 0.620 \pm 0.001, a = 0.0015 \pm 0.00002$) in the same index range. Thus for Twitter network the PageRank is more delocalized compared to CheiRank (e.g. $P(1) < P^*(1)$) while usually one has the opposite relation (e.g. for Wikipedia $P(1) > P^*(1)$). We attribute this to the enormously high inter-connectivity between the top PageRank nodes $K \leq 10^4$ which is well visible in Figure 1.

We should also point out a specific property of PageRank and CheiRank vectors which has been already noted in [21]: there are some degenerate plateaus in $P(K(i))$ or $P^*(K^*(i))$ with absolutely the same values of P or P^* for a few nodes. For example, for the Twitter network we have the appearance of the first degenerate plateau at $P = 7.639 \times 10^{-7}$ for $196489 \leq K \leq 196491$. As a result the PageRank index K can be ordered in various ways. We attribute this phenomenon to the fact that the matrix elements of G are composed from rational elements that leads to such type of degeneracy. However, the sizes of such degenerate plateaus are relatively short and they do not influence significantly the PageRank order. Indeed, on large scales the curves of $P(K)$, $P^*(K^*)$ are rather smooth being characterized by a finite slope (see Fig. 3). Similar type of degenerate plateaus exits for networks of Wikipedia, Cambridge and Oxford Universities.

Other eigenvectors of G and G^* of Twitter network are shown by color curves in Figure 3. We see that the shape of eigenstates with λ_1 and λ_2, shown as a function of their monotonic decrease index K_i, is well pronounced in $P(K)$. Indeed, these vectors have a rather small gap separating

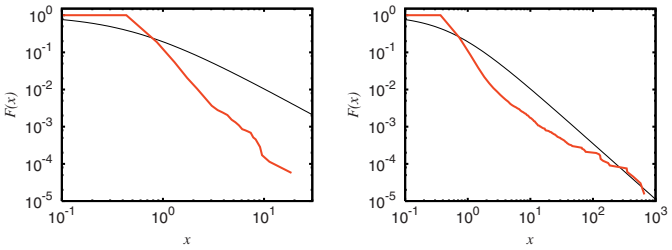


Fig. 4. (Color online) Fraction of invariant subspaces F with dimensions larger than d as a function of the rescaled variable $x = d/\langle d\rangle$, where $\langle d\rangle$ is the average subspace dimension. Left (right) panel corresponds to the matrix S (S^*) for the Twitter network (thick red curve) with $\langle d\rangle = 2.30$ (2.72). The tail can be fitted for $x \geq 0.5$ ($x \geq 10$) by the power law $F(x) = a/x^b$ with $a = 0.092 \pm 0.011$ and $b = 2.60 \pm 0.07$ ($a = 0.0125 \pm 0.0008$ and $b = 0.94 \pm 0.02$). The thin black line is $F(x) = (1+2x)^{-1.5}$ which corresponds to the universal behavior of $F(x)$ found in reference [11] for the WWW of British university networks.

them from unity ($|\Delta\lambda| \sim 2 \times 10^{-5}$) and thus they significantly contribute to the PageRank at $\alpha = 0.85$. At the same time we note that the gap values are significantly smaller than those for certain British Universities (see e.g. Fig. 4 in [11]). We argue that a larger number of links ξ_ℓ for Twitter is at the origin of moderate spectral gap between the core space spectrum and $\lambda = 1$. The eigenvectors of G^* have less slope variations and their decay is rather similar to the decay of CheiRank vector $P^*(K^*)$.

Finally, in Figure 4 we use the approach developed in [11] and analyze the dependence of the fraction of invariant subspaces $F(x)$ with dimensions larger than d on the rescaled variable $x = d/\langle d\rangle$ where $\langle d\rangle$ is the average subspace dimension. In [11] it was found that the British University networks are characterized by a universal functional distribution $F(x) = 1/(1+2x)^{3/2}$. For the Twitter network we find significant deviations from such a dependence as it is well seen in Figure 4. The tail can be fitted by the power law $F(x) \sim x^{-b}$ with the exponent $b = 2.60$ for G and $b = 0.94$ for G^*. It seems that with the increase of number of links per node ξ_ℓ we start to see deviations from the above universal distribution: it is visible for Wikipedia network (see Fig. 7 in [11]) and becomes even more pronounced for the Twitter network. We assume that a large value of ξ_ℓ for Twitter leads to a change of the percolation properties of the network generating other type of distribution F which properties should be studied in more detail in further.

4 CheiRank versus PageRank of Twitter

As discussed in [5,12,15] each network node i has its own PageRank index $K(i)$ and CheiRank index $K^*(i)$ and, hence, the ranking of network nodes becomes a two-dimensional (2DRanking). The distribution of Twitter nodes in the PageRank-CheiRank plane (K, K^*) is shown in Figure 5 (left column) in comparison to the case of the Wikipedia network from [5,15] (right column). There are much more nodes inside the square of size $K, K^* \leq 1000$

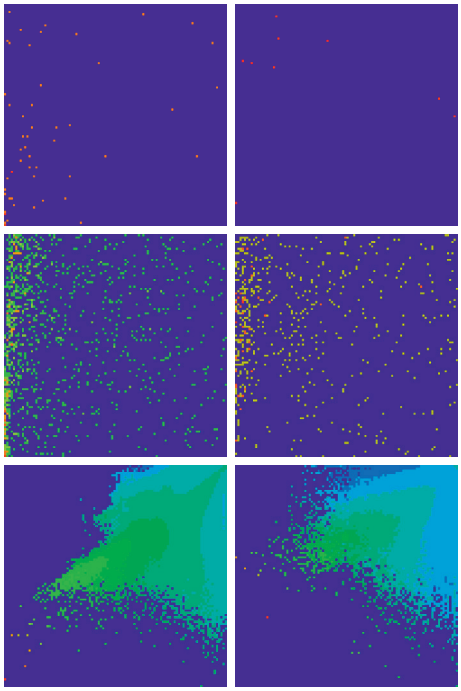


Fig. 5. (Color online) Density of nodes $W(K, K^*)$ on PageRank-CheiRank plane (K, K^*) for Twitter (left panels) and Wikipedia (right panels). Top panels show density in the range $1 \leq K, K^* \leq 1000$ with averaging over cells of size 10×10; middle panels show the range $1 \leq K, K^* \leq 10^4$ with averaging over cells of size 100×100; bottom panels show density averaged over 100×100 logarithmically equidistant grids for $0 \leq \ln K, \ln K^* \leq \ln N$, 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 blue at zero value to red at maximal density value. At each panel the x-axis corresponds to K (or $\ln K$ for the bottom panels) and the y-axis to K^* (or $\ln K^*$ for the bottom panels).

for Twitter as compared to the case of Wikipedia. For the squares of larger sizes the densities become comparable. The global logarithmic density distribution is shown in the bottom panels of Figure 5 for both networks. The two densities have certain similarities in their distributions: both have a maximal density along a certain ridge along a line $\ln K^* = \ln K + \text{const}$. However, for the Twitter network we have a significantly larger number of nodes at small values $K, K^* < 1000$ while in the Wikipedia network this area is practically empty.

The striking difference between the Twitter and Wikipedia networks is in the number of points N_K, located inside a square area of size $K \times K$ in the PageRank-CheiRank plane. This is directly illustrated in Figure 6: at $K = 500$ there are 40 times more nodes for Twitter, at $K = 1000$ we have this ratio around 6. We note that a similar dependence N_K was studied in [15] for Wikipedia, British Universities and Linux Kernel networks (see Fig. 8 there), where in all cases the initial growth of N_K was significantly smaller as compared to the Twitter network considered here.

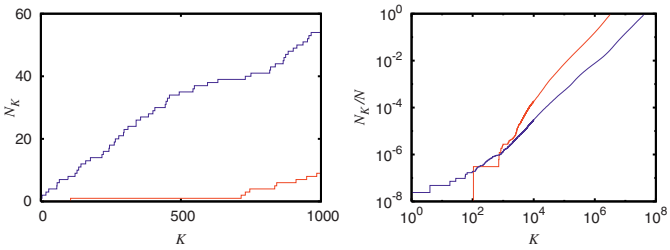


Fig. 6. (Color online) Dependence of number of nodes N_K, counted inside the square of size $K \times K$ on PageRank-CheiRank plane, on K for Twitter (blue curve) and Wikipedia (red curve); left panel shows data for $1 \leq K \leq 1000$ in linear scale, right panel shows data in log-log scale for the whole range of K.

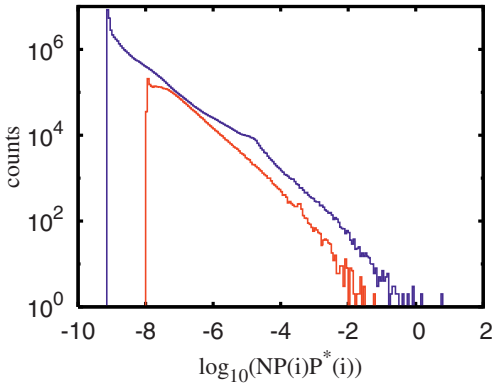


Fig. 7. (Color online) Histogram of frequency appearance of correlator components $\kappa_i = NP(K(i))P^*(K^*(i))$ for networks of Twitter (blue) and Wikipedia (red). For the histogram the whole interval $10^{-10} \leq \kappa_i \leq 10^2$ is divided in 240 cells of equal size in logarithmic scale.

Another important characteristics of 2DRanking is the correlator κ (2) between PageRank and CheiRank vectors. We find for Twitter the value $\kappa = 112.60$ which is by a factor 30−60 larger compared to this value for Wikipedia (4.08), Cambridge and Oxford University networks of 2006 considered in [5,11,15]. The origin of such a large value of κ for the Twitter network becomes more clear from the analysis of the distribution of individual node contributions $\kappa_i = NP(K(i))P^*(K^*(i))$ in the correlator sum (2) shown in Figure 7. We see that there are certain nodes with very large κ_i values and even if there are only few of them still they give a significant contribution to the total correlator value. We note that there is a similar feature for the Cambridge University network in 2011 as discussed in [15] even if there one finds a smaller value $\kappa = 30$. Thus we see that for certain nodes we have strongly correlated large values of $P(K(i))$ and $P^*(K^*(i))$ explaining the largest correlator value κ among all networks studied up to now. We will argue below that this is related to a very strong inter-connectivity between top K PageRank users of the Twitter network.

5 Discussion

In this work we study the statistical properties of the Google matrix of Twitter network including its spectrum,

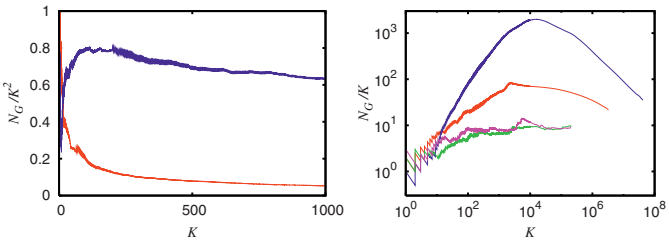


Fig. 8. (Color online) Left panel: dependence of the area density $g_K = N_G/K^2$ of nonzero elements of the adjacency matrix among top PageRank nodes on the PageRank index K for Twitter (blue curve) and Wikipedia (red curve) networks, data are shown in linear scale. Right panel: linear density N_G/K of 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 $K = 100$).

eigenstates and 2DRanking of PageRank and CheiRank vectors. The comparison with Wikipedia shows that for Twitter we have much stronger correlations between PageRank and CheiRank vectors. Thus for the Twitter network there are nodes which are very well known by the community of users and at the same time they are very communicative being strongly connected with top PageRank nodes. We attribute the origin of this phenomenon to a very strong connectivity between top K nodes for Twitter as compared to the Wikipedia network. This property is illustrated in Figure 8 where we show the number of nonzero elements N_G of the Google matrix, taken at $\alpha = 1$ and counted in the top left corner with indexes being smaller or equal to K (elements in columns of dangling nodes are not taken into account). We see that for $K \leq 1000$ we have for Twitter the 2D density of nonzero elements to be on a level of 70% while for Wikipedia this density is by a factor 10 smaller. For these two networks the dependence of N_G on K at $K \leq 1000$ is well described by a power law $N_G = aN^b$ with $a = 0.72 \pm 0.01$, $b = 1.993 \pm 0.002$ for Twitter and $a = 2.10 \pm 0.01$, $b = 1.469 \pm 0.001$ for Wikipedia. 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. An effective number of links per node for top K nodes is given by the ratio N_G/K which is equal to ξ_ℓ at $K = N$. The dependence of this ratio on K is shown in Figure 8 in right panel. We see a striking difference between Twitter network and networks of Wikipedia, Cambridge and Oxford Universities. For Twitter the maximum value of N_G/K is by two orders of magnitude larger as compared to the Universities networks, and by a factor 20 larger than for Wikipedia. Thus the Twitter network is characterized by a very strong connectivity between top PageRank nodes which can be considered as the Twitter elite [20].

It is interesting to note that for $K \leq 20$ the Wikipedia network has a larger value of the ratio N_G/K^2 compared to the Twitter network, but the situation is changed for larger values of $K > 20$. In fact the first top 20 nodes of Wikipedia network are mainly composed from

world countries (see [5]) which are strongly interconnected due to historical reasons. However, at larger values of K Wikipedia starts to have articles on various subjects and the ratio N_G/K^2 drops significantly. On the other hand, for the Twitter network we see that a large group of very important persons (VIP) with $K < 10^4$ is strongly interconnected. This dominant VIP structure has certain similarities with the structure of transnational corporations and their ownership network dominated by a small tightly-knit core of financial institutions [22]. The existence of a solid phase of industrially developed, strongly linked countries is also established for the world trade network obtained from the United Nations COMTRADE data base [23]. It is possible that such super concentration of links between top Twitter users results from a global increase of number of links per node characteristic for such type of social networks. Indeed, the recent analysis of the Facebook network shows a significant decrease of degree of separation during the time evolution of this network [24]. Also the number of friendship links per node reaches as high value as $\xi_\ell \approx 100$ at the current Facebook snapshot (see Tab. 2 in [24]). This significant growth of ξ_ℓ during the time evolution of social networks leads to an enormous concentration of links among society elite at top PageRank users and may significantly influence the process of strategic decisions on such networks in the future. The growth of ξ_ℓ leads also to a significant decrease of the exponent β of algebraic decay of PageRank which is known to be $\beta \approx 0.9$ for the WWW (see e.g. [3,4,7]) while for the Twitter network we find $\beta \approx 0.5$ (see also [20]). This tendency may be a precursor of a delocalization transition of the PageRank vector emerging at a large values of ξ_ℓ. Such a delocalization would lead to a flat PageRank probability distribution and a strong drop of the efficiency of the information retrieval process. It is known that for the Ulam networks of dynamical maps such a delocalization indeed takes place under certain conditions [19,25].

Our results show that the strong inter-connectivity of VIP users with about top 1000 PageRank indexes dominates the information flow on the network. This result is in line with the recent studies of opinion formation of the Twitter network [20] showing that the top 1300 PageRank users of Twitter can impose their opinion for the whole network of 41 million size. Thus we think that the statistical analysis presented here plays a very important role for a better understanding of decision making and opinion formation on the modern social networks.

The present size of the Twitter network is by a factor 3.5 larger as compared to its size in 2009 analyzed in this work. Thus it would be very interesting to extend the present analysis to the current status of the Twitter network which now includes all layers of the world society. Such an analysis will allow to understand in an better way the process of information flow and decision making on social networks.

This work is supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No. 288956). We thank S.Vigna for providing us a friendly access to the Twitter dataset[1] [2]. We also acknowledge the France-Armenia collaboration grant CNRS/SCS No. 24943 (IE-017) on "Classical and quantum chaos".

References

1. Wikipedia (The Free Encyclopedia) Twitter, `http://en.wikipedia.org/wiki/Twitter` (2012)
2. H. Kwak, C. Lee, H. Park, S. Moon, *Proc. 19th Int. Conf. WWW2010* (ACM, New York, 2010), p. 591
3. D. Donato, L. Laura, S. Leonardi, S. Millozzi, Eur. Phys. J. B **38**, 239 (2004)
4. G. Pandurangan, P. Raghavan, E. Upfal, Internet Math. **3**, 1 (2005)
5. A.O. Zhirov, O.V. Zhirov, D.L. Shepelyansky, Eur. Phys. J. B **77**, 523 (2010)
6. S. Brin, L. Page, Comput. Netw. ISDN Syst. **30**, 107 (1998)
7. A.M. Langville, C.D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings* (Princeton University Press, Princeton, 2006)
8. G.W. Stewart, *Matrix Algorithms Eigensystems* (SIAM, 2001), Vol. II
9. G.H. Golub, C. Greif, BIT Num. Math. **46**, 759 (2006)
10. K.M. Frahm, D.L. Shepelyansky, Eur. Phys. J. B **76**, 57 (2010)
11. K.M. Frahm, B. Georgeot, D.L. Shepelyansky, J. Phys. A: Math. Theor. **44**, 465101 (2011)
12. A.D. Chepelianskii, `arXiv:1003.5455[cs.SE]` (2010)
13. L. Ermann, A.D. Chepelianskii, D.L. Shepelyansky, Eur. Phys. J. B **79**, 115 (2011)
14. L. Ermann, D.L. Shepelyansky, Acta Phys. Polonica A **120**, A158 (2011)
15. L. Ermann, A.D. Chepelianskii, D.L. Shepelyansky, J. Phys. A: Math. Theor. **45**, 275101 (2012)
16. N. Litvak, W.R.W. Scheinhardt, Y. Volkovich, Lect. Notes Comput. Sci. **4936**, 72 (2008)
17. K. Zyczkowski, M. Kus, W. Slomczynski, H.-J. Sommers, J. Phys. A: Math. Gen. **36**, 3425 (2003)
18. F.L. Metz, I. Neri, D. Bolle, Phys. Rev. E **84**, 055101(R) (2011)
19. L. Ermann, D.L. Shepelyansky, Phys. Rev. E **81**, 036221 (2010)
20. V. Kandiah, D.L. Shepelyansky, Physica A **391**, 5779 (2012)
21. K.M. Frahm, A.D. Chepelianskii, D.L. Shepelyansky, J. Phys. A: Math. Theor. **45**, 405101 (2012)
22. S. Vitali, J.B. Glattfelder, S. Battiston, PLoS ONE **6**, e25995 (2011)
23. L. Ermann, D.L. Shepelyansky, Acta Phys. Polonica A **120**, A158 (2011)
24. L. Backstrom, P. Boldi, M. Rosa, J. Ugander, S. Vigna, `arXiv:1111.4570v3 [cs.SI]` (2012)
25. D.L. Shepelyansky, O.V. Zhirov, Phys. Rev. E **81**, 036213 (2010)

EPJ manuscript No.
(will be inserted by the editor)

Time evolution of Wikipedia network ranking

Young-Ho Eom[1], Klaus M. Frahm[1], András Benczúr[2], and Dima L. Shepelyansky[1]

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

[2] Informatics Laboratory, Institute for Computer Science and Control,
Hungarian Academy of Sciences (MTA SZTAKI), Pf. 63, H-1518 Budapest, Hungary

Received: April 24, 2013

Abstract. We study the time evolution of ranking and spectral properties of the Google matrix of English Wikipedia hyperlink network during years 2003 - 2011. The statistical properties of ranking of Wikipedia articles via PageRank and CheiRank probabilities, as well as the matrix spectrum, are shown to be stabilized for 2007 - 2011. A special emphasis is done on ranking of Wikipedia personalities and universities. We show that PageRank selection is dominated by politicians while 2DRank, which combines PageRank and CheiRank, gives more accent on personalities of arts. The Wikipedia PageRank of universities recovers 80 percents of top universities of Shanghai ranking during the considered time period.

PACS. 89.75.Fb Structures and organization in complex systems – 89.75.Hc Networks and genealogical trees – 89.20.Hh World Wide Web, Internet

1 Introduction

At present Wikipedia [1] became the world largest Encyclopedia with open public access to its contain. A recent review [2] represents a detailed description of publications and scientific research of this modern Library of Babel, which stores an enormous amount of information, approaching the one described by Jorge Luis Borges [3]. The hyperlinks of citations between Wikipedia articles represent a directed network which reminds the structure of the World Wide Web (WWW). Hence, the mathematical tools developed for WWW search engines, based on the Markov chains [4], Perron-Frobenius operators [5] and the PageRank algorithm of the corresponding Google matrix [6,7], give solid mathematical grounds for analysis of information flow on the Wikipedia network. In this work we perform the Google matrix analysis of Wikipedia network of English articles extending the results presented in [8,9],[10,11]. The main new element of this work is the study of time evolution of Wikipedia network during the years 2003 to 2011. We analyze how the ranking of Wikipedia articles and the spectrum of the Google matrix G of Wikipedia are changed during this period.

The directed network of Wikipedia articles is constructed in a usual way: a directed link is formed from an article j to an article 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 ($\sum_j S_{ij} = 1$) 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 [6,7]:

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

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 $\alpha \approx 0.85$ [7]. The matrix G belongs to the class of Perron-Frobenius operators [5,7], 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 nonnegative elements $P(i)$ and gives a probability $P(i)$ to find a random surfer at site i. It is possible to 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 their ranking, placing the most popular articles or nodes at the top values $K = 1, 2, 3....$

Due to the gap $1-\alpha \approx 0.15$ between the largest eigenvalue $\lambda = 1$ and other eigenvalues the PageRank algorithm permits an efficient and simple determination of the PageRank by the power iteration method [7]. It is also possible to use the powerful Arnoldi method [12,13],[14] to compute efficiently the eigenspectrum λ_i of the Google matrix:

$$\sum_{k=1}^{N} G_{jk}\psi_i(k) = \lambda_i \psi_i(j) \ . \qquad (2)$$

The Arnoldi method allows to find a several thousands of eigenvalues λ_i with maximal $|\lambda|$ for a matrix size N as large as a few tens of millions [10,11], [14,15]. Usually,

at $\alpha = 1$ the largest eigenvalue $\lambda = 1$ is highly degenerate [15] due to many invariant subspaces which define many independent Perron-Frobenius operators providing (at least) one eigenvalue $\lambda = 1$.

In addition to a given directed network A_{ij} it is useful to analyze an inverse network with inverted direction of links with elements of adjacency matrix $A_{ij} \to A_{ji}$. The Google matrix G^* of the inverse network is then constructed via corresponding matrix S^* according to the relations (1) using the same value of α as for the G matrix. This time inversion approach was used in [16,17] but the statistical properties and correlations between direct and inversed ranking were not analyzed there. In [18], on an example of the Linux Kernel network, it was shown thus this approach allows to obtain an additional interesting characterization of information flow on directed networks. Indeed, the right eigenvector of G^* at eigenvalue $\lambda = 1$ gives a probability $P^*(i)$, called CheiRank vector [8]. It determines a complementary rank index $K^*(i)$ of network nodes in a decreasing order of probability $P^*(K^*(i))$ [8, 9],[10,18]. It is known that the PageRank probability is proportional to the number of ingoing links characterizing how popular or known is a given node. In a similar way the CheiRank probability is proportional to the number of outgoing links highlighting the node communicativity (see e.g. [7,19], [20,21],[8,9]). The statistical properties of distribution of indexes $K(i), K^*(i)$ on the PageRank-CheiRank plane are described in [9].

In this work we apply the above mathematical methods to the analysis of time evolution of Wikipedia network ranking using English Wikipedia snapshots dated by December 31 of years 2003, 2005, 2007, 2009, 2011. In addition we use the snapshot of August 2009 (200908) analyzed in [8]. The parameters of networks with the number of articles (nodes) N, number of links N_ℓ and other information are given in Tables 1,2 with the description of notations given in Appendix.

The paper is composed as following: the statistical properties of PageRank and CheiRank are analyzed in Section 2, ranking of Wikipedia personalities and universities are considered in Sections 3, 4 respectively, the properties of spectrum of Google matrix are considered in Section 5, the discussion of the results is presented in Section 6, Appendix Section 7 gives network parameters.

2 CheiRank versus PageRank

The dependencies of PageRank and CheiRank probabilities $P(K)$ and $P^*(K^*)$ on their indexes K, K^* at different years are shown in Fig. 1. The top positions of K are occupied by countries starting from *United States* while at the top positions of K^* we find various listings (e.g. geographical names, prime ministers etc.; in 2011 we have appearance of listings of listings). Indeed, the countries accumulate links from all types of human activities and nature, that make them most popular Wikipedia articles, while listings have the largest number of outgoing links making them the most communicative articles.

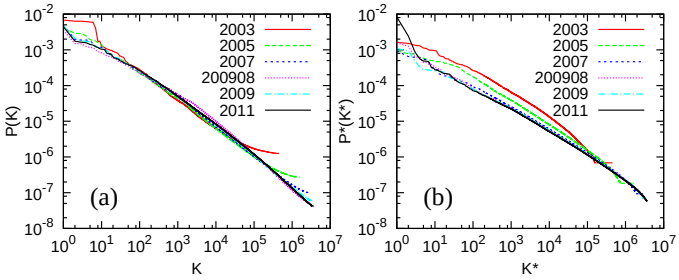


Fig. 1. PageRank probability $P(K)$ (left panel) and CheiRank probability $P^*(K^*)$ (right panel) are shown as a function of the corresponding rank indexes K and K^* for English Wikipedia articles at years 2003, 2005, 2007, 200908, 2009, 2011; here the damping factor is $\alpha = 0.85$.

The data of Fig. 1 show that the global behavior of $P(K)$ remains stable from 2007 to 2011. The probability $P^*(K^*)$ is stable in the time interval 2007 - 2009 while at 2011 we see the appearance of peak at $1 \leq K^* < 10$ that is related to introduction of listings of listings which were absent at earlier years. At the same time the behavior of $P^*(K^*)$ in the range $10 \leq K^* \leq 10^6$ remains stable for 2007 - 2011.

Each article i has its PageRank and CheiRank indexes $K(i)$, $K^*(i)$ so that all articles are distributed on two-dimensional plane of PageRank-CheiRank indexes. Following [8,9] we present the density of articles in the 2D plane (K, K^*) in Fig. 2. The density is computed for 100×100 logarithmically equidistant cells which cover the whole plane (K, K^*) for each year. The density distribution is globally stable for years 2007-2011 even if there are articles which change their location in 2D plane. We see an appearance of a mountain like ridge of probability along a line $\ln K^* \approx \ln K + 4.6$ that indicate the presence of correlation between $P(K(i))$ and $P^*(K^*(i))$. Following [8,9, 18] we characterize the interdependence of PageRank and CheiRank vectors by the correlator

$$\kappa = N \sum_{i=1}^{N} P(K(i)) P^*(K^*(i)) - 1 \ . \qquad (3)$$

We find the following values of the correlator at various time slots: $\kappa = 2.837(2003), 3.894(2005), 4.121(2007), 4.084(200908), 6.629(2009), 5.391(2011)$. During that period the size of the network increased almost by 10 times while κ increased less than 2 times. This confirms the stability of the correlator κ during the time evolution of the Wikipedia network.

In the next two Sections we analyze the time variation of ranking of personalities and universities.

3 Ranking of personalities

To analyze the time evolution of ranking of Wikipedia personalities (persons or humans) we chose the top 100 persons appearing in the ranking list of Wikipedia 200908

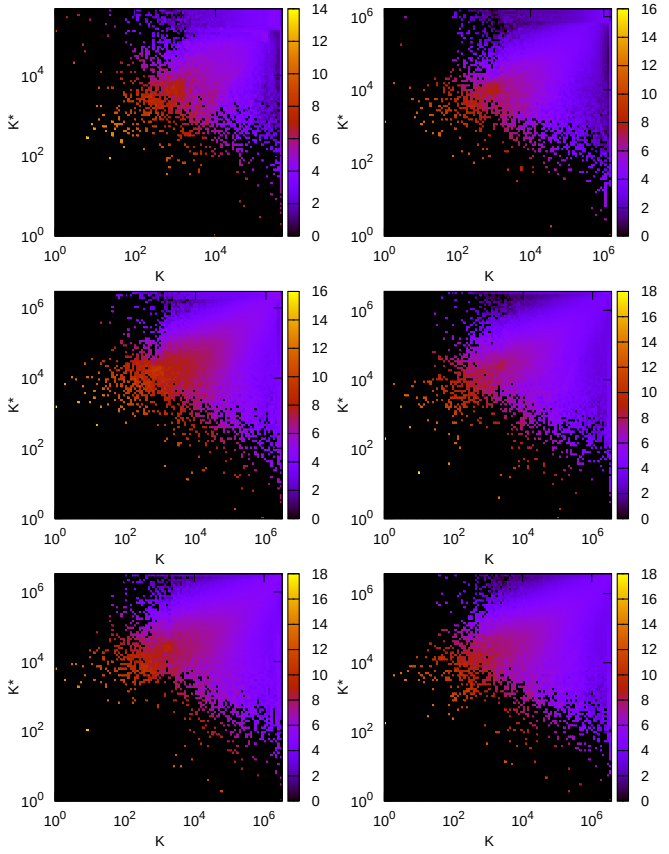


Fig. 2. Density of Wikipedia articles in the CheiRank versus PageRank plane at different years. Color is proportional to logarithm of density changing from minimal nonzero density (dark) to maximal one (white), zero density is shown by black (distribution is computed for 100×100 cells equidistant in logarithmic scale; bar shows color variation of natural logarithm of density); left column panels are for years 2003, 2007, 200908 and right column panels are for 2005, 2009, 2011 (from top to bottom).

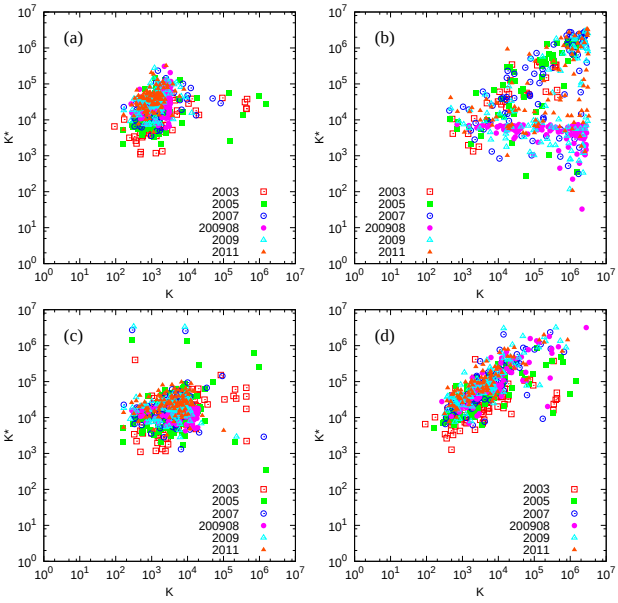


Fig. 3. Change of locations of top-rank persons of Wikipedia in K-K* plane. Each list of top ranks is determined by data of top 100 personalities of time slot 200908 in corresponding rank. Data sets are shown for (a) PageRank, (b) CheiRank, (c) 2DRank, (d) rank from Hart [22].

given in [8] in order of PageRank, CheiRank and 2DRank. We remind that 2DRank K_2 is obtained by counting nodes in order of their appearance on ribs of squares in (K, K^*) plane with their size growing from $K = 1$ to $K = N$ [8].

The distributions of personalities in PageRank-CheiRank plane is shown at various time slots in Fig. 3. There are visible fluctuations of distribution of nodes for years 2003, 2005 when the Wikipedia size has rapid growth. For other years the distribution of top 100 nodes of PageRank and 2DRank is stable even if individual nodes change their ranking. For top 100 of CheiRank the fluctuations remain strong during all years. Indeed, the number of outgoing links is more easy to be modified by authors writing a given article, while a modification of ingoing links depends on authors of other articles.

In Fig. 3 we also show the distribution of top 100 personalities from Hart's book [22] (the list of names is also available at the web page [8]). This distribution also remains stable in years 2007-2011. It is interesting to note that while top PageRank and 2DRank nodes form a kind of droplet in (K, K^*) plane, the distribution of Hart's personalities approximately follows the ridge along the line $\ln K^* \approx \ln K + 4.6$.

The time evolution of top 10 personalities of slot 200908 is shown in Fig. 4 for PageRank and 2DRank. For PageRank the main part of personalities keeps their rank position in time, e.g. G.W.Bush remains at first-second position. B.Obama significantly improves his ranking as a result of president elections. There are strong variations for Elizabeth II which we relate to modification of article name during the considered time interval. We also see a steady improvement of ranking of C.Linnaeus that we attribute to a growth of various botanic descriptions and listings at Wikipedia articles which quote his name. For 2DRank we observe stronger variations of K_2 index with time. Such a politician as R.Nixon has increasing K_2 index with time since the period of his presidency goes in the past. At the same time singers and artists remain at approximately constant level of K_2.

In [8] it was pointed out that the top personalities of PageRank are dominated by politicians while for 2DRank the dominant component of human activity is represented by artists. We analyze the time evolution of the distribution of top 30 personalities over 6 categories of human activity (*politics, arts, science, religion, sport and etc (or others)*). The category *etc* contains only C.Columbus. The results are presented in Fig. 5. They clearly show that the PageRank personalities are dominated by politicians whose percentage increases with time, while the percent of arts decreases. For 2DRank we see that the arts are dominant even if their percentage decreases with time. We also see the appearance of sport which is absent in

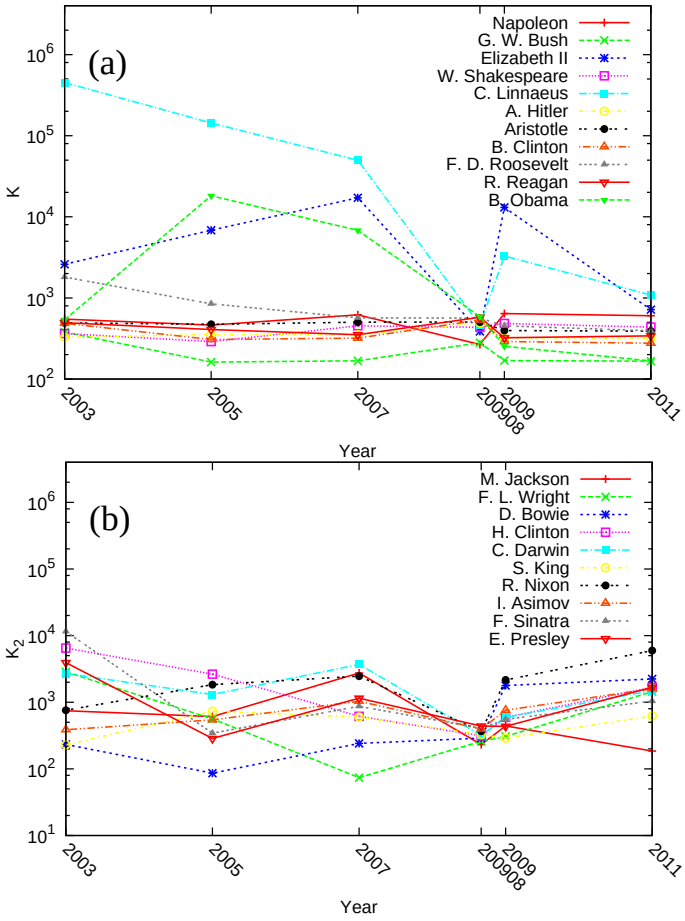


Fig. 4. Time evolution of top 10 personalities of year 200908 in indexes of PageRank K (a) and 2DRank K_2 (b); B.Obama is added in panel (a).

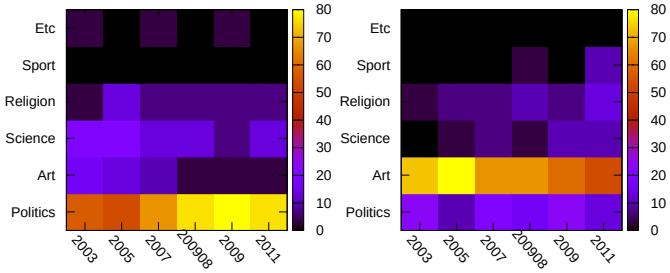


Fig. 5. Left panel: distribution of top 30 PageRank personalities over 6 activity categories at various years of Wikipedia. Right panel: distribution of top 30 2DRank personalities over the same activity categories at same years. Categories are politics, art, science, religion, sport, etc (other). Color shows the number of personalities for each activity expressed in percents.

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 increase their ranking since they are listed in various music albums, movies sport competition results. Due to that the component of arts gets higher positions in 2DRank in contrast to politics dominance in PageRank. Thus the two-dimensional ranking on PageRank-CheiRank plane allows to select qualities of nodes according to their popularity and communicativity.

4 Ranking of universities

The local ranking of top 100 universities is shown in Fig. 6 for years 2003, 2005, 2007 and in Fig. 7 for 2009, 200908, 2011. The local ranking is obtained by selecting top 100 universities appearing in PageRank listing so that they get their university ranking K from 1 to 100. The same procedure is done for CheiRank listing of universities obtaining their local CheiRank index K^* from 1 to 100. Those universities which enter inside 100×100 square on the local index plane (K, K^*) are shown in Figs. 6, 7.

The data show that the top PageRank universities are rather stable in time, e.g. U Harvard is always on the first top position. At the same time the positions in K^* are strongly changing in time. To understand the origin of this variations in CheiRank we consider the case of U Cambridge. Its Wikipedia article in 2003 is rather short but it contains the list of all 31 Colleges with direct links to their corresponding articles. This leads to a high position of U Cambridge with university $K^* = 4$ in 2003 (Fig. 8). However, with time the direct links remain only to about 10 Colleges while the whole number of Colleges are presented by a list of names without links. This leads to a significant increase of index up to $K^* \approx 40$ at Dec 2009. However, at Dec 2011 U Cambridge again improves significantly its CheiRank obtaining $K^* = 2$. The main reason of that is the appearance of section of "Notable alumni and academics" which provides direct links to articles about outstanding scientists studied and/or worked at U Cambridge that leads to second position at $K^* = 2$ among all universities. We note that in 2011 the top CheiRank University is George Mason University with university $K^* = 1$. The main reason of this high ranking is the presence of detailed listings of alumni in politics, media, sport with direct links to articles about corresponding personalities (including former director of CIA). These two examples show that the links, kept with a large number of university alumni, significantly increase CheiRank position of university. We note that artistic and politically oriented universities usually preserve more links with their alumni.

The time evolution of global ranking of top 10 universities of year 200908 for PageRank and 2DRank is shown in Fig. 8. The results show the stability of PageRank order with a clear tendency of top universities (e.g. Harvard) to go with time to higher and higher top positions of K. Thus for U Harvard the global value of K changes from $K \approx 300$ in 2003 to $K \approx 100$ in 2011, while the whole size N of the Wikipedia network increases almost by a factor 10 during this time interval. Since Wikipedia ranks all human knowledge, the stable improvement of PageRank indexes of universities reflects the global growing impor-

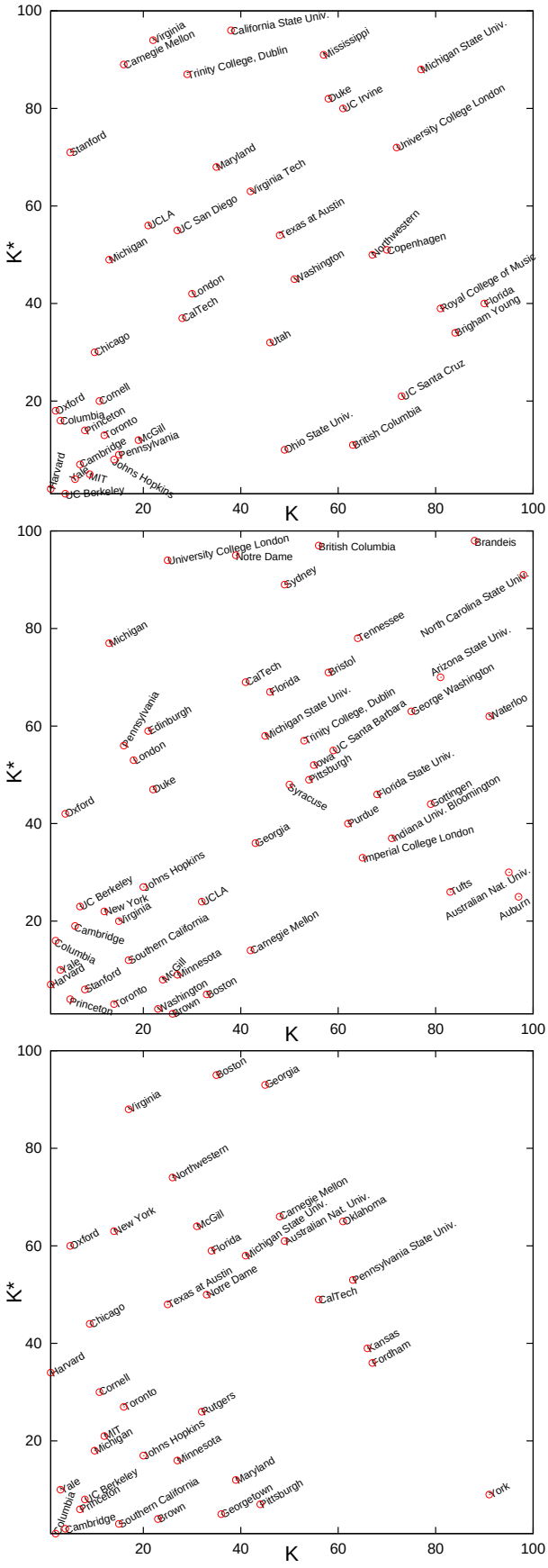


Fig. 6. University of Wikipedia articles in the local CheiRank versus PageRank plane at different years; panels are for years 2003, 2005, 2007 (from top to bottom).

Fig. 7. Same as in Fig. 6 for years 2009, 200908, 2011 (from top to bottom).

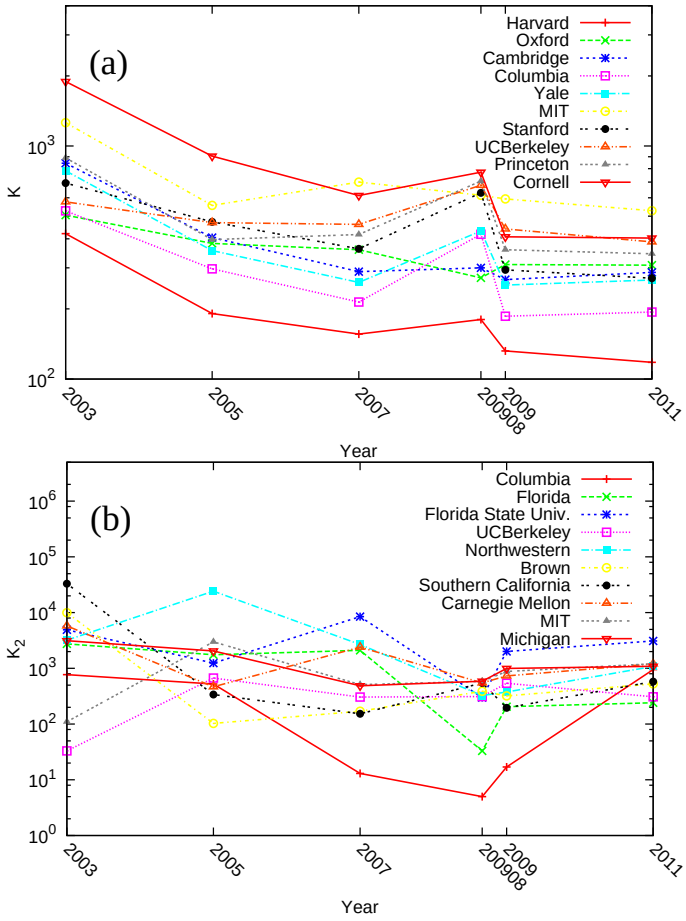


Fig. 8. Time evolution of global ranking of top 10 Universities of year 200908 in indexes of PageRank K (a) and 2DRank K_2 (b).

tance of universities in the world of human activity and knowledge.

The time evolution of the same universities in 2DRank remains stable in time showing certain interchange of their ranking order. We think that an example of U Cambridge considered above explains the main reasons of these fluctuations. In view of 10 times increase of the whole network size during the period 2003 - 2011 the average stability of 2DRank of universities also confirms the significant importance of their place in human activity.

Finally we compare the Wikipedia ranking of universities in their local PageRank index K with those of Shanghai university ranking [23]. In the top 10 of Shanghai university rank the Wikipedia PageRank recovers 9 (2003), 9 (2005), 8 (2007), 7 (2009), 7 (2011). This shows that the Wikipedia ranking of universities gives the results being very close to the real situation. A small decrease of overlap with time can be attributed to earlier launched activity of leading universities on Wikipedia.

5 Google matrix spectrum

Finally we discuss the time evolution of the spectrum of Wikipedia Google matrix taken at $\alpha = 1$. We perform the numerical diagonalization based on the Arnoldi method [12,13] using the additional improvements described in [14,15] with the Arnold dimension $n_A = 6000$. The Google matrix is reduced to the form

$$S = \begin{pmatrix} S_{ss} & S_{sc} \\ 0 & S_{cc} \end{pmatrix} \qquad (4)$$

where S_{ss} describes disjoint subspaces V_j of dimension d_j invariant by applications of S; S_{cc} depicts the remaining part of nodes forming the wholly connected *core space*. We note that S_{ss} is by itself composed of many small diagonal blocks for each invariant subspace and hence those eigenvalues can be efficiently obtained by direct ("exact") numerical diagonalization. The total subspace size N_s, the number of independent subspaces N_d, the maximal subspace dimension $d_{\max}$ and the number N_1 of S eigenvalues with $\lambda = 1$ are given in Table 2 (See also Appendix). The spectrum and eigenstates of the core space S_{cc} are determined by the Arnoldi method with Arnoldi dimension n_A giving the eigenvalues λ_i of S_{cc} with largest modulus. Here we restrict ourselves to the statistical analysis of the spectrum λ_i. The analysis of eigenstates ψ_i ($G\psi_i = \lambda_i\psi_i$), which has been done in [11] for the slot 200908, is left for future studies.

The spectrum for all Wikipedia time slots is shown in Fig. 9 for G and in Fig. 10 for G^*. We see that the spectrum remains stable for the period 2007 - 2001 even if there is a small difference of slot 200908 due to a slightly different cleaning link procedure (see Appendix). For the spectrum of G^* in 2007 - 2001 we observe a well pronounced 3-6 arrow star structure. This structure is very similar to those found in random unistochastic matrices of side 3-4 [24] (see Fig.4 therein). This fact has been pointed in [11] for the slot 200908. Now we see that this is a generic phenomenon which remains stable in time. This indicates that there are dominant groups of 3-4 nodes which have structure similar to random unistochastic matrices with strong ties between 3-4 nodes and various random permutations with random hidden complex phases. The spectral arrow star structure is significantly more pronounce for the case of G^* matrix. We attribute this to more significant fluctuations of outgoing links that probably makes sectors of G^* to be more similar to elements of unistochastic matrices. A further detailed analysis will be useful to understand these arrow star structure and its links with various communities inside Wikipedia.

As it is shown in [11] the eigenstates of G and G^* select certain well defined communities of the Wikipedia network. Such an eigenvector detection of the communities provides a new method of communities detection in addition to more standard methods developed in network science and described in [25]. However, the analysis of eigenvectors represents a separate detailed research and in this work we restrict ourselves to PageRank and CheiRank vectors.

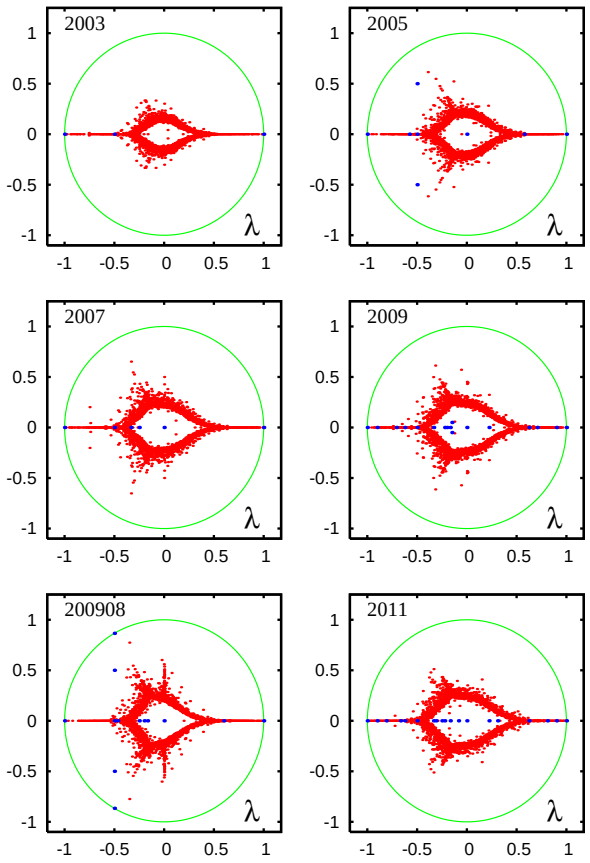


Fig. 9. Spectrum of eigenvalues λ of the Google matrix G of Wikipedia at different years. Red dots are core space eigenvalues, blue dots are subspace eigenvalues and the full green curve shows the unit circle. The core space eigenvalues were calculated by the projected Arnoldi method with Arnoldi dimensions $n_A = 6000$.

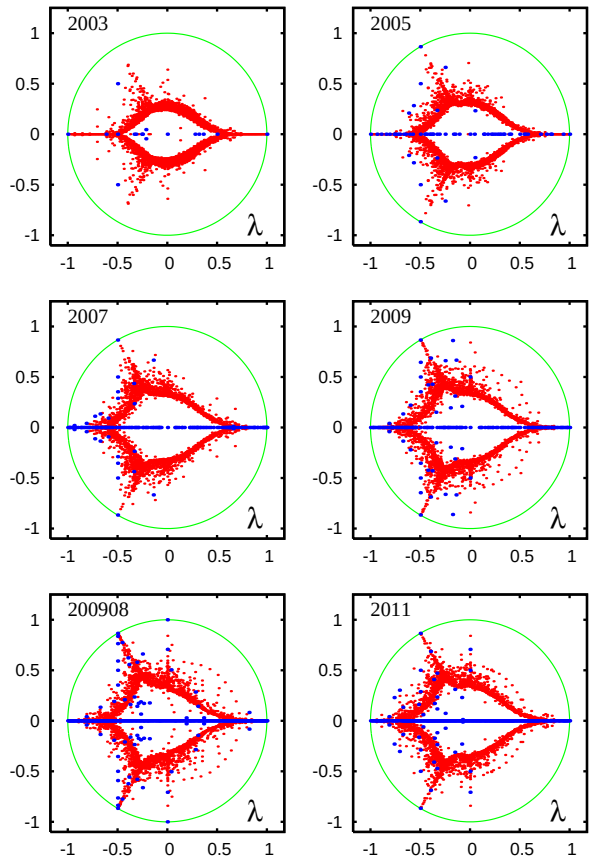


Fig. 10. Same as in Fig. 9 but for the spectrum of matrix G^*.

Finally we note that the fraction of isolated subspaces is very small for G matrix. It is increased approximately by a factor of order 10 for G^* but still it remains very small compared to the networks of UK universities analyzed in [15]. This fact reflects a strong connectivity of network of Wikipedia articles.

6 Discussion

In this work we analyzed the time evolution of ranking of network of English Wikipedia articles. Our study demonstrates the stability of such statistical properties as PageRank and CheiRank probabilities, the article density distribution in PageRank-CheiRank plane during the period 2007 - 2011. The analysis of human activities in different categories shows that PageRank gives main accent to politics while the combined 2DRank gives more importance to arts. We find that with time the number of politicians in the top positions increases. Our analysis of ranking of universities shows that on average the global ranking of top universities goes to higher and higher positions. This clearly marks the growing importance of universities for the whole range of human activities and knowledge. We find that Wikipedia PageRank recovers 70 - 80 % of top 10 universities from Shanghai ranking [23]. This confirms the reliability of Wikipedia ranking.

We also find that the spectral structure of the Wikipedia Google matrix remains stable during the time period 2007 -2011 and show that its arrow star structure reflects certain features of small size unistochastic matrices.

Acknowledgments: 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 allocations 2012-P0110, 2013-P0110. We also acknowledge the France-Armenia collaboration grant CNRS/SCS *No* 24943 (IE-017) on "Classical and quantum chaos".

7 Appendix

The tables with all network parameters used in this work are given in the text of the paper. The notations used in the tables are: N is network size, N_ℓ is the number of links, n_A is the Arnoldi dimension used for the Arnoldi method for the core space eigenvalues, N_d is the number of invariant subspaces, $d_{\max}$ gives a maximal subspace dimension, $N_{\rm circ.}$ notes number of eigenvalues on the unit

	N	N_ℓ	n_A
2003	455436	2033173	6000
2005	1635882	11569195	6000
2007	2902764	34776800	6000
2009	3484341	52846242	6000
200908	3282257	71012307	6000
2011	3721339	66454329	6000

Table 1. Parameters of all Wikipedia networks at different years considered in the paper.

	N_s	N_d	$d_{\max}$	$N_{\rm circ.}$	N_1
2003	15	7	3	11	7
2003*	940	162	60	265	163
2005	152	97	4	121	97
2005*	5966	1455	1997	2205	1458
2007	261	150	6	209	150
2007*	10234	3557	605	5858	3569
2009	285	121	8	205	121
2009*	11423	4205	134	7646	4221
200908	515	255	11	381	255
200908*	21198	5355	717	8968	5365
2011	323	131	8	222	131
2011*	14500	4637	1323	8591	4673

Table 2. G and G^* eigespectrum parameters for all Wikipedia networks, year marks spectrum of G, year with star marks spectrum of G^*.

circle with $|\lambda_i| = 1$, N_1 notes number of unit eigenvalues with $\lambda_i = 1$. We remark that $N_s \geq N_{\rm circ.} \geq N_1 \geq N_d$ and $N_s \geq d_{\max}$. The data for G are marked by the corresponding year of the time slot, the data for G^* are marked by the year with a star. Links cleaning procedure eliminates all redirects (nodes with one outgoing link), this procedure is slightly different from the one used for the slot 200908 in [8]. All data sets and high resolution figures are available at the web page [26].

References

1. Wikipedia, *Wikipedia*, `en.wikipedia.org/wiki/Wikipedia`
2. F.A. Nielsen, *Wikipedia research and tools: review and comments*, (2012), available at SSRN: `dx.doi.org/10.2139/ssrn.2129874`
3. J.L. Borges, *The Library of Babel* in *Ficciones* (Grove Press. N.Y. 1962).
4. A.A. Markov, *Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga*, Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya, **15** (1906) 135 (in Russian) [English trans.: *Extension of the limit theorems of probability theory to a sum of variables connected in a chain* reprinted in Appendix B of: R.A. Howard *Dynamic Probabilistic Systems*, volume 1: *Markov models*, Dover Publ. (2007)]
5. M. Brin and G. Stuck, *Introduction to dynamical systems*, Cambridge Univ. Press, Cambridge, UK (2002)
6. S. Brin and L. Page, Computer Networks and ISDN Systems **30**, 107 (1998)
7. A. M. Langville and C. D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton (2006)
8. A.O.Zhirov, O.V.Zhirov and D.L.Shepelyansky, Eur. Phys. J. B **77**, 523 (2010); `www.quantware.ups-tlse.fr/QWLIB/2drankwikipedia/`
9. L.Ermann, A.D.Chepelianskii and D.L.Shepelyansky, J. Phys. A: Math. Theor. **45**, 275101 (2012); `www.quantware.ups-tlse.fr/QWLIB/dvvadi/`
10. K.M.Frahm and D.L.Shepelyansky, Eur. Phys. J. B **85**, 355 (2012); `www.quantware.ups-tlse.fr/QWLIB/twittermatrix/`
11. L.Ermann, K.M. Frahm and D.L.Shepelyansky, *Spectral properties of Google matrix of Wikipedia and other networks*, arXiv:1212.1068 [cs.IR] (2012) (Eur. Phys. J. B in press).
12. G.W. Stewart, *Matrix Algorithms Eigensystems*, (SIAM, 2001), Vol. II
13. G.H. Golub and C. Greif, *BIT Num. Math.* **46**, 759 (2006)
14. K.M. Frahm and D.L. Shepelyansky, *Eur. Phys. J. B* **76**, 57 (2010)
15. K.M.Frahm, B.Georgeot and D.L.Shepelyansky, J. Phys, A: Math. Theor. **44**, 465101 (2011)
16. D. Fogaras, *Lect. Notes Comp. Sci.* **2877**, 65 (2003)
17. V. Hrisitidis, H. Hwang and Y. Papakonstantino, *ACM Trans. Database Syst.* **33**, 1 (2008)
18. A. D. Chepelianskii, *Towards physical laws for software architecture*, arXiv:1003.5455[cs.SE] (2010); `www.quantware.ups-tlse.fr/QWLIB/linuxnetwork/`
19. D. Donato, L. Laura, S. Leonardi and S. Millozzi, Eur. Phys. J. B **38**, 239 (2004)
20. G. Pandurangan, P. Raghavan and E. Upfal, Internet Math. **3**, 1 (2005)
21. N. Litvak, W.R.W. Scheinhardt, and Y. Volkovich, Lecture Notes in Computer Science, **4936**, 72 (2008).
22. M.H. Hart, *The 100: ranking of the most influential persons in history*, Citadel Press, N.Y. (1992).
23. `www.shanghairanking.com/`
24. K. Zyczkowski, M. Kus, W. Slomczynski and H.-J. Sommers, J. Phys. A: Math. Gen. **36**, 3425 (2003)
25. S. Fortunato, Phys. Rep. *486*, 75 (2010)
26. `www.quantware.ups-tlse.fr/QWLIB/wikirankevolution/`

OPEN ACCESS Freely available online

PLOS ONE

Highlighting Entanglement of Cultures via Ranking of Multilingual Wikipedia Articles

Young-Ho Eom, Dima L. Shepelyansky*

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

Abstract

How different cultures evaluate a person? Is an important person in one culture is also important in the other culture? We address these questions via ranking of multilingual Wikipedia articles. With three ranking algorithms based on network structure of Wikipedia, we assign ranking to all articles in 9 multilingual editions of Wikipedia and investigate general ranking structure of PageRank, CheiRank and 2DRank. In particular, we focus on articles related to persons, identify top 30 persons for each rank among different editions and analyze distinctions of their distributions over activity fields such as politics, art, science, religion, sport for each edition. We find that local heroes are dominant but also global heroes exist and create an effective network representing entanglement of cultures. The Google matrix analysis of network of cultures shows signs of the Zipf law distribution. This approach allows to examine diversity and shared characteristics of knowledge organization between cultures. The developed computational, data driven approach highlights cultural interconnections in a new perspective. Dated: June 26, 2013

Citation: Eom Y-H, Shepelyansky DL (2013) Highlighting Entanglement of Cultures via Ranking of Multilingual Wikipedia Articles. PLoS ONE 8(10): e74554. doi:10.1371/journal.pone.0074554

Editor: Matjaz Perc, University of Maribor, Slovenia

Received June 26, 2013; **Accepted** August 5, 2013; **Published** October 3, 2013

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 funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dima@irsamc.ups-tlse.fr

Introduction

Wikipedia, the online collaborative encyclopedia, is an amazing example of human collaboration for knowledge description, characterization and creation. Like the Library of Babel, described by Jorge Luis Borges [1], Wikipedia goes to accumulate the whole human knowledge. Since every behavioral 'footprint' (log) is recorded and open to anyone, Wikipedia provides great opportunity to study various types of social aspects such as opinion consensus [2,3], language complexity [4], and collaboration structure [5–7]. A remarkable feature of Wikipedia is its existence in various language editions. In a first approximation we can attribute each language to an independent culture, leaving for future refinements of cultures inside one language. Although Wikipedia has a neutral point of view policy, cultural bias or reflected cultural diversity is inevitable since knowledge and knowledge description are also affected by culture like other human behaviors [8–11]. Thus the cultural bias of contents [12] becomes an important issue. Similarity features between various Wikipedia editions has been discussed at [13]. However, the cross-cultural difference between Wikipedia editions can be also a valuable opportunity for a cross-cultural empirical study with quantitative approach. Recent steps in this direction, done for biographical networks of Wikipedia, have been reported in [14].

Here we address the question of how importance (ranking) of an article in Wikipedia depends on cultural diversity. In particular, we consider articles about persons. For instance, is an important person in English Wikipedia is also important in Korean Wikipedia? How about French? Since Wikipedia is the product of collective intelligence, the ranking of articles about persons is a collective evaluation of the persons by Wikipedia users. For the ranking of Wikipedia articles we use PageRank algorithm of Brin and Page [15], CheiRank and 2Drank algorithms used in [16–18], which allow to characterize the information flows with incoming and outgoing links. We also analyze the distribution of top ranked persons over main human activities attributed to politics, science, art, religion, sport, etc (all others), extending the approach developed in [17,19] to multiple cultures (languages). The comparison of different cultures shows that they have distinct dominance of these activities.

We attribute belongings of top ranked persons at each Wikipedia language to different cultures (native languages) and in this way construct the network of cultures. The Google matrix analysis of this network allows us to find interconnections and entanglement of cultures. We believe that our computational and statistical analysis of large-scale Wikipedia networks, combined with comparative distinctions of different languages, generates novel insights on cultural diversity.

Methods

We consider Wikipedia 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 adjacency matrix A_{ij}. If there is a link (one or more quotations) 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.

Google matrix

The matrix S_{ij} of Markov chain transitions is constructed from adjacency matrix A_{ij} by normalizing sum of elements of each column to unity ($S_{ij}=A_{ij}/\sum_i A_{ij}$, $\sum_i S_{ij}=1$) and replacing columns with only zero elements (*dangling nodes*) by $1/N$, with N being the matrix size. Then the Google matrix of this directed network has the form [15,20]:

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

In the WWW context the damping parameter α describes the probability $(1-\alpha)$ to jump to any article (node) for a random walker. The matrix G belongs to the class of Perron-Frobenius operators, it naturally appears in dynamical systems [21]. 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 walker at site i. It is possible to rank all nodes in a decreasing order of PageRank probability $P(K(i))$ so that the PageRank index $K(i)$ sorts all N nodes i according their ranks. For large size networks the PageRank vector and several other eigenvectors can be numerically obtained using the powerful Arnoldi algorithm as described in [22]. The PageRank vector can be also obtained by a simple iteration method [20]. Here, we use here the standard value of $\alpha=0.85$ [20].

To rank articles of Wikipedia, we use three ranking algorithms based on network structure of Wikipedia articles. Detail description of these algorithms and their use for English Wikipedia articles are given in [17–19,22].

PageRank algorithm

PageRank algorithm is originally introduced for Google web search engine to rank web pages of the World Wide Web (WWW) [15]. Currently PageRank is widely used to rank nodes of network systems including scientific papers [23], social network services [24] and even biological systems [25]. 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 of N nodes is given by

$$\begin{aligned} P(i,t) &= \sum_j G_{ij}\ P(j,t-1)\ ,\ P(i,t) \\ &= (1-\alpha)/N+\alpha \sum_j A_{ij}\ P(j,t-1)/k_{out}(j). \end{aligned} \quad (2)$$

The stationary state $P(i)$ of $P(i,t)$ is the PageRank of node i. More detail information about PageRank algorithm is described in [20]. Ordering all nodes by their decreasing probability $P(i)$ we obtain the PageRank index $K(i)$.

The essential idea of PageRank algorithm is to use a directed link as a weighted 'recommendation'. Like in academic citation network, more cited nodes are considered to be more important. In addition, recommendations by highly ranked articles are more important. Therefore high PageRank nodes in the network have many incoming links from other nodes or incoming links from high PageRank nodes.

CheiRank algorithm

While the PageRank algorithm uses information of incoming links to node i, CheiRank algorithm considers information of outgoing links from node i [16–18]. Thus CheiRank is complementary to PageRank in order to rank nodes in directed networks. The CheiRank vector $P^*(i,t)$ of a node at iteration time t is given by

$$P^*(i)=(1-\alpha)/N+\alpha \sum_j A_{ji}\ P^*(j)/k_{in}(j) \quad (3)$$

We also point out that the CheiRank is the right eigenvector with maximal eigenvalue $\lambda=1$ satisfying the equation $P^*(i)=\sum_j G^*_{ij}\ P^*(j)$, where the Google matrix G^* is built for the network with inverted directions of links via the standard definition of G given above.

Like for PageRank, we consider the stationary state $P^*(i)$ of $P^*(i,t)$ as the CheiRank probability of node i at $\alpha=0.85$. High CheiRank nodes in the network have a large out-degree. Ordering all nodes by their decreasing probability $P^*(i)$ we obtain the CheiRank index $K^*(i)$.

We note that PageRank and CheiRank naturally appear in the world trade network corresponding to import and export in a commercial exchange between countries [26].

The correlation between PageRank and CheiRank vectors can be characterized by the correlator κ [16–18] defined by

$$\kappa=N \sum_i P(i)\ P^*(i)-1 \quad (4)$$

The value of correlator for each Wikipedia edition is represented in Table 1. All correlators are positive and distributed in the interval (1,8).

Table 1. Considered Wikipedia networks from language editions: English (EN), French (FR), German (DE), Italian (IT), Spanish (ES), Dutch (NL), Russian (RU), Hungarian (HU), Korean (KO).

Edition	N_A	N_L	κ	Date
EN	3920628	92878869	3.905562	Mar. 2012
FR	1224791	30717338	3.411864	Feb. 2012
DE	1396293	32932343	3.342059	Mar. 2012
IT	917626	22715046	7.953106	Mar. 2012
ES	873149	20410260	3.443931	Feb. 2012
NL	1034912	14642629	7.801457	Feb. 2012
RU	830898	17737815	2.881896	Feb. 2012
HU	217520	5067189	2.638393	Feb. 2012
KO	323461	4209691	1.084982	Feb. 2012

Here N_A is number of articles, N_L is number of hyperlinks between articles, κ is the correlator between PageRank and CheiRank. Date represents the time in which data are collected.

doi:10.1371/journal.pone.0074554.t001

2DRank algorithm

With PageRank $P(i)$ and CheiRank $P^*(i)$ probabilities, we can assign PageRank ranking $K(i)$ and CheiRank ranking $K^*(i)$ to each article, respectively. From these two ranks, we can construct 2-dimensional plane of K and K^*. The two dimensional ranking K_2 is defined by counting nodes in order of their appearance on ribs of squares in (K,K^*) plane with the square size growing from $K=1$ to $K=N$ [17]. A direct detailed illustration and description of this algorithm is given in [17]. Briefly, nodes with high PageRank and CheiRank both get high 2DRank ranking.

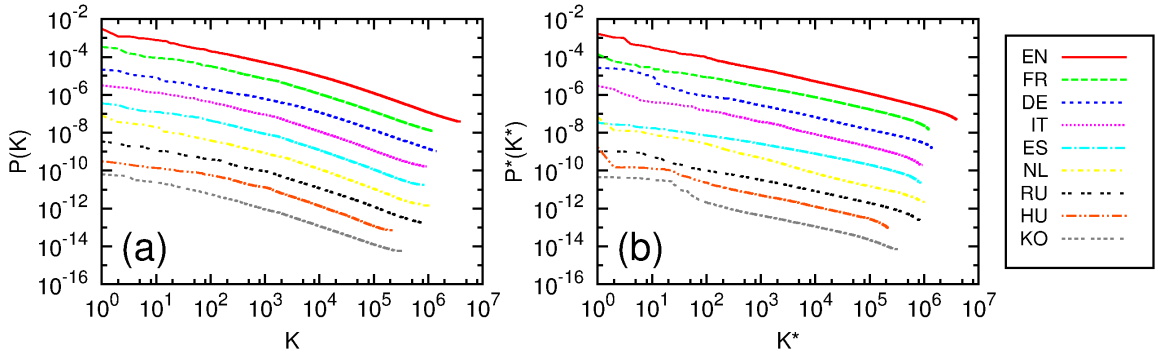


Figure 1. PageRank probability $P(K)$ as function of PageRank index K (a) and CheiRank probability $P^*(K^*)$ as function of CheiRank index K^* (b). For a better visualization each PageRank P and CheiRank P^* curve is shifted down by a factor 10^0 (EN), 10^1 (FR), 10^2 (DE), 10^3 (IT), 10^4 (ES), 10^5 (NL), 10^6 (RU), 10^7 (HU), 10^8 (KO).
doi:10.1371/journal.pone.0074554.g001

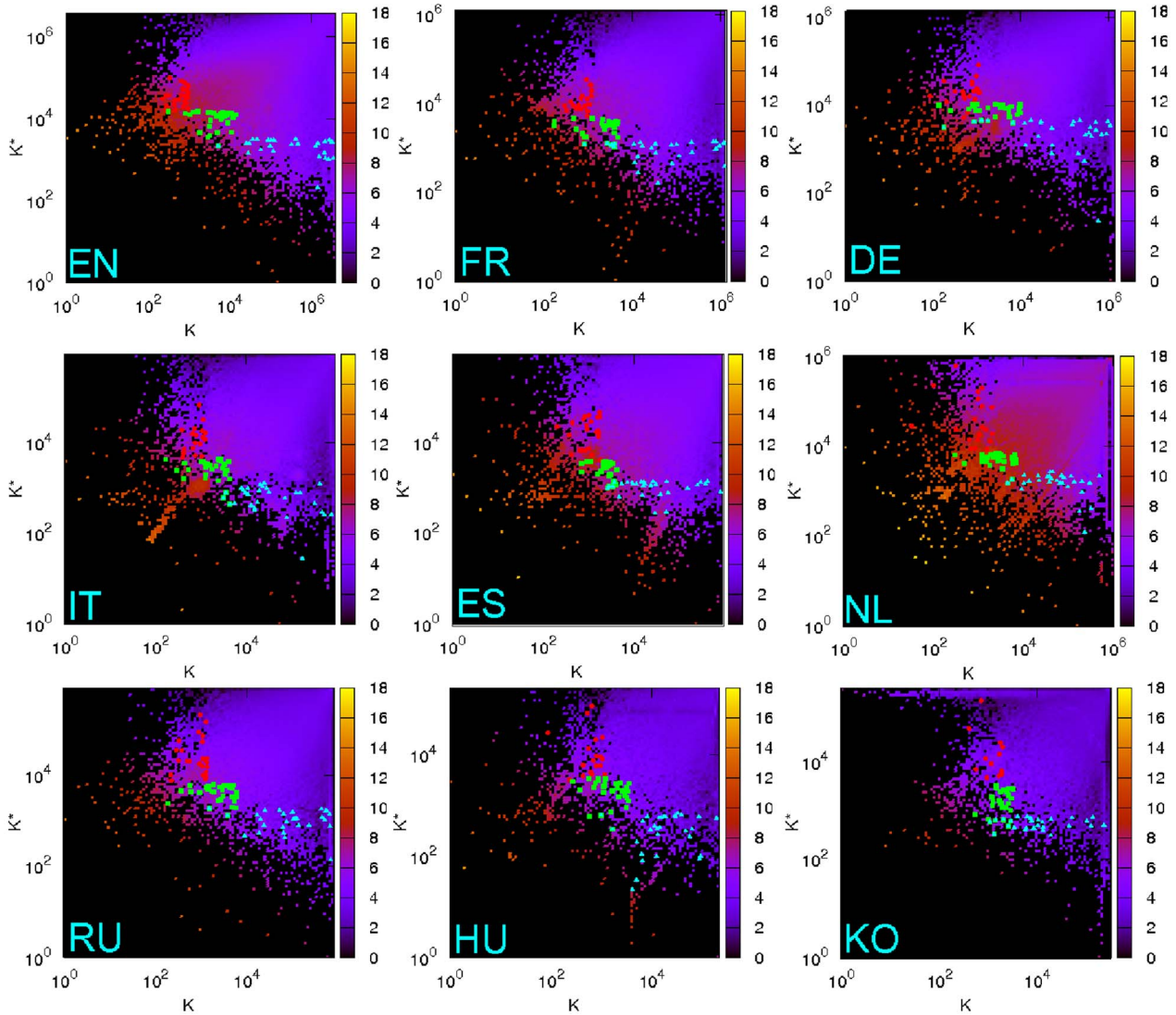


Figure 2. Density of Wikipedia articles in the PageRank ranking K versus CheiRank ranking K^* plane for each Wikipedia edition. The red points are top PageRank articles of persons, the green points are top 2DRank articles of persons and the cyan points are top CheiRank articles of persons. Panels show: English (top-left), French (top-center), German (top-right), Italian (middle-left), Spanish (middle-center), Dutch (middle-left), Russian (bottom-left), Hungarian (bottom-center), Korean (bottom-right). Color bars shown natural logarithm of density, changing from minimal nonzero density (dark) to maximal one (white), zero density is shown by black.
doi:10.1371/journal.pone.0074554.g002

Table 2. Example of list of top 10 persons by PageRank for English Wikipedia with their field of activity and native language.

$R_{EN,PageRank}$	Person	Field	Culture	Locality
1	Napoleon	Politics	FR	Non-local
2	Carl Linnaeus	Science	WR	Non-local
3	George W. Bush	Politics	EN	Local
4	Barack Obama	Politics	EN	Local
5	Elizabeth II	Politics	EN	Local
6	Jesus	Religion	WR	Non-local
7	William Shakespeare	Art	EN	Local
8	Aristotle	Science	WR	Non-local
9	Adolf Hitler	Politics	DE	Non-local
10	Bill Clinton	Politics	EN	Local

doi:10.1371/journal.pone.0074554.t002

Data Description

We consider 9 editions of Wikipedia including English (EN), French (FR), German (DE), Italian (IT), Spanish (ES), Dutch (NL), Russian (RU), Hungarian (HU) and Korean (KO). Since Wikipedia has various language editions and language is a most fundamental part of culture, the cross-edition study of Wikipedia can give us insight on cultural diversity. The overview summary of parameters of each Wikipedia is represented in Table 1.

The corresponding networks of these 9 editions are collected and kindly provided to us by S.Vigna from LAW, Univ. of Milano. The first 7 editions in the above list represent mostly spoken European languages (except Polish). Hungarian and Korean are additional editions representing languages of not very large population on European and Asian scales respectively. They allow us to see interactions not only between large cultures but also to see links on a small scale. The KO and RU editions allow us to compare views from European and Asian continents. We also note that in part these 9 editions reflect the languages present in the EC NADINE collaboration.

We understand that the present selection of Wikipedia editions does represent a complete view of all 250 languages present at Wikipedia. However, we think that this selection allows us to perform the quantitative statistical analysis of interactions between cultures making a first step in this direction.

To analyze these interactions we select the fist top 30 persons (or articles about persons) appearing in the top ranking list of each of 9 editions for 3 ranking algorithms of PageRank, CheiRank and 2DRank. We select these 30 persons manually analyzing each list. We attribute each of 30 persons to one of 6 fields of human activity: politics, science, art, religion, sport, and etc (here "etc" includes all other activities). In addition we attribute each person to one of 9 selected languages or cultures. We place persons belonging to other languages inside the additional culture WR (world) (e.g. Plato). Usually a belonging of a person to activity field

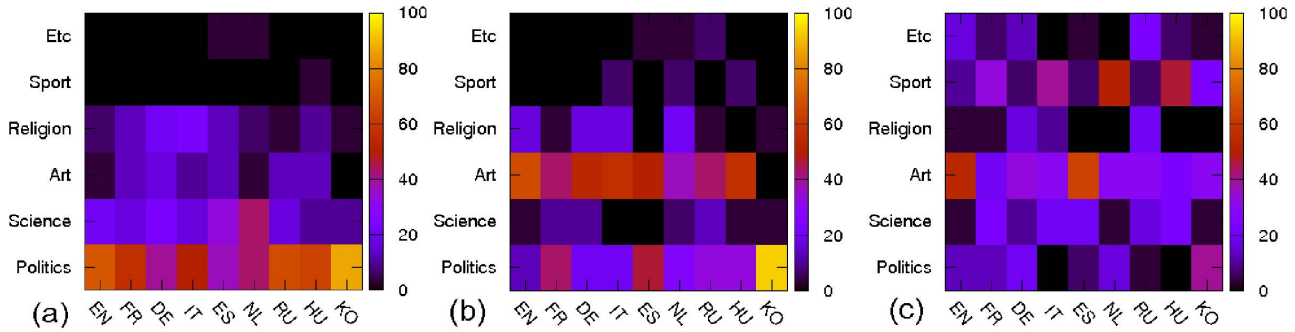


Figure 3. Distribution of top 30 persons in each rank over activity fields for each Wikipedia edition. Panels correspond to (a) PageRank, (b) 2DRank, (3) CheiRank. The color bar shows the values in percents.
doi:10.1371/journal.pone.0074554.g003

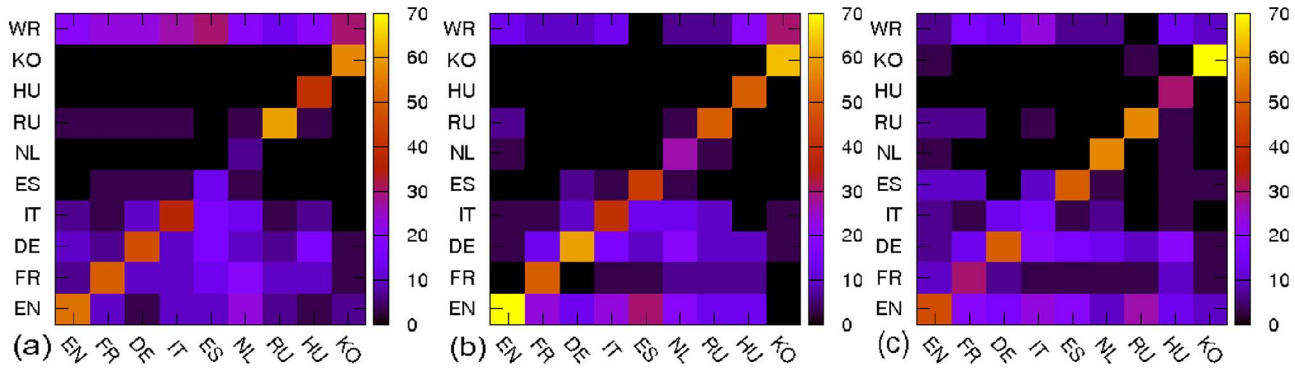


Figure 4. Distributions of top 30 persons over different cultures corresponding to Wikipedia editions, "WR" category represents all other cultures which do not belong to considered 9 Wikipedia editions. Panels show ranking by (a) PageRank, (b) 2DRank, (3) CheiRank. The color bar shows the values in percents.
doi:10.1371/journal.pone.0074554.g004

Table 3. PageRank contribution per link and in-degree of PageRank local and non-local heroes i for each edition.

Edition	N_{Local}	$[P(j)/k(j)_{out}]_L$		$[P(j)/k(j)_{out}]_{NL}$	$[k(L)_{in}]$		$[k(NL)_{in}]$
EN	16	1.43×10^{-8}	<	2.18×10^{-8}	5.3×10^{3}	>	3.1×10^{3}
FR	15	3.88×10^{-8}	<	5.69×10^{-8}	2.6×10^{3}	>	2.0×10^{3}
DE	14	3.48×10^{-8}	<	4.29×10^{-8}	2.6×10^{3}	>	2.1×10^{3}
IT	11	7.00×10^{-8}	<	7.21×10^{-8}	1.9×10^{3}	>	1.5×10^{3}
ES	4	5.44×10^{-8}	<	8.58×10^{-8}	2.2×10^{3}	>	1.2×10^{3}
NL	2	7.77×10^{-8}	<	14.4×10^{-8}	1.0×10^{3}	>	6.7×10^{2}
RU	18	6.67×10^{-8}	<	10.2×10^{-8}	1.7×10^{3}	>	1.5×10^{3}
HU	12	21.1×10^{-8}	<	32.3×10^{-8}	8.1×10^{2}	>	5.3×10^{2}
KO	17	16.6×10^{-8}	<	35.5×10^{-8}	4.7×10^{2}	>	2.3×10^{2}

$[P(j)/k(j)_{out}]_L$ and $[P(j)/k(j)_{out}]_{NL}$ are median PageRank contribution of a local hero L and non-local hero NL by a article j which cites local heroes L and non-local heroes NL respectively. $[k(L)_{in}]$ and $[k(NL)_{in}]$ are median number of in-degree $k(L)_{in}$ and $k(NL)_{in}$ of local hero L and non-local hero NL, respectively. N_{Local} is number local heroes in given edition.
doi:10.1371/journal.pone.0074554.t003

and language is taken from the English Wikipedia article about this person. If there is no such English Wikipedia article then we use an article of a Wikipedia edition language which is native for such a person. Usually there is no ambiguity in the distribution over activities and languages. Thus Christopher Columbus is attributed to IT culture and activity field etc, since English Wikipedia describes him as "italian explorer, navigator, and colonizer". By our definition politics includes politicians (e.g. Barak Obama), emperors (e.g. Julius Caesar), kings (e.g. Charlemagne). Arts includes writers (e.g. William Shakespeare), singers (e.g. Frank Sinatra), painters (Leonardo da Vinci), architects, artists, film makers (e.g. Steven Spielberg). Science includes physicists, philosophers (e.g. Plato), biologists, mathematicians and others. Religion includes such persons as Jesus, Pope John Paul II. Sport includes sportsmen (e.g. Roger Federer). All other activities are placed in activity etc (e.g. Christopher Columbus, Yuri Gagarin). Each person belongs only to one language and one activity field. There are only a few cases which can be questioned, e.g. Charles V, Holy Roman Emperor who is attributed to ES language since from early long times he was the king of Spain. All listings of person distributions over the above categories are presented at the web page given at Supporting Information (SI) file and in 27 tables given in File S1.

Unfortunately, we were obliged to construct these distributions manually following each person individually at the Wikipedia ranking listings. Due to that we restricted our analysis only to top 30 persons. We think that this number is sufficiently large so that the statistical fluctuations do not generate significant changes. Indeed, we find that our EN distribution over field activities is close to the one obtained for 100 top persons of English Wikipedia dated by Aug 2009 [17].

To perform additional tests we use the database of about 250000 person names in English, Italian and Dutch from the research work [14] provided to us by P.Aragón and A.Kaltenbrunner. Using this database we were able to use computerized (automatic) selection of top 100 persons from the ranking lists and to compare their distributions over activities and languages with our case of 30 persons. The comparison is presented in figures S1,S2,S3 in File S1. For these 3 cultures we find that our top 30 persons data are statistically stable even if the fluctuations are larger for CheiRank lists. This is in an agreement with the fact that the CheiRank probabilities. related to the outgoing links, are more fluctuating (see discussion at [19]).

Of course, it would be interesting to extend the computerized analysis of personalities to a larger number of top persons and larger number of languages. However, the database of persons in various languages still should be cleaned and checked and also attribution of persons to various activities and languages still requires a significant amount of work. Due to that we present here our analysis only for 30 top persons. But we note that by itself it represents an interesting case study since here we have the most important persons for each ranking. May be the top 1000 persons would be statistically more stable but clearly a person at position 30 is more important than a one at position 1000. Thus we think that the top 30 persons already give an interesting information on links and interactions between cultures. This information can be used in future more extended studies of a larger number of persons and languages.

Finally we note that the language is the primary element of culture even if, of course, culture is not reduced only to language. In this analysis we use in a first approximation an equivalence between language and culture leaving for future studies the refinement of this link which is of course much more complex. In this approximation we consider that a person like Mahatma Gandhi belongs to EN culture since English is the official language of India. A more advanced study should take into account Hindi

Table 4. List of local heroes by PageRank for each Wikipedia edition.

Edition	1st	2nd	3rd
EN	George W. Bush	Barack Obama	Elizabeth II
FR	Napoleon	Louis XIV of France	Charles de Gaulle
DE	Adolf Hitler	Martin Luther	Immanuel Kant
IT	Augustus	Dante Alighieri	Julius Caesar
ES	Charles V, Holy Roman Emperor	Philip II of Spain	Francisco Franco
NL	William I of the Netherlands	Beatrix of the Netherlands	William the Silent
RU	Peter the Great	Joseph Stalin	Alexander Pushkin
HU	Matthias Corvinus	Szentágothai János	Stephen I of Hungary
KO	Gojong of the Korean Empire	Sejong the Great	Park Chung-hee

All names are represented by article titles in English Wikipedia. Here "William the Silent" is the third local hero in Dutch Wikipedia but he is out of top 30 persons.
doi:10.1371/journal.pone.0074554.t004

Table 5. List of local heroes by CheiRank for each Wikipedia edition.

Edition	1st	2nd	3rd
EN	C. H. Vijayashankar	Matt Kelley	William Shakespeare (inventor)
FR	Jacques Davy Duperron	Jean Baptiste Eblé	Marie-Magdeleine Aymé de La Chevrelière
DE	Harry Pepl	Marc Zwiebler	Eugen Richter
IT	Nduccio	Vincenzo Olivieri	Mina (singer)
ES	Che Guevara	Arturo Mercado	Francisco Goya
NL	Hans Renders	Julian Jenner	Marten Toonder
RU	Aleksander Vladimirovich Sotnik	Aleksei Aleksandrovich Bobrinsky	Boris Grebenshchikov
HU	Csernus Imre	Kati Kovács	Pléh Csaba
KO	Lee Jong-wook (baseball)	Kim Dae-jung	Kim Kyu-sik

All names are represented by article titles in English Wikipedia.
doi:10.1371/journal.pone.0074554.t005

Table 6. List of local heroes by 2DRank for each Wikipedia edition.

Edition	1st	2nd	3rd
EN	Frank Sinatra	Paul McCartney	Michael Jackson
FR	François Mitterrand	Jacques Chirac	Honoré de Balzac
DE	Adolf Hitler	Otto von Bismarck	Ludwig van Beethoven
IT	Giusppe Garibaldi	Raphael	Benito Mussolini
ES	Simón Bolívar	Francisco Goya	Fidel Castro
NL	Albert II of Belgium	Johan Cruyff	Rembrandt
RU	Dmitri Mendeleev	Peter the Great	Yaroslav the Wise
HU	Stephen I of Hungary	Sándor Petöfi	Franz Liszt
KO	Gojong of the Korean Empire	Sejong the Great	Park Chung-hee

All names are represented by article titles in English Wikipedia.
doi:10.1371/journal.pone.0074554.t006

Wikipedia edition and attribute this person to this edition. Definitely our statistical study is only a first step in Wikipedia based statistical analysis of network of cultures and their interactions.

We note that any person from our top 30 ranking belongs only to one activity field and one culture. We also define local heros as those who in a given language edition are attributed to this language, and non-local heros as those who belong in a given edition to other languages. We use category WR (world) where we

Table 7. List of global heroes by PageRank and 2DRank for all 9 Wikipedia editions.

Rank	PageRank global heroes	Θ_{PR}	N_A	2DRank global heroes	Θ_{2D}	N_A
1st	Napoleon	259	9	Micheal Jackson	119	5
2nd	Jesus	239	9	Adolf Hitler	93	6
3rd	Carl Linnaeus	235	8	Julius Caesar	85	5
4th	Aristotle	228	9	Pope Benedict XVI	80	4
5th	Adolf Hitler	200	9	Wolfgang Amadeus Mozart	75	5
6th	Julius Caesar	161	8	Pope John Paul II	71	4
7th	Plato	119	6	Ludwig van Beethoven	69	4
8th	Charlemagne	111	8	Bob Dylan	66	4
9th	William Shakespeare	110	7	William Shakespeare	57	3
10th	Pope John Paul II	108	6	Alexander the Great	56	3

All names are represented by article titles in English Wikipedia. Here, Θ_A is the ranking score of the algorithm A (5); N_A is the number of appearances of a given person in the top 30 rank for all editions.
doi:10.1371/journal.pone.0074554.t007

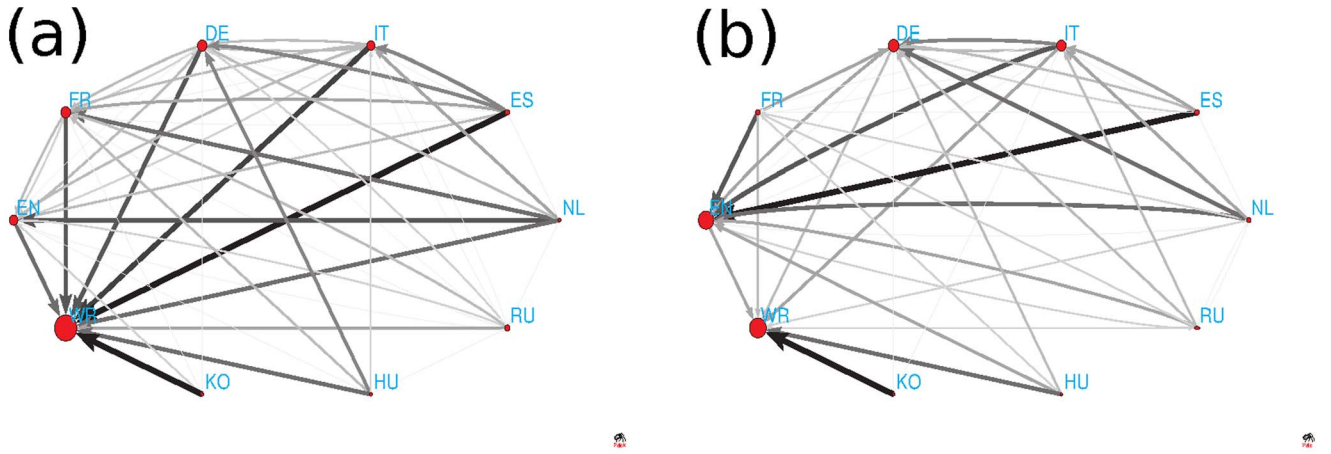


Figure 5. Network of cultures obtained from 9 Wikipedia languages and the remaining world (WR) selecting 30 top persons of PageRank (a) and 2DRank (b) in each culture. The link width and darkness are proportional to a number of foreign persons quoted in top 30 of a given culture, the link direction goes from a given culture to cultures of quoted foreign persons, quotations inside cultures are not considered. The size of nodes is proportional to their PageRank.
doi:10.1371/journal.pone.0074554.g005

place persons who do not belong to any of our 9 languages (e.g. Pope John Paul II belongs to WR since his native language is Polish).

Results

We investigate ranking structure of articles and identify global properties of PageRank and CheiRank vectors. The detailed analysis is done for top 30 persons obtained from the global list of ranked articles for each of 9 languages. The distinctions and common characteristics of cultures are analyzed by attributing top 30 persons in each language to human activities listed above and to their native language.

General ranking structure

We calculate PageRank and CheiRank probabilities and indexes for all networks of considered Wikipedia editions. The PageRank and CheiRank probabilities as functions of ranking indexes are shown in Fig. 1. The decay is compatible with an approximate algebraic decrease of a type $P \sim 1/K^{\beta}$, $P^* \sim 1/K^{*\beta}$ with $\beta \sim 1$ for PageRank and $\beta \sim 0.6$ for CheiRank. These values are similar to those found for the English Wikipedia of 2009 [17]. The difference of β values originates from asymmetric nature between in-degree and out-degree distributions, since PageRank is based on incoming edges while CheiRank is based on outgoing edges. In-degree distribution of Wikipedia editions is broader than out-degree distribution of the same edition. Indeed, the CheiRank probability is proportional to frequency of outgoing links which has a more rapid decay compared to incoming one (see discussion in [17]). The PageRank (CheiRank) probability distributions are similar for all editions. However, the fluctuations of P^* are stronger that is related to stronger fluctuations of outgoing edges [19].

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.

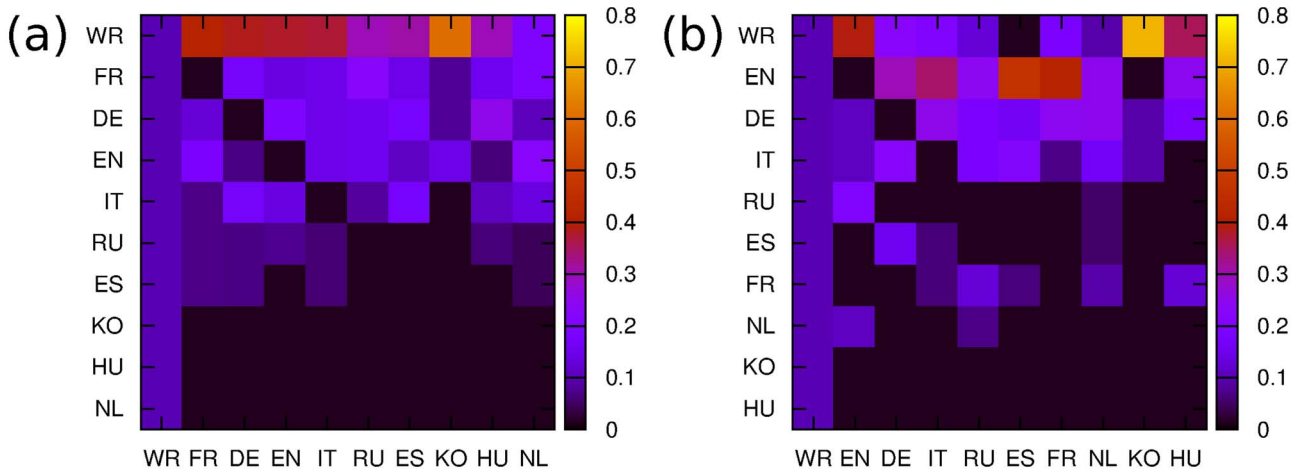


Figure 6. Google matrix of network of cultures from Fig. 5, shown respectively for panels (a),(b)**.** The matrix elements G_{ij} are shown by color at the damping factor $\alpha = 0.85$, index j is chosen as the PageRank index K of PageRank vector so that the top cultures with $K = K' = 1$ are located at the top left corner of the matrix.
doi:10.1371/journal.pone.0074554.g006

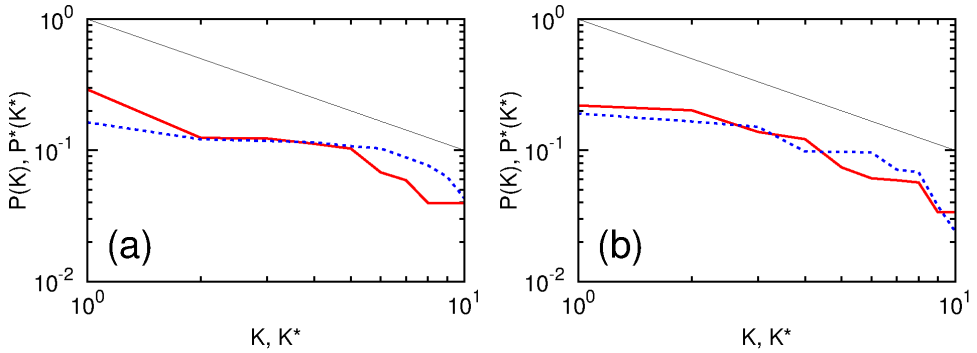


Figure 7. Dependence of probabilities of PageRank P (red) and CheiRank P^* (blue) on corresponding indexes K and K^*. The probabilities are obtained from the network and Google matrix of cultures shown in Fig. 5 and Fig. 6 for corresponding panels (a),(b). The straight lines indicate the Zipf law $P \sim 1/K; P^* \sim 1/K^*$.
doi:10.1371/journal.pone.0074554.g007

Since each article has its PageRank ranking K and CheiRank ranking K^*, we can assign two dimensional coordinates to all the articles. Fig. 2 shows the density of articles in the two dimensional plane (K,K^*) for each Wikipedia edition. The density is computed for 100×100 logarithmically equidistant cells which cover the whole plane (K,K^*). The density plot represents the locations of articles in the plane. We can observe high density of articles around line $K = K^* + const$ that indicates the positive correlation between PageRank and CheiRank. However, there are only a few articles within the region of top both PageRank and CheiRank indexes. We also observe the tendency that while high PageRank articles $(K < 100)$ have intermediate CheiRank $(10^2 < K^* < 10^4)$, high CheiRank articles $(K^* < 100)$ have broad PageRank rank values.

Ranking of articles for persons

We choose top 30 articles about persons for each edition and each ranking. In Fig. 2, they are shown by red circles (PageRank), green squares (2DRank) and cyan triangles (CheiRank). We assign local ranking $R_{E,A}$ $(1 \ldots 30)$ to each person in the list of top 30 persons for each edition E and ranking algorithm A. An example of $E = EN$ and $A = PageRank$ are given in Table 2.

From the lists of top persons, we identify the "fields" of activity for each top 30 rank person in which he/she is active on. We categorize six activity fields - politics, art, science, religion, sport and etc (here "etc" includes all other activities). As shown in Fig. 3, 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, art becomes dominant and politics is secondarily dominant. In case of CheiRank, art and sport are dominant fields. 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 giving him a high CheiRank.

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 politician 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.

Next we investigate distributions over "cultures" to which persons belong. We determined the culture of person based on the language the person mainly used (mainly native language). We consider 10 culture categories - EN, FR, DE, IT, ES, NL, RU, HU, KO and WR. Here "WR" category represents all other cultures which do not belong to considered 9 Wikipedia editions.

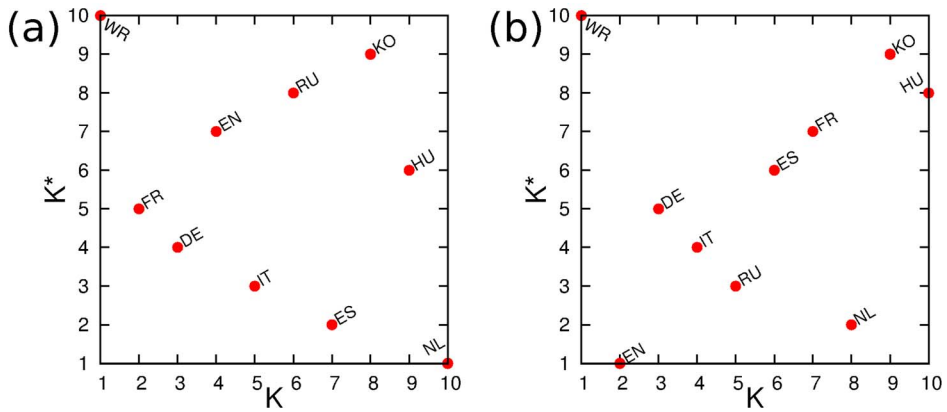


Figure 8. PageRank versus CheiRank plane of cultures with corresponding indexes K and K^* obtained from the network of cultures for corresponding panels (a),(b).
doi:10.1371/journal.pone.0074554.g008

Comparing with the culture of persons at various editions, we can assign "locality" to each 30 top rank persons for a given Wikipedia edition and ranking algorithm. For example, as shown in Table 2, *George W. Bush* belongs to "Politics", "English" and "Local" for English Wikipedia and PageRank, while *Jesus* belongs to "Religion", "World" WR and "Non-local".

As shown in Fig. 4, regardless of ranking algorithms, main part of top 30 ranking persons of each edition belong to the culture of the edition (usually about 50%). For example, high PageRank persons in English Wikipedia are mainly English (53.3%). This corresponds to the self-focusing effect discussed in [6]. It is notable that top ranking persons in Korean Wikipedia are not only mainly Korean (56.7%) but also the most top ranking non Korean persons in Korean Wikipedia are Chinese and Japanese (20%). Although there is a strong tendency that each edition favors its own persons, there is also overlap between editions. For PageRank, on average, 23.7 percent of top persons are overlapping while for CheiRank , the overlap is quite low, only 1.3 percent. For 2DRank, the overlap is 6.3 percent. The overlap of list of top persons implies the existence of cross-cultural 'heroes'.

To understand the difference between local and non-local top persons for each edition quantitatively, we consider the PageRank case because it has a large fraction of non-local top persons. From Eq. (2), a citing article j contributes $\langle P(j)/k_{out}(j)\rangle$ to PageRank of a node i. So the PageRank $P(i)$ can be high if the node i has many incoming links from citing articles j or it has incoming links from high PageRank nodes j with low out-degree $k_{out}(j)$. Thus we can identify origin of each top person's PageRank using the average PageRank contribution $\langle P(j)/k_{out}(j)\rangle$ by nodes j to person i and average number of incoming edges (in-degree) $k_{in}(i)$ of person i .

As represented in Table 3, considering median, local top persons have more incoming links than non-local top persons but the PageRank contribution of the corresponding links are lower than links of non-local top persons. This indicates that local top persons are cited more than non-local top persons but non-local top persons are cited more high weighted links (i.e. cited by important articles or by articles which don't have many citing links).

Global and local heroes

Based on cultural dependency on rankings of persons, we can identify global and local heroes in the considered Wikipedia editions. However, for CheiRank the overlap is very low and our statistics is not sufficient for selection of global heroes. Hence we consider only PageRank and 2DRank cases. We determine the local heroes for each ranking and for each edition as top persons of the given ranking who belongs to the same culture as the edition. Top 3 local heroes for each ranking and each edition are represented in Table 4 (PageRank), Table 5 (CheiRank) and Table 6 (2DRank), respectively.

In order to identify the global heroes, we define ranking score $\Theta_{P,A}$ for each person P and each ranking algorithm A. Since every person in the top person list has relative ranking $R_{P,E,A}$ for each Wikipedia edition E and ranking algorithm A (For instance, in Table 2, $R_{Napoleon,EN,PageRank}=1$). The ranking score $\Theta_{P,A}$ of a person P is give by

$$\Theta_{P,A}=\sum_{E}\,(31-R_{P,E,A}) \qquad (5)$$

According to this definition, a person who appears more often in the lists of editions and has top ranking in the list gets high ranking score. We sort this ranking score for each algorithm. In this way obtain a list of global heroes for each algorithm. The result is shown in Table 7. Napoleon is the 1st global hero by PageRank and Micheal Jackson is the 1st global hero by 2DRank.

Network of cultures

To characterize the entanglement and interlinking of cultures we use the data of Fig. 4 and from them construct the network of cultures. The image of networks obtained from top 30 persons of PageRank and 2DRank listings are shown in Fig. 5 (we do not consider CheiRank case due to small overlap of persons resulting in a small data statistics). The weight of directed Markov transition, or number of links, from a culture A to a culture B is given by a number of persons of a given culture B (e.g FR) appearing in the list of top 30 persons of PageRank (or 2DRank) in a given culture A (e.g. EN). Thus e.g. for transition from EN to FR in PageRank we find 2 links (2 French persons in PageRank top 30 persons of English Wikipedia); for transition from FR to EN in PageRank we have 3 links (3 English persons in PageRank top 30 persons of French Wikipedia). The transitions inside each culture (persons of the same language as language edition) are omitted since we are analyzing the interlinks between cultures. Then the Google matrix of cultures is constructed by the standard rule for the directed networks: all links are treated democratically with the same weight, sum of links in each column is renormalized to unity, $\alpha=0.85$. Even if this network has only 10 nodes we still can find for it PageRank and CheiRank probabilities P and P^* and corresponding indexes K and K^*. The matrix elements of G matrix, written in order of index K, are shown in Fig. 6 for the corresponding networks of cultures presented in Fig. 5. We note that we consider all cultures on equal democratic grounds.

The decays of PageRank and CheiRank probabilities with the indexes K,K^* are shown in Fig. 7 for the culture networks of Fig. 5. On a first glance a power decay like the Zipf law [27] $P\sim 1/K$ looks to be satisfactory. The formal power law fit $P\sim 1/K^z, P^*\sim 1/(K^*)^{z^*}$, done in log–log-scale for $1\leq K,K^*\leqslant q10$, gives the exponents $z=0.85\pm 0.09, z^*=0.45\pm 0.09$ (Fig. 7a), $z=0.88\pm 0.10, z^*=0.77\pm 0.16$ (Fig. 7b). However, the error bars for these fits are relatively large. Also other statistical tests (e.g. the Kolmogorov-Smirnov test, see details in [28]) give low statistical accuracy (e.g. statistical probability $p\approx 0.2; 0.1$ and $p\approx 0.01; 0.01$ for exponents $z,z^*=0.79,0.42$ and $0.75,0.65$ in Fig. 7a and Fig. 7b respectively). It is clear that 10 cultures is too small to have a good statistical accuracy. Thus, a larger number of cultures should be used to check the validity of the generalized Zipf law with a certain exponent. We make a conjecture that the Zipf law with the generalized exponents z,z^* will work in a better way for a larger number of multilingual Wikipedia editions which now have about 250 languages.

The distributions of cultures on the PageRank - CheiRank plane (K,K^*) are shown in Fig. 8. For the network of cultures constructed from top 30 PageRank persons we obtain the following ranking. The node WR is located at the top PageRank $K=1$ and it stays at the last CheiRank position $K^*=10$. This happens due to the fact that such persons as *Carl Linnaeus, Jesus, Aristotle, Plato, Alexander the Great, Muhammad* are not native for our 9 Wikipedia editions so that we have many nodes pointing to WR node, while WR has no outgoing links. The next node in PageRank is FR node at $K=2,K^*=5$, then DE node at $K=3,K^*=4$ and only then we find EN node at $K=4,K^*=7$. The node EN is not at all at top PageRank positions since it has many American politicians that does not count for links between cultures. After the world WR the top position is taken by French (FR) and then German (DE) cultures which have strong links inside the continental Europe.

However, the ranking is drastically changed when we consider top 30 2DRank persons. Here, the dominant role is played by art and science with singers, artists and scientists. The world WR here remains at the same position at $K=1,K^*=10$ but then we obtain English EN $(K=2,K^*=1)$ and German DE $(K=3,K^*=5)$ cultures while FR is moved to $K=K^*=7$.

Discussion

We investigated cross-cultural diversity of Wikipedia via ranking of Wikipedia articles. Even if the used ranking algorithms are purely based on network structure of Wikipedia articles, we find cultural distinctions and entanglement of cultures obtained from the multilingual editions of Wikipedia.

In particular, we analyze raking of articles about persons and identify activity field of persons and cultures to which persons belong. Politics is dominant in top PageRank persons, art is dominant in top 2DRank persons and in top CheiRank persons art and sport are dominant. We find that each Wikipedia edition favors its own persons, who have same cultural background, but there are also cross-cultural non-local heroes, and even ''global heroes''. We establish that local heroes are cited more often but non-local heroes on average are cited by more important articles.

Attributing top persons of the ranking list to different cultures we construct the network of cultures and characterize entanglement of cultures on the basis of Google matrix analysis of this directed network.

We considered only 9 Wikipedia editions selecting top 30 persons in a ''manual'' style. It would be useful to analyze a larger number of editions using an automatic computerized selection of persons from prefabricated listing in many languages developing lines discussed in [14]. This will allow to analyze a large number of persons improving the statistical accuracy of links between different cultures.

The importance of understanding of cultural diversity in globalized world is growing. Our computational, data driven approach can provide a quantitative and efficient way to understand diversity of cultures by using data created by millions of Wikipedia users. We believe that our results shed a new light on how organized interactions and links between different cultures.

Supporting Information

File S1 Presents Figures S1, S2, S3 in SI file showing comparison between probability distributions over activity fields and language for top 30 and 100 persons for EN, IT, NK respectively; tables S1, S2, … S27 in SI file showing top 30 persons in PageRank, CheiRank and 2DRank for all 9 Wikipedia editions. All names are given in English. Supplementary methods, tables, ranking lists and figures are available at http://www.quantware.ups-tlse.fr/QWLIB/wikiculturenetwork/; data sets of 9 hyperlink networks are available at [29] by a direct request addressed to S.Vigna.
(PDF)

Acknowledgments

We thank Sebastiano Vigna [29] who kindly provided to us the network data of 9 Wikipedia editions, collected in the frame of FET NADINE project. We thank Pablo Aragón and Andreas Kaltenbrunner for the list of persons in EN, IT, NL which we used to obtain supporting Figs.S1,S2,S3 in File S1.

Author Contributions

Conceived and designed the experiments: DLS. Performed the experiments: YHE. Analyzed the data: YHE DLS. Contributed reagents/materials/analysis tools: YHE DLS. Wrote the paper: YHE DLS.

References

1. Borges JL (1962) *The Library of Babel* in *Ficciones*, Grove Press, New York
2. Kaltenbrunner A, Laniado D (2012) *There is no deadline - time evolution of Wikipedia discussions*, Proc. of the 8th Intl. Symposium on Wikis and Open Collaboration, Wik- iSym12, Linz
3. Torok J, Iniguez G, Yasseri T, San Miguel M, Kaski K, et al. (2013) *Opinion, conflicts and consensus: modeling social dynamics in a collaborative enviroment* Phys Rev Lett 110: 088701
4. Yasseri T, Kornai A, Kertész J (2012) *A practical approach to language complexity: a Wikipedia case study* PLoS ONE, 7: e48386
5. Brandes U, Kenis P, Lerner U, van Raaij D (2009) *Network analysis of collaboration structure in Wikipedia* Proc. 18th Intl. Conf. WWW, :731
6. 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 :11
7. 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
8. Norenzayan A (2011) *Explaining human behavioral diversity*, Science, 332: 1041
9. Gelfand MJ, Raver JL, Nishii L, Leslie LM, Lun J, et al. (2011) *Differences between tight and loose cultures: a 33-nation study*, Science, 332: 1100
10. 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]
11. 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
12. Callahan ES, Herriing SC (2011) *Cultural bias in Wikipedia content on famous persons*, Journal of the American society for information science and technology 62: 1899
13. 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 Eighth Annual International Symposium on Wikis and Open Collaboration (WikiSym 2012), ACM, New York No 20
14. Aragón P, Laniado D, Kaltenbrunner A, Volkovich Y (2012) *Biographical social networks on Wikipedia: a cross-cultural study of links that made history*, Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration (WikiSym 2012), ACM, New York No 19; arXiv:1204.3799v2[cs.SI]
15. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine Computer Networks and ISDN Systems 30: 107
16. Chepelianskii AD (2010) *Towards physical laws for software architecture* arXiv:1003.5455 [cs.SE]
17. Zhirov AO, Zhirov OV, Shepelyansky DL (2010) *Two-dimensional ranking of Wikipedia articles*, Eur Phys J B 77: 523
18. Ermann L, Chepelianskii AD, Shepelyansky DL (2012) *Toward two-dimensional search engines*, J Phys A: Math Theor 45: 275101
19. Eom YH, Frahm KM, Benczúr A, Shepelyansky DL (2013) *Time evolution of Wikipedia network ranking* arXiv:1304.6601 [physics.soc-ph]
20. Langville AM, Meyer CD (2006) *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton
21. Brin M, Stuck G (2002) *Introduction to dynamical systems*, Cambridge Univ. Press, Cambridge, UK
22. Ermann L, Frahm KM, Shepelyansky DL (2013) *Spectral properties of Google matrix of Wikipedia and other networks*, Eur Phys J D 86: 193
23. Chen P, Xie H, Maslov S, Redner S (2007) *Finding scientific gems with Googles PageRank algorithm* Jour Informetrics, 1: 8
24. 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 :591
25. Kandiah V, Shepelyansky DL (2013) *Google matrix analysis of DNA sequences*, PLoS ONE 8(5): e61519
26. Ermann L, Shepelyansky DL (2011) *Google matrix of the world trade network*, Acta Physica Polonica A 120(6A), A158
27. Zipf GK (1949) *Human behavior and the principle of least effort*, Addison-Wesley, Boston
28. Clauset A, Shalizi CR, Newman MEJ (2009) *Power-law distributions in empirical data*, SIAM Rev 51(4): 661
29. Personal website of Sebastiano Vigna. Available: http://vigna.dsi.unimi.it/. Accessed 2013 Jun 26.

SUPPORTING INFORMATION FOR:
Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles

Young-Ho Eom[1], Dima L. Shepelyansky[1,*]

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

∗ Webpage: www.quantware.ups-tlse.fr/dima

1 Additional data

Supplementary methods, tables, ranking lists and figures are available at
http://www.quantware.ups-tlse.fr/QWLIB/wikiculturenetwork/;
data sets of 9 hyperlink networks are available at
http://vigna.dsi.unimi.it/
by a direct request addressed to S.Vigna.

Here we present additional figures and tables for the main part of the paper.

Figures S1, S2, S3 show comparison between probability distributions over activity fields and language for top 30 and 100 persons for EN, IT, NK respectively.

Tables show top 30 persons in PageRank, CheiRank and 2DRank for all 9 Wikipedia editions. All names are given in English.

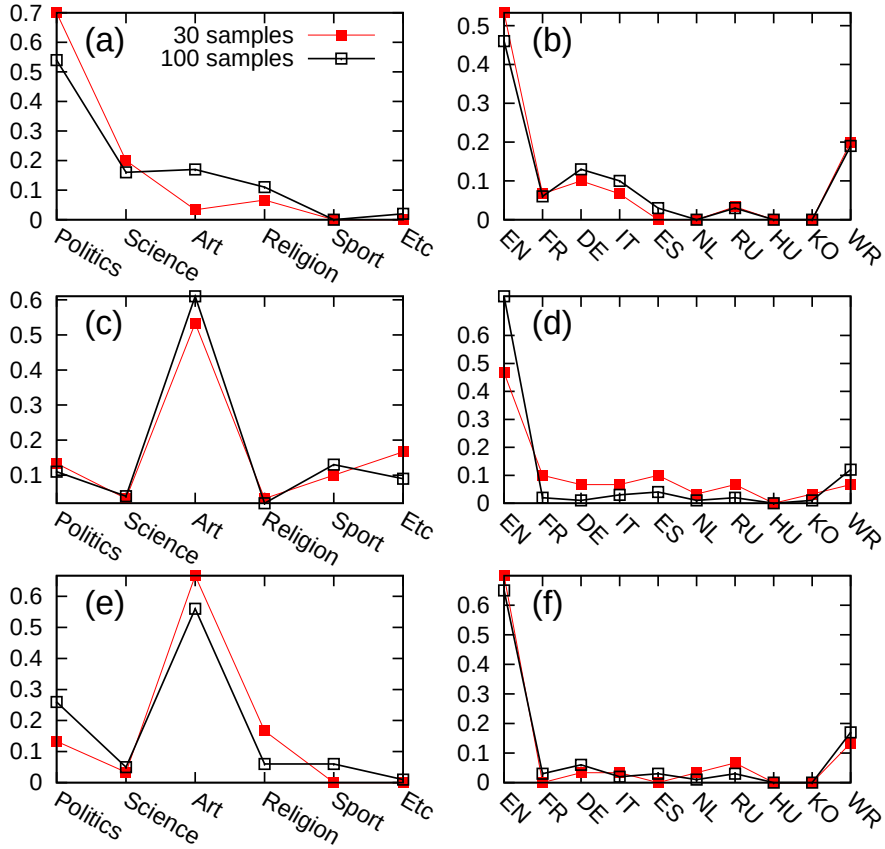


Figure S1: Probability distributions of activity fields and languages of top 30 persons and top 100 persons in English Wikipedia EN (total probability is normalized to unity): (a) Distribution of activity fields of PageRank top persons (b) Distribution of langauge of PageRank top persons. (c) Distribution of activity fields of CheiRank top persons (d) Distribution of langauge of CheiRank top persons. (e) Distribution of activity fields of 2DRank top persons (f) Distribution of langauge of 2DRank top persons.

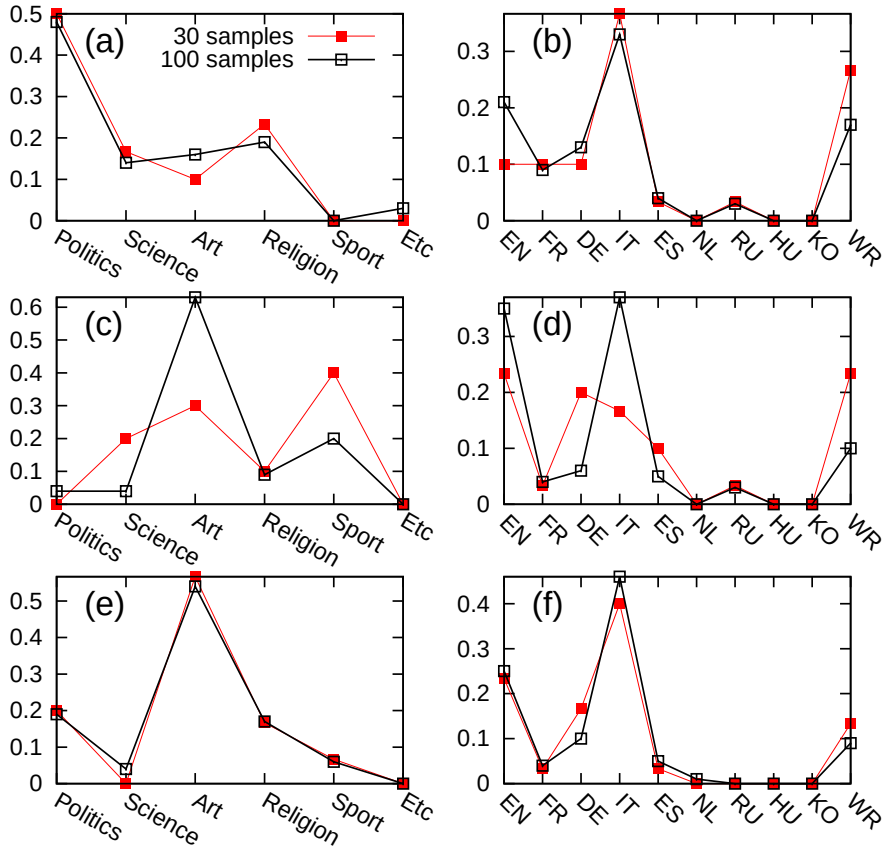


Figure S2: Same as in Fig.SI1 for Italian Wikipedia IT.

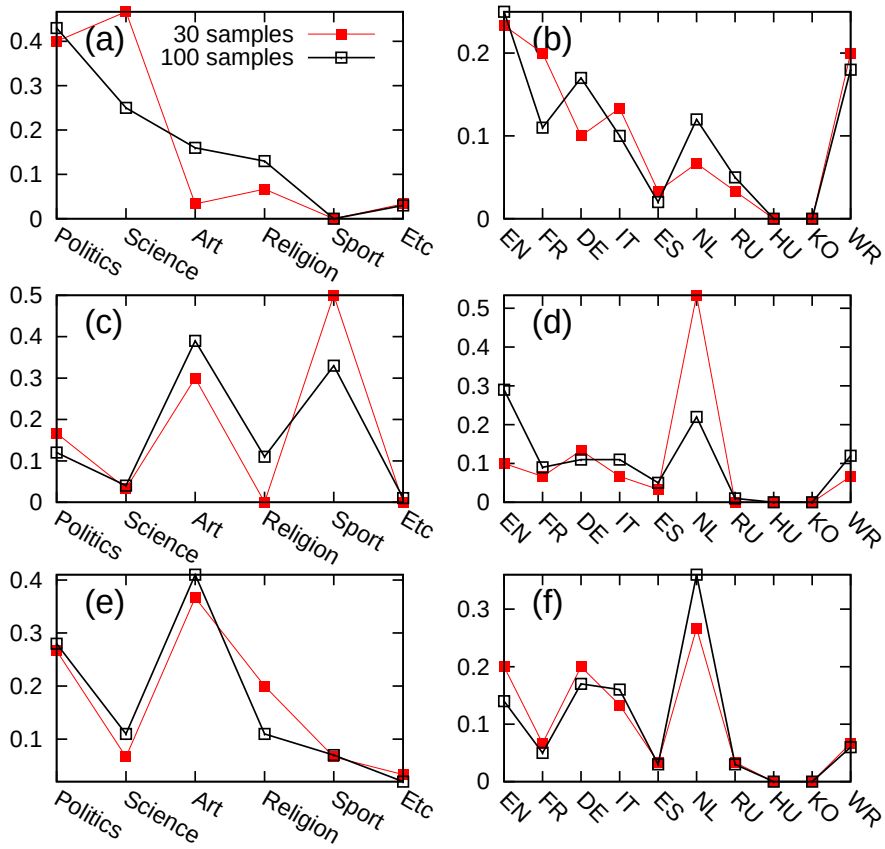


Figure S3: Same as in Fig.SI1 for Dutch Wikipedia NL.

Table S1: Top 30 persons by PageRank for English Wikipedia with their field of activity and native language.

$R_{EN,PageRank}$	Person	Field	Culture
1	Napoleon	Politics	FR
2	Carl Linnaeus	Science	WR
3	George W. Bush	Politics	EN
4	Barack Obama	Politics	EN
5	Elizabeth II	Politics	EN
6	Jesus	Religion	WR
7	William Shakespeare	Art	EN
8	Aristotle	Science	WR
9	Adolf Hitler	Politics	DE
10	Bill Clinton	Politics	EN
11	Franklin D. Roosevelt	Politics	EN
12	Ronald Reagan	Politics	EN
13	George Washington	Politics	EN
14	Plato	Science	WR
15	Richard Nixon	Politics	EN
16	Abraham Lincoln	Politics	EN
17	Joseph Stalin	Politics	RU
18	Winston Churchill	Politics	EN
19	John F. Kennedy	Politics	EN
20	Henry VIII of England	Politics	EN
21	Muhammad	Religion	WR
22	Thomas Jefferson	Politics	EN
23	Albert Einstein	Science	DE
24	Alexander the Great	Politics	WR
25	Augustus	Politics	IT
26	Charlemagne	Politics	FR
27	Karl Marx	Science	DE
28	Charles Darwin	Science	EN
29	Elizabeth I of England	Politics	EN
30	Julius Caesar	Politics	IT

Table S2: Top 30 persons by 2DRank for English Wikipedia with their field of activity and native language.

$R_{EN,2DRank}$	Person	Field	Culture
1	Frank Sinatra	Art	EN
2	Paul McCartney	Art	EN
3	Michael Jackson	Art	EN
4	Steven Spielberg	Art	EN
5	Pope Pius XII	Religion	IT
6	Vladimir Putin	Politics	RU
7	Mariah Carey	Art	EN
8	John Kerry	Politics	EN
9	Isaac Asimov	Art	EN
10	Stephen King	Art	EN
11	Dolly Parton	Art	EN
12	Prince (musician)	Art	EN
13	Robert Brown (botanist)	Science	EN
14	Vincent van Gogh	Art	NL
15	Lady Gaga	Art	EN
16	Beyoncé Knowles	Art	EN
17	Pope John Paul II	Religion	WR
18	Lord Byron	Art	EN
19	Muhammad	Religion	WR
20	Johnny Cash	Art	EN
21	Alice Cooper	Art	EN
22	Catherine the Great	Politics	RU
23	14th Dalai Lama	Religion	WR
24	Christina Aguilera	Art	EN
25	Marilyn Monroe	Art	EN
26	David Bowie	Art	EN
27	John McCain	Politics	EN
28	Bob Dylan	Art	EN
29	Johann Sebastian Bach	Art	DE
30	Jesus	Religion	WR

Table S2: Top 30 persons by CheiRank for English Wikipedia with their field of activity and native language.

$R_{EN,CheiRank}$	Person	Field	Culture
1	Roger Calmel	Art	FR
2	C. H. Vijayashankar	Politics	EN
3	Matt Kelley	ETC	EN
4	Alberto Cavallari	ETC	IT
5	Yury Chernavsky	Art	RU
6	William Shakespeare (inventor)	ETC	EN
7	Kelly Clarkson	Art	EN
8	Park Ji-Sung	Sport	KO
9	Mithun Chakraborty	Art	EN
10	Olga Sedakova	Sport	RU
11	Sara García	Art	ES
12	Pope Pius XII	Religion	IT
13	Andy Kerr	Politics	EN
14	Joe-Max Moore	Sport	EN
15	Josef Kemr	Art	WR
16	Darius Milhaud	Art	FR
17	Jan Crull, Jr.	ETC	EN
18	Farshad Fotouhi	Science	EN
19	Swaroop Kanchi	Art	EN
20	Jacques Lancelot	Art	FR
21	František Martin Pecháček	Art	DE
22	George Stephanekoulosech	ETC	EN
23	Chano Urueta	Art	ES
24	Franz Pecháček	Art	DE
25	Nicolae Iorga	Politics	WR
26	Arnold Houbraken	Art	NL
27	August Derleth	Art	EN
28	Javier Solana	Politics	ES
29	Drew Barrymore	Art	EN
30	Kevin Bloody Wilson	Art	EN

Table S4: Top 30 persons by PageRank for French Wikipedia with their field of activity and native language.

$R_{FR,PageRank}$	Person	Field	Culture
1	Napoleon	Politics	FR
2	Carl Linnaeus	Science	WR
3	Louis XIV of France	Politics	FR
4	Jesus	Religion	WR
5	Aristotle	Science	WR
6	Julius Caesar	Politics	IT
7	Charles de Gaulle	Politics	FR
8	Pope John Paul II	Religion	WR
9	Adolf Hitler	Politics	DE
10	Plato	Science	WR
11	Charlemagne	Politics	FR
12	Joseph Stalin	Politics	RU
13	Charles V, Holy Roman Emperor	Politics	ES
14	Napoleon III	Politics	FR
15	Nicolas Sarkozy	Politics	FR
16	Franois Mitterrand	Politics	FR
17	Victor Hugo	Art	FR
18	Jacques Chirac	Politics	FR
19	Honore de Balzac	Art	FR
20	Mary (mother of Jesus)	Religion	WR
21	Voltaire	Art	FR
22	George W. Bush	Politics	EN
23	Elizabeth II	Politics	EN
24	Muhammad	Religion	WR
25	Francis I of France	Politics	FR
26	William Shakespeare	Art	EN
27	Louis XVI of France	Politics	FR
28	Rene Descartes	Science	FR
29	Karl Marx	Science	DE
30	Louis XV of France	Politics	FR

Table S5: Top 30 persons by 2DRank for French Wikipedia with their field of activity and native language.

$R_{FR,2DRank}$	Person	Field	Culture
1	Franois Mitterrand	Politics	FR
2	Jacques Chirac	Politics	FR
3	Honore de Balzac	Art	FR
4	Nicolas Sarkozy	Politics	FR
5	Napoleon III	Politics	FR
6	Otto von Bismarck	Politics	DE
7	Michael Jackson	Art	EN
8	Adolf Hitler	Politics	DE
9	Ludwig van Beethoven	Art	DE
10	Johnny Hallyday	Art	FR
11	Napoleon	Politics	FR
12	Leonardo da Vinci	Art	IT
13	Jules Verne	Art	FR
14	Jacques-Louis David	Art	FR
15	Thomas Jefferson	Politics	EN
16	Sigmund Freud	Science	DE
17	Madonna (entertainer)	Art	EN
18	Serge Gainsbourg	Art	FR
19	14th Dalai Lama	Religion	WR
20	Alfred Hitchcock	Art	EN
21	Georges Clemenceau	Politics	FR
22	Carl Linnaeus	Science	WR
23	Steven Spielberg	Art	EN
24	J. R. R. Tolkien	Art	EN
25	Arthur Rimbaud	Art	FR
26	Charles Darwin	Science	EN
27	Maximilien de Robespierre	Politics	FR
28	Nelson Mandela	Politics	WR
29	Henry IV of France	Politics	FR
30	Charles de Gaulle	Politics	FR

Table S6: Top 30 persons by CheiRank for French Wikipedia with their field of activity and native language.

$R_{FR,CheiRank}$	Person	Field	Culture
1	John Douglas Lynch	Science	EN
2	Roger Federer	Sport	DE
3	Richard Upjohn Light	Science	EN
4	Jacques Davy Duperron	Art	FR
5	Rafael Nadal	Sport	ES
6	Martina Navratilova	Sport	EN
7	Michael Ilmari Saaristo	Science	WR
8	Kevin Bacon	Art	EN
9	Jean Baptiste Eble	Etc	FR
10	Marie-Magdeleine Ayme de La Chevreliere	Politics	FR
11	Nataliya Pyhyda	Sport	RU
12	Max Wolf	Science	DE
13	14th Dalai Lama	Religion	WR
14	Francoise Hardy	Art	FR
15	Ghislaine N. H. Sathoud	Etc	FR
16	Frank Glaw	Science	DE
17	Johnny Hallyday	Art	FR
18	Juan A. Rivero	Science	ES
19	Valentino Rossi	Sport	IT
20	Sheila (singer)	Art	FR
21	Franois Mitterrand	Politics	FR
22	Christopher Walken	Art	EN
23	Georges Clemenceau	Politics	FR
24	Elgin Loren Elwais	Sport	WR
25	Otto von Bismarck	Politics	DE
26	Edward Drinker Cope	Science	EN
27	Rashidi Yekini	Sport	WR
28	Tofiri Kibuuka	Sport	WR
29	Paola Espinosa	Sport	ES
30	Aksana Drahun	Sport	RU

Table S7: Top 30 persons by PageRank for German Wikipedia with their field of activity and native language.

$R_{DE,PageRank}$	Person	Field	Culture
1	Napoleon	Politics	FR
2	Carl Linnaeus	Science	WR
3	Adolf Hitler	Politics	DE
4	Aristotle	Science	WR
5	Johann Wolfgang von Goethe	Art	DE
6	Martin Luther	Religion	DE
7	Jesus	Religion	WR
8	Immanuel Kant	Science	DE
9	Charlemagne	Politics	FR
10	Plato	Science	WR
11	Pope John Paul II	Religion	WR
12	Karl Marx	Science	DE
13	Julius Caesar	Politics	IT
14	Augustus	Politics	IT
15	Louis XIV of France	Politics	FR
16	Friedrich Schiller	Art	DE
17	Wolfgang Amadeus Mozart	Art	DE
18	William Shakespeare	Art	EN
19	Josef Stalin	Politics	RU
20	Pope Benedict XVI	Religion	DE
21	Otto von Bismarck	Politics	DE
22	Cicero	Politics	IT
23	Wilhelm II, German Emperor	Politics	DE
24	Johann Sebastian Bach	Art	DE
25	Max Weber	Science	DE
26	Charles V, Holy Roman Emperor	Politics	ES
27	Frederick the Great	Politics	DE
28	Georg Wilhelm Friedrich Hegel	Science	DE
29	Mary (mother of Jesus)	Religion	WR
30	Augustine of Hippo	Religion	WR

Table S8: Top 30 persons by 2DRank for German Wikipedia with their field of activity and native language.

$R_{DE,2DRank}$	Person	Field	Culture
1	Adolf Hitler	Politics	DE
2	Otto von Bismarck	Politics	DE
3	Pope Paul VI	Religion	IT
4	Ludwig van Beethoven	Art	DE
5	Franz Kafka	Art	DE
6	George Frideric Handel	Art	DE
7	Gerhart Hauptmann	Art	DE
8	Bob Dylan	Art	EN
9	Johann Sebastian Bach	Art	DE
10	Alexander the Great	Politics	WR
11	Martin Luther	Religion	DE
12	Julius Caesar	Politics	IT
13	Joseph Beuys	Art	DE
14	Pope Leo XIII	Religion	IT
15	Carl Friedrich Gauss	Science	DE
16	Andy Warhol	Art	EN
17	Alfred Hitchcock	Art	EN
18	Thomas Mann	Art	DE
19	John Lennon	Art	EN
20	Augustus II the Strong	Politics	DE
21	Pope Benedict XVI	Religion	DE
22	Ferdinand II of Aragon	Politics	ES
23	Arthur Schnitzler	Art	DE
24	Martin Heidegger	Science	DE
25	Albrecht Dürer	Art	DE
26	Carl Linnaeus	Science	WR
27	Pablo Picasso	Art	ES
28	Rainer Werner Fassbinder	Art	DE
29	Wolfgang Amadeus Mozart	Art	DE
30	Historical Jesus	Religion	WR

Table S9: Top 30 persons by CheiRank for German Wikipedia with their field of activity and native language.

$R_{DE,CheiRank}$	Person	Field	Culture
1	Diomede Carafa	Religion	IT
2	Harry Pepl	Art	DE
3	Marc Zwiebler	Sport	DE
4	Eugen Richter	Politics	DE
5	John of Nepomuk	Religion	WR
6	Pope Marcellus II	Religion	IT
7	Karl Wilhelm Reinmuth	Science	WR
8	Johannes Molzahn	Art	DE
9	Georges Vanier	ETC	FR
10	Arthur Willibald Königsheim	ETC	DE
11	Thomas Fitzsimons	Politics	EN
12	Nelson W. Aldrich	Politics	EN
13	Ma Jun	ETC	WR
14	Michael Psellos	Religion	WR
15	Adolf Hitler	Politics	DE
16	Edoardo Fazzioli	ETC	IT
17	Ray Knepper	Sport	EN
18	Frédéric de Lafresnaye	Science	FR
19	Joan Crawford	Art	EN
20	Stephen King	Art	EN
21	Gerhart Hauptmann	Art	DE
22	Paul Moder	Politics	DE
23	Erni Mangold	Art	DE
24	Robert Stolz	Art	DE
25	Otto von Bismarck	Politics	DE
26	Christine Holstein	Art	DE
27	Pope Paul VI	Religion	IT
28	Franz Buxbaum	Science	DE
29	Gustaf Gründgens	Art	DE
30	Ludwig van Beethoven	Art	DE

Table S10: Top 30 persons by PageRank for Italian Wikipedia with their field of activity and native language.

$R_{IT,PageRank}$	Person	Field	Culture
1	Napoleon	Politics	FR
2	Jesus	Religion	WR
3	Aristotle	Science	WR
4	Augustus	Politics	IT
5	Pope John Paul II	Religion	WR
6	Dante Alighieri	Art	IT
7	Adolf Hitler	Politics	DE
8	Julius Caesar	Politics	IT
9	Benito Mussolini	Politics	IT
10	Charlemagne	Politics	FR
11	Mary (mother of Jesus)	Religion	WR
12	Plato	Science	WR
13	Isaac Newton	Science	EN
14	Charles V, Holy Roman Emperor	Politics	ES
15	Galileo Galilei	Science	IT
16	Louis XIV of France	Politics	FR
17	Constantine the Great	Politics	IT
18	Cicero	Politics	IT
19	Alexander the Great	Politics	WR
20	Paul the Apostle	Politics	WR
21	Albert Einstein	Science	DE
22	Joseph Stalin	Politics	RU
23	George W. Bush	Politics	EN
24	Silvio Berlusconi	Politics	IT
25	William Shakespeare	Art	EN
26	Augustine of Hippo	Religion	WR
27	Pope Paul VI	Religion	IT
28	Pope Benedict XVI	Religion	DE
29	Giuseppe Garibaldi	Politics	IT
30	Leonardo da Vinci	Science	IT

Table S11: Top 30 persons by 2DRank for Italian Wikipedia with their field of activity and native language.

$R_{IT,2DRank}$	Person	Field	Culture
1	Pope John Paul II	Religion	WR
2	Pope Benedict XVI	Religion	DE
3	Giuseppe Garibaldi	Politics	IT
4	Raphael	Art	IT
5	Jesus	Religion	WR
6	Benito Mussolini	Politics	IT
7	Michelangelo	Art	IT
8	Leonardo da Vinci	Art	IT
9	Pier Paolo Pasolini	Art	IT
10	Michael Jackson	Art	EN
11	Martina Navratilova	Sport	EN
12	Saint Peter	Religion	WR
13	Pope Paul III	Religion	IT
14	Wolfgang Amadeus Mozart	Art	DE
15	John Lennon	Art	EN
16	Bob Dylan	Art	EN
17	Mina (singer)	Art	IT
18	William Shakespeare	Art	EN
19	Julius Caesar	Politics	IT
20	Titian	Art	IT
21	Silvio Berlusconi	Politics	IT
22	Alexander the Great	Politics	WR
23	Pablo Picasso	Art	ES
24	Antonio Vivaldi	Art	IT
25	Ludwig van Beethoven	Art	DE
26	Napoleon	Politics	FR
27	Madonna (entertainer)	Art	EN
28	Roger Federer	Sport	DE
29	Johann Sebastian Bach	Art	DE
30	Walt Disney	Art	EN

Table S12: Top 30 persons by CheiRank for Italian Wikipedia with their field of activity and native language.

$R_{IT,CheiRank}$	Person	Field	Culture
1	Ticone di Amato	Religion	WR
2	John the Merciful	Religion	WR
3	Nduccio	Art	IT
4	Vincenzo Olivieri	Art	IT
5	Leo Baeck	Religion	DE
6	Karl Wilhelm Reinmuth	Science	DE
7	Freimut Börngen	Science	DE
8	Nikolai Chernykh	Science	RU
9	Edward L. G. Bowell	Science	EN
10	Roger Federer	Sport	DE
11	Michel Morganella	Sport	WR
12	Rafael Nadal	Sport	ES
13	Robin Söderling	Sport	WR
14	Iván Zamorano	Sport	ES
15	Martina Navratilova	Sport	EN
16	Venus Williams	Sport	EN
17	Goran Ivanišević	Sport	WR
18	Javier Pastore	Sport	ES
19	Stevan Jovetić	Sport	WR
20	Mina (singer)	Art	IT
21	George Ade	Art	EN
22	Kazuro Watanabe	Sport	WR
23	Andy Roddick	Sport	EN
24	Johann Strauss II	Art	DE
25	Max Wolf	Science	DE
26	Isaac Asimov	Art	EN
27	Georges Simenon	Art	FR
28	Alice Joyce	Art	EN
29	Pietro De Sensi	Sport	IT
30	Noemi (singer)	Art	IT

Table S13: Top 30 persons by PageRank for Spanish Wikipedia with their field of activity and native language.

$R_{ES,PageRank}$	Person	Field	Culture
1	Carl Linnaeus	Scinece	WR
2	Napoleon	Politics	FR
3	Jesus	Religion	WR
4	Aristotle	Science	WR
5	Charles V, Holy Roman Emperor	Politics	ES
6	Adolf Hitler	Politics	DE
7	Julius Caesar	Politics	IT
8	Philip II of Spain	Politics	ES
9	William Shakespeare	Art	EN
10	Plato	Science	WR
11	Albert Einstein	Science	DE
12	Augustus	Politics	IT
13	Pope John Paul II	Religion	WR
14	Christopher Columbus	ETC	IT
15	Karl Marx	Science	DE
16	Alexander the Great	Politics	WR
17	Isaac Newton	Science	EN
18	Francisco Franco	Politics	ES
19	Charlemagne	Politics	FR
20	Immanuel Kant	Science	DE
21	Charles Darwin	Science	EN
22	Louis XIV of France	Politics	FR
23	Mary (mother of Jesus)	Religion	WR
24	Wolfgang Amadeus Mozart	Art	DE
25	Galileo Galilei	Science	IT
26	Cicero	Politics	IT
27	Homer	Art	WR
28	Paul the Apostle	Religion	WR
29	René Descartes	Science	FR
30	Miguel de Cervantes	Art	ES

Table S14: Top 30 persons by 2DRank for Spanish Wikipedia with their field of activity and native language.

$R_{ES,2DRank}$	Person	Field	Culture
1	Wolfgang Amadeus Mozart	Art	DE
2	Julius Caesar	Politics	IT
3	Simón Bolívar	Politics	ES
4	Francisco Goya	Art	ES
5	Madonna (entertainer)	Art	EN
6	Bob Dylan	Art	EN
7	Barack Obama	Politics	EN
8	Fidel Castro	Politics	ES
9	Michael Jackson	Art	EN
10	Richard Wagner	Art	DE
11	Augusto Pinochet	Politics	ES
12	Trajan	Politics	IT
13	Jorge Luis Borges	Art	ES
14	Juan Perón	Politics	ES
15	Porfirio Díaz	Politics	ES
16	Michelangelo	Art	IT
17	J. R. R. Tolkien	Art	EN
18	Paul McCartney	Art	EN
19	Adolf Hitler	Politics	DE
20	John Lennon	Art	EN
21	Hugo Chávez	Politics	ES
22	Elizabeth II	Politics	EN
23	Lope de Vega	Art	ES
24	Francisco Franco	Politics	ES
25	Christopher Columbus	ETC	IT
26	Diego Velázquez	Art	ES
27	Pablo Picasso	Art	ES
28	Edgar Allan Poe	Art	EN
29	Charlemagne	Politics	FR
30	Juan Carlos I of Spain	Politics	ES

Table S15: Top 30 persons by CheiRank for Spanish Wikipedia with their field of activity and native language.

$R_{ES,CheiRank}$	Person	Field	Culture
1	Max Wolf	Science	DE
2	Monica Bellucci	Art	IT
3	Che Guevara	Politics	ES
4	Steve Buscemi	Art	EN
5	Johann Palisa	Science	DE
6	Auguste Charlois	Science	FR
7	José Flávio Pessoa de Barros	Science	WR
8	Arturo Mercado	Art	ES
9	Francisco Goya	Art	ES
10	Bob Dylan	Art	EN
11	Jorge Luis Borges	Art	ES
12	Brian May	Art	EN
13	Virgilio Barco Vargas	Politics	ES
14	Mariano Bellver	ETC	ES
15	Demi Lovato	Art	EN
16	Joan Manuel Serrat	Art	ES
17	Mary Shelley	Art	EN
18	Ana Belén	Art	ES
19	Aki Misato	Art	WR
20	Carl Jung	Science	DE
21	Roger Federer	Sport	DE
22	Antoni Gaudí	Art	ES
23	Rafael Nadal	Sport	ES
24	Hans Melchior	Science	DE
25	Paulina Rubio	Art	ES
26	Paul McCartney	Art	EN
27	Julieta Venegas	Art	ES
28	Fermin Muguruza	Art	ES
29	Belinda (entertainer)	Art	ES
30	Patricia Acevedo	Art	ES

Table S16: Top 30 persons by PageRank for Dutch Wikipedia with their field of activity and native language.

$R_{NL,PageRank}$	Person	Field	Culture
1	Carl Linnaeus	Science	WR
2	Pierre Andre Latreille	Science	FR
3	Napoleon	Politics	FR
4	Eugene Simon	Science	FR
5	Jesus	Religion	WR
6	Charles Darwin	Science	EN
7	Julius Caesar	Politics	IT
8	Adolf Hitler	Politics	DE
9	Aristotle	Science	WR
10	Charlemagne	Politics	FR
11	Plato	Science	WR
12	Jean-Baptiste Lamarck	Science	FR
13	Ernst Mayr	Science	DE
14	Alexander the Great	Politics	WR
15	Louis XIV of France	Politics	FR
16	Pope John Paul II	Religion	WR
17	Alfred Russel Wallace	Science	EN
18	Charles V, Holy Roman Emperor	Politics	ES
19	Thomas Robert Malthus	Science	EN
20	Augustus	Politics	IT
21	William I of the Netherlands	Politics	NL
22	Joseph Stalin	Politics	RU
23	Albert Einstein	Science	DE
24	Beatrix of the Netherlands	Politics	NL
25	Christopher Columbus	Etc	IT
26	Elizabeth II	Politics	EN
27	Isaac Newton	Science	EN
28	Wolfgang Amadeus Mozart	Art	DE
29	J. B. S. Haldane	Science	EN
30	Cicero	Politics	IT

Table S17: Top 30 persons by 2DRank for Dutch Wikipedia with their field of activity and native language.

$R_{NL,2DRank}$	Person	Field	Culture
1	Pope Benedict XVI	Religion	DE
2	Elizabeth II	Politics	EN
3	Charles Darwin	Science	EN
4	Albert II of Belgium	Politics	NL
5	Albert Einstein	Science	DE
6	Pope John Paul II	Religion	WR
7	Michael Jackson	Art	EN
8	Johann Sebastian Bach	Art	DE
9	Saint Peter	Religion	WR
10	Johan Cruyff	Sport	NL
11	William Shakespeare	Art	EN
12	Christopher Columbus	Etc	IT
13	Augustus	Politics	IT
14	Frederick the Great	Politics	DE
15	Rembrandt	Art	NL
16	Eddy Merckx	Sport	NL
17	Ludwig van Beethoven	Art	DE
18	Pope Pius XII	Religion	IT
19	Peter Paul Rubens	Art	NL
20	Napoleon	Politics	FR
21	Wolfgang Amadeus Mozart	Art	DE
22	Igor Stravinsky	Art	RU
23	Martin of Tours	Religion	FR
24	Geert Wilders	Politics	NL
25	J.R.R. Tolkien	Art	EN
26	Pierre Cuypers	Art	NL
27	Charles V, Holy Roman Emperor	Politics	ES
28	Pope Pius IX	Religion	IT
29	Juliana of the Netherlands	Politics	NL
30	Elvis Presley	Art	EN

Table S18: Top 30 persons by CheiRank for Dutch Wikipedia with their field of activity and native language.

$R_{NL,CheiRank}$	Person	Field	Culture
1	Pier Luigi Bersani	Politics	IT
2	Francesco Rutelli	Politics	IT
3	Hans Renders	Science	NL
4	Julian Jenner	Sport	NL
5	Marten Toonder	Art	NL
6	Uwe Seeler	Sport	DE
7	Stefanie Sun	Art	WR
8	Roger Federer	Sport	DE
9	Theo Janssen	Sport	NL
10	Zazie	Art	FR
11	Albert II of Belgium	Politics	NL
12	Denny Landzaat	Sport	NL
13	Paul Biegel	Art	NL
14	Guido De Padt	Politics	NL
15	Jan Knippenberg	Sport	NL
16	Michael Schumacher	Sport	DE
17	Hans Werner Henze	Art	DE
18	Lionel Messi	Sport	ES
19	Johan Cruijff	Sport	NL
20	Eva Janssen (actrice)	Art	NL
21	Marion Zimmer Bradley	Art	EN
22	Graham Hill	Sport	EN
23	Rick Wakeman	Art	EN
24	Mihai Nesu	Sport	NL
25	Freddy De Chou	Politics	NL
26	Rubens Barrichello	Sport	WR
27	Ismail Aissati	Sport	NL
28	Marco van Basten	Sport	NL
29	Paul Geerts	Art	NL
30	Ibrahim Afellay	Sport	NL

Table S19: Top 30 persons by PageRank for Russian Wikipedia with their field of activity and native language.

$R_{RU,PageRank}$	Person	Field	Culture
1	Peter the Great	Politics	RU
2	Napoleon	Politics	FR
3	Carl Linnaeus	Science	WR
4	Joseph Stalin	Politics	RU
5	Alexander Pushkin	Art	RU
6	Vladimir Lenin	Politics	RU
7	Catherine the Great	Politics	RU
8	Jesus	Religion	WR
9	Aristotle	Science	WR
10	Vladimir Putin	Politics	RU
11	Julius Caesar	Politics	IT
12	Adolf Hitler	Politics	DE
13	Boris Yeltsin	Politics	RU
14	William Shakespeare	Art	EN
15	Ivan the Terrible	Politics	RU
16	Alexander II of Russia	Politics	RU
17	Nicholas II of Russia	Politics	RU
18	Karl Marx	Science	DE
19	Louis XIV of France	Politics	FR
20	Nicholas I of Russia	Politics	RU
21	Alexander I of Russia	Politics	RU
22	Alexander the Great	Politics	WR
23	Charlemagne	Politics	FR
24	William Herschel	Science	EN
25	Mikhail Gorbachev	Politics	RU
26	Paul I of Russia	Politics	RU
27	Leo Tolstoy	Art	RU
28	Nikolai Gogol	Art	RU
29	Dmitry Medvedev	Politics	RU
30	Lomonosov	Science	RU

Table S20: Top 30 persons by 2DRank for Russian Wikipedia with their field of activity and native language.

$R_{RU,2DRank}$	Person	Field	Culture
1	Dmitri Mendeleev	Science	RU
2	Peter the Great	Politics	RU
3	Justinian I	Politics	WR
4	Yaroslav the Wise	Politics	RU
5	Elvis Presley	Art	EN
6	Yuri Gagarin	Etc	RU
7	William Shakespeare	Art	EN
8	Albert Einstein	Science	DE
9	Adolf Hitler	Politics	DE
10	Christopher Columbus	Etc	IT
11	Catherine the Great	Politics	RU
12	Vladimir Vysotsky	Art	RU
13	Louis de Funes	Art	FR
14	Lomonosov	Science	RU
15	Alla Pugacheva	Art	RU
16	Viktor Yanukovych	Politics	RU
17	Nikolai Gogol	Art	RU
18	Felix Dzerzhinsky	Politics	RU
19	Aleksandr Solzhenitsyn	Art	RU
20	Pope Benedict XVI	Religion	DE
21	Maxim Gorky	Art	RU
22	Julius Caesar	Politics	IT
23	George Harrison	Art	EN
24	Bohdan Khmelnytsky	Politics	RU
25	Rembrandt	Art	NL
26	John Lennon	Art	EN
27	Jules Verne	Art	FR
28	Benito Mussolini	Politics	IT
29	Nicholas Roerich	Art	RU
30	Niels Bohr	Science	WR

Table S21: Top 30 persons by CheiRank for Russian Wikipedia with their field of activity and native language.

$R_{RU,CheiRank}$	Person	Field	Culture
1	Aleksander Vladimirovich Sotnik	Etc	RU
2	Aleksei Aleksandrovich Bobrinsky	Politics	RU
3	Boris Grebenshchikov	Art	RU
4	Karl Wilhelm Reinmuth	Science	DE
5	Ronnie O'Sullivan	Sport	EN
6	Max Wol	Science	DE
7	Ivan Egorovich Sizykh	Etc	RU
8	Vladimir Mikhilovich Popkov	Art	RU
9	Sun Myung Moon	Religion	KO
10	Mikhail Pavlovich Tolstoi	Etc	RU
11	Perry Como	Art	EN
12	John Heenan	Religion	EN
13	Petr Aleksandrovich Ivaschenko	Art	RU
14	Andrey Vlasov	Etc	RU
15	Christian Heinrich Friedrich Peters	Science	DE
16	Auguste Charlois	Science	FR
17	Damian (Marczhuk)	Religion	RU
18	Yuri Gagarin	Etc	RU
19	Stephen Hendry	Sport	EN
20	Ivan Grigorevich Donskikh	Etc	RU
21	Anna Semenovna Kamenkova-Pavlova	Art	RU
22	Ivan Nikolaevich Shulga	Art	RU
23	George Dwyer	Religion	EN
24	William Wheeler (bishop)	Religion	EN
25	Vladimir Vladimirovitsch Antonik	Art	RU
26	Leonid Parfyonov	Art	RU
27	Vincent Nichols	Religion	EN
28	Dmitri Mendeleev	Science	RU
29	Boris Vladimirovich Bakin	Etc	RU
30	George Harrison	Art	EN

Table S22: Top 30 persons by PageRank for Hungarian Wikipedia with their field of activity and native language.

$R_{HU,PageRank}$	Person	Field	Culture
1	Carl Linnaeus	Science	WR
2	Jesus	Religion	WR
3	Napoleon	Politics	FR
4	Aristotle	Science	WR
5	Julius Caesar	Politics	IT
6	Matthias Corvinus	Politics	HU
7	Szentagothai Janos	Science	HU
8	William Shakespeare	Art	EN
9	Adolf Hitler	Politics	DE
10	Stephen I of Hungary	Politics	HU
11	Augustus	Politics	IT
12	Michael Schumacher	Sport	DE
13	Miklos Rethelyi	Politics	HU
14	Sigismund, Holy Roman Emperor	Politics	HU
15	Lajos Kossuth	Politics	HU
16	Charles I of Hungary	Politics	HU
17	Bela IV of Hungary	Politics	HU
18	Maria Theresa	Politics	DE
19	Joseph Stalin	Politics	RU
20	Franz Joseph I of Austria	Politics	DE
21	Louis I of Hungary	Politics	HU
22	Francis II Rakoczi	Politics	HU
23	Mary (mother of Jesus)	Religion	WR
24	Sandor Petofi	Art	HU
25	Pope John Paul II	Religion	WR
26	Johann Wolfgang von Goethe	Art	DE
27	Alexander the Great	Politics	WR
28	Bela Bartok	Art	HU
29	Charlemagne	Politics	FR
30	Louis XIV of France	Politics	FR

Table S23: Top 30 persons by 2DRank for Hungarian Wikipedia with their field of activity and native language.

$R_{HU,2DRank}$	Person	Field	Culture
1	Stephen I of Hungary	Politics	HU
2	Sandor Petofi	Art	HU
3	Franz Liszt	Art	HU
4	Kati Kovacs	Art	HU
5	Alexander the Great	Politics	WR
6	Attila Jozsef	Art	HU
7	Aristotle	Science	WR
8	Kimi Raikkonen	Sport	WR
9	Rubens Barrichello	Sport	WR
10	Lajos Kossuth	Politics	HU
11	Bela Bartok	Art	HU
12	Charlemagne	Politics	FR
13	Sandor Weores	Art	HU
14	Mariah Carey	Art	EN
15	Wolfgang Amadeus Mozart	Art	DE
16	Josip Broz Tito	Politics	WR
17	Charles I of Hungary	Politics	HU
18	Isaac Asimov	Art	EN
19	Napoleon	Politics	FR
20	Bonnie Tyler	Art	EN
21	Miklos Radnoti	Art	HU
22	Jay Chou	Art	WR
23	Janos Kodolanyi	Art	HU
24	Louis I of Hungary	Politics	HU
25	Zsuzsa Koncz	Art	HU
26	Adolf Hitler	Politics	HU
27	Stephen King	Art	EN
28	Mor Jokai	Art	HU
29	Ferenc Erkel	Art	HU
30	Franz Joseph I of Austria	Politics	DE

Table S24: Top 30 persons by CheiRank for Hungarian Wikipedia with their field of activity and native language.

$R_{HU,CheiRank}$	Person	Field	Culture
1	Edward L. G. Bowell	Science	EN
2	Karl Wilhelm Reinmuth	Science	DE
3	Max Wolf	Science	DE
4	Benjamin Boukpeti	Sport	FR
5	Urata Takesi	Science	WR
6	Wilfred Bungei	Sport	WR
7	Henri Debehogne	Science	FR
8	Lee "Scratch" Perry	Art	WR
9	Karl Golsdorf	Etc	DE
10	Johann Palisa	Science	DE
11	Dirk Kuijt	Sport	NL
12	Roger Federer	Sport	DE
13	Csernus Imre	Etc	HU
14	Kati Kovacs	Art	HU
15	Rafael Nadal	Sport	ES
16	Venus Williams	Sport	EN
17	Sebastien Loeb	Sport	FR
18	Pleh Csaba	Science	HU
19	Tibor Antalpeter	Sport	HU
20	Serena Williams	Sport	EN
21	Csore Gabor	Art	HU
22	Pirmin Schwegler	Sport	DE
23	Olivia Newton-John	Art	EN
24	Petter Solberg	Sport	WR
25	Orosz Anna	Art	HU
26	Zsambeki Gabor	Art	HU
27	Vera Igorevna Zvonarjova	Sport	RU
28	Sandor Petofi	Art	HU
29	Roberta Vinci	Sport	IT
30	Flavia Pennetta	Sport	HU

Table S25: Top 30 persons by PageRank for Korean Wikipedia with their field of activity and native language.

$R_{KO,PageRank}$	Person	Field	Culture
1	Carl Linnaeus	Science	WR
2	Gojong of the Korean Empire	Politics	KO
3	Jesus	Religion	WR
4	John Edward Gray	Science	EN
5	Aristotle	Science	WR
6	Napoleon	Politics	FR
7	Sejong the Great	Politics	KO
8	Park Chung-hee	Politics	KO
9	Emperor Wu of Han	Politics	WR
10	Seonjo of Joseon	Politics	KO
11	Taejong of Joseon	Politics	KO
12	Syngman Rhee	Politics	KO
13	Kim Dae-jung	Politics	KO
14	Roh Moo-hyun	Politics	KO
15	Yeongjo of Joseon	Politics	KO
16	Adolf Hitler	Politics	DE
17	Taejo of Joseon	Politics	KO
18	Sukjong of Joseon	Politics	KO
19	Kim Il-sung	Politics	KO
20	Qianlong Emperor	Politics	WR
21	Kim Jong-il	Politics	KO
22	Kangxi Emperor	Politics	WR
23	Emperor Gaozu of Han	Politics	WR
24	Chun Doo-hwan	Politics	KO
25	Taejo of Goryeo	Politics	KO
26	George W. Bush	Politics	EN
27	Qin Shi Huang	Politics	WR
28	Jeongjo of Joseon	Politics	KO
29	Sunjo of Joseon	Politics	KO
30	Cao Cao	Politics	WR

Table S26: Top 30 persons by 2DRank for Korean Wikipedia with their field of activity and native language.

$R_{KO,2DRank}$	Person	Field	Culture
1	Gojong of the Korean Empire	Politics	KO
2	Sejong the Great	Politics	KO
3	Park Chung-hee	Politics	KO
4	Taejong of Joseon	Politics	KO
5	Kim Dae-jung	Politics	KO
6	Roh Moo-hyun	Politics	KO
7	Syngman Rhee	Politics	KO
8	Kim Il-sung	Politics	KO
9	Qianlong Emperor	Politics	WR
10	Kangxi Emperor	Politics	WR
11	Taejo of Goryeo	Politics	KO
12	Seonjo of Joseon	Politics	KO
13	Jeongjo of Joseon	Politics	KO
14	Kim Young-sam	Politics	KO
15	Julius Caesar	Politics	IT
16	Chun Doo-hwan	Politics	KO
17	Injo of Joseon	Politics	KO
18	Tokugawa Ieyasu	Politics	WR
19	Lee Myung-bak	Politics	KO
20	Seongjong of Joseon	Politics	KO
21	Cao Cao	Politics	WR
22	Confucius	Science	WR
23	Mao Zedong	Politics	WR
24	Taejo of Joseon	Politics	KO
25	Toyotomi Hideyoshi	Politics	WR
26	Heungseon Daewongun	Politics	KO
27	Liu Bei	Politics	WR
28	Yeongjo of Joseon	Politics	KO
29	Pope John Paul II	Religion	WR
30	Adolf Hitler	Politics	DE

Table S27: Top 30 persons by CheiRank for Korean Wikipedia with their field of activity and native language.

$R_{KO,CheiRank}$	Person	Field	Culture
1	Lee Jong-wook (baseball)	Sport	KO
2	Kim Dae-jung	Politics	KO
3	Lionel Messi	Sport	ES
4	Kim Kyu-sik	Politics	KO
5	Johannes Kepler	Science	DE
6	Yun Chi-young	Politics	KO
7	Michael Jackson	Art	EN
8	Yi Sun-sin	ETC	KO
9	Chang Myon	Politics	KO
10	IU (singer)	Art	KO
11	Kim Seo-yeong	Art	KO
12	Tokugawa Ieyasu	Politics	WR
13	Jeremy Renner	Art	EN
14	Zhao Deyin	Politics	WR
15	Yang Joon-Hyu	Sport	KO
16	Zhang Gui (Tang Dynasty)	Politics	WR
17	Zinedine Zidane	Sport	FR
18	Park Chung-hee	Politics	KO
19	Heungseon Daewongun	Politics	KO
20	Ahn Ji-hwan	Art	KO
21	Lee Seung-Yeop	Sport	KO
22	Roh Moo-hyun	Politics	KO
23	Britney Spears	Art	EN
24	Kim Young-sam	Politics	KO
25	Jeong Hyeong-don	Art	KO
26	Kim Yu-Na	Sport	KO
27	Park Jong-Seol	Art	KO
28	Lim Taekyoung	Art	KO
29	Park Ji-Sung	Sport	KO
30	Yuh Woon-Hyung	Politics	KO

Uncovering disassortativity in large scale-free networks

Nelly Litvak*
University of Twente

Remco van der Hofstad†
Eindhoven University of Technology

(Dated: February 22, 2013)

Abstract

Mixing patterns in large self-organizing networks, such as the Internet, the World Wide Web, social and biological networks are often characterized by degree-degree dependencies between neighbouring nodes. In this paper we propose a new way of measuring degree-degree dependencies. One of the problems with the commonly used assortativity coefficient is that in disassortative networks its magnitude decreases with the network size. We mathematically explain this phenomenon and validate the results on synthetic graphs and real-world network data. As an alternative, we suggest to use rank correlation measures such as Spearman's rho. Our experiments convincingly show that Spearman's rho produces consistent values in graphs of different sizes but similar structure, and it is able to reveal strong (positive or negative) dependencies in large graphs. In particular, we discover much stronger negative degree-degree dependencies in Web graphs than was previously thought. Rank correlations allow us to compare the assortativity of networks of different sizes, which is impossible with the assortativity coefficient due to its genuine dependence on the network size. We conclude that rank correlations provide a suitable and informative method for uncovering network mixing patterns.

* n.litvak@ewi.utwente.nl
† r.w.v.d.hofstad@TUE.nl

I. INTRODUCTION

This paper proposes a new way of measuring mixing patterns in large self-organizing networks, such as the Internet, the World Wide Web, social and biological networks. Most of these real-world networks are scale-free, i.e., their degree distribution has huge variability and closely follows a power law (the fraction of nodes with degree k is roughly proportional to $k^{-\gamma-1}$, $\gamma > 0$). We study correlations between degrees of two nodes connected by an edge. This problem, first posed in [1, 2], has received vast attention in the networks literature, in particular in physics, sociology, biology and computer science. We show however, analytically and on the data, that the presence of power laws makes currently used measures inadequate for comparison of mixing patterns in networks of different sizes, and provide an alternative that is free from this disadvantage.

Adequate measuring and comparison of degree-degree correlations is important because mixing patterns define many of the network's properties. For instance, the Internet topology is not sufficiently specified by the degree distribution; the negative degree-degree correlations in the Internet graph have a great influence on the robustness to failures [3], efficiency of Internet protocols [4], as well as distances and betweenness [5]. This is totally different from the mixing patterns in networks of bank transactions [6] where the core of 25 most important banks is entirely connected. The correlation between in- and out-degree of tasks plays and important role in the dynamics of production and development systems [7]. Mixing patterns affect epidemic spread [8, 9] and Web ranking [10].

In his seminal papers, Newman [1, 2] proposed to measure degree-degree correlations using the *assortativity* coefficient, which is, in fact, an empirical estimate of the Pearson's correlation coefficient between the degrees at either ends of a random edge. A network is *assortative* when neighbouring nodes are likely to have a similar number of connections. In *disassortative* networks, high-degree nodes mostly have neighbours with small number of connections. The empirical data in [1, Table I] suggest that social networks tend to be assortative (which is indicated by the positive assortativity coefficient), while technological and biological networks tend to be disassortative.

In [1, Table I], it is striking that larger disassortative networks typically have an assortativity coefficient that is closer to 0 and therefore appear to have approximately *uncorrelated* degrees across edges. Similar conclusions can be drawn from [2, Table II]. In recent literature [11, 12] the issue was raised that the Pearson's correlation coefficient in scale-free networks decreases with the network size. In this paper we demonstrate analytically and on the data that in *all* scale-free disassortative networks with a realistic value of the power-law exponent, the assortativity coefficient decreases in magnitude with the size of the graph. In assortative networks, on the other hand, the assortativity coefficient can show two types of behaviour. It either decreases with graph size, or it shows a considerable dispersion in values, even if large networks are constructed by the same mechanism.

We suggest an alternative solution based on the classical Spearman's rho measure [13] that is the correlation coefficient computed on the *ranks* of degrees. The huge advantage of such dependency measures is that they work well *independently* of the degree distribution, while the assortativity coefficient, despite the fact that it is always in $[-1, 1]$, suffers from a strong dependence on the extreme values of the degrees. The usefullness of the rank correlation approach to discover dependencies in skewed distributions has already been postulated in the 1936 paper by H. Hotelling and M.R. Pabst [14]: *'Certainly where there is complete absence of knowledge of the form of the bivariate distribution, and especially if it is believed not to be normal, the rank correlation coefficient is to be strongly recommended as a means of testing the existence of relationship.'*

We compute Spearman's rho on artificially generated random graphs and on real data

from web and social networks. Our results agree with [1] concerning the presence of positive or negative correlations, but Spearman's rho has two important advantages: (1) it is able to reveal strong disassortativity in large networks; (2) it produces consistent values on the graphs created by the same mechanism, e.g. on preferential attachment graphs [15] of different sizes. Thus, Spearman's rho correctly and consistently captures the underlying connection patterns and tendencies. We conclude that when networks are large, or two networks of difference sizes must be compared (e.g. in web crawls or social networks from different countries), Spearman's rho is a preferred method for measuring and comparing degree-degree correlations.

The closing section discusses further challenges in the evaluation of network mixing patterns.

II. NO DISASSORTATIVE SCALE-FREE RANDOM GRAPH SEQUENCES

In this section we present a simple analytical argument that in disassortative networks the assortativity coefficient always decreases in magnitude with the size of the graph. Formal proofs can be found in [16].

Assortativity in networks is usually measured using the assortativity coefficient, which is in fact a statistical estimator of a Pearson's correlation coefficient for the degrees on the two ends of an arbitrary edge in a graph. Let $G = (V, E)$ be a graph with vertex set V, where $|V| = n$ denotes the size of the network, and edge set E. The assortativity coefficient of G is equal to (see, e.g., [1, (4)])

$$\rho_n = \frac{\frac{1}{|E|}\sum_{ij\in E} d_i d_j - \left(\frac{1}{|E|}\sum_{ij\in E}\frac{1}{2}(d_i+d_j)\right)^2}{\frac{1}{|E|}\sum_{ij\in E}\frac{1}{2}(d_i^2+d_j^2) - \left(\frac{1}{|E|}\sum_{ij\in E}\frac{1}{2}(d_i+d_j)\right)^2}, \tag{II.1}$$

where the sum is over directed edges of G, i.e., ij and ji are two distinct edges, and d_i is the degree of vertex i. We compute that

$$\frac{1}{|E|}\sum_{ij\in E}\frac{1}{2}(d_i+d_j) = \frac{1}{|E|}\sum_{i\in V} d_i^2, \qquad \frac{1}{|E|}\sum_{ij\in E}\frac{1}{2}(d_i^2+d_j^2) = \frac{1}{|E|}\sum_{i\in V} d_i^3.$$

Thus, ρ_n can be written as

$$\rho_n = \frac{\sum_{ij\in E} d_i d_j - \frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}{\sum_{i\in V} d_i^3 - \frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}. \tag{II.2}$$

In practice, all quantities in (II.2) are finite, and ρ_n can always be computed. However, since many real-life networks are very large, a relevant question is how ρ_n behaves when n becomes large.

In the literature, many examples are reported of real-world networks where the degree distribution obeys a power law [17, 18]. In particular, for scale-free networks, the observed proportion of vertices of degree k is close to $f(k) = c_0 k^{-\gamma-1}$, and most values of γ found in real-world networks are in $(1, 3)$, see e.g., [17, Table I] or [18, Table I]. For $p < \gamma$, let $\mu_p = \sum_k k^p f(k)$, and note that the series diverges if $p \geq \gamma$; let $a \sim b$ denote that $a/b \to 1$. Then we can expect that, as n grows large,

$$|E| = \sum_{i\in V} d_i \sim \mu_1 n, \quad \sum_{i\in V} d_i^p \sim \mu_p n, \quad p < \gamma,$$

while $\max_{i\in V} d_i$ is of the order $n^{1/\gamma}$. As a direct consequence,

$$cn \le |E| \le Cn, \tag{II.3}$$

$$cn^{1/\gamma} \le \max_{i\in[n]} d_i \le Cn^{1/\gamma}, \tag{II.4}$$

$$cn^{\max\{p/\gamma,1\}} \le \sum_{i\in[n]} d_i^p \le Cn^{\max\{p/\gamma,1\}}, \quad p = 2,3, \tag{II.5}$$

for $\gamma \in (1,3)$ and some constants $0 < c < C < \infty$. We emphasize that conditions (II.3) – (II.5) are very general and hold for any scale-free network of growing size, independently of its mixing patterns. From (II.2) we simply write

$$\rho_n \ge \rho_n^- \equiv -\frac{\frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2}{\sum_{i\in V} d_i^3 - \frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2},$$

and notice that

$$\sum_{i\in V} d_i^3 \ge (\max_{i\in[n]} d_i)^3 \ge c^3 n^{3/\gamma},$$

whereas

$$\frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2 \le (C^2/c)n^{2\max\{2/\gamma,1\}-1} = (C^2/c)n^{\max\{4/\gamma-1,1\}}.$$

Since $\gamma \in (1,3)$ we have $\max\{4/\gamma - 1, 1\} < 3/\gamma$, so that

$$\frac{\sum_{i\in V} d_i^3}{\frac{1}{|E|}\Big(\sum_{i\in V} d_i^2\Big)^2} \to \infty \quad \text{as } t \to \infty.$$

Hence, the lower bound ρ_n^- is of the order $n^{\max\{1/\gamma-1,1-3/\gamma\}}$. It is now easy to check that if $\gamma \in (1,3)$, then ρ_n^- converges to zero when the graph size increases. This means that any limit point of the assortativity coefficients ρ_n is non-negative. Note also that ρ_n^- is defined by the degree sequence, and it does not depend on the mixing pattern at all. We conclude that by looking only at the value of ρ_n one cannot discover even very strong disassortativity in large scale-free graphs. We will confirm this finding in Section IV on artificially generated random graphs, and in Section V on real-world networks.

We note that if $\gamma > 3$, then all terms in (II.1) converge to a number, and ρ_n does not scale with the network size. In practice this means that the dependence of ρ_n on the graph size is observed when node degrees have a broad distribution, and this range increases when the network gets bigger. This is the case in most real-life networks and models for them, as is e.g. obviously the case for preferential attachment models.

We further notice that (II.3)–(II.5) imply that

$$\sum_{ij\in E} d_i d_j \le \Big(\max_{i\in[n]} d_i\Big) \sum_{ij\in E_n} d_i = \max_{i\in[n]} d_i \Big(\sum_{i\in V_n} d_i^2\Big) \le C^2 n^{1/\gamma+\max\{2/\gamma,1\}}. \tag{II.6}$$

Mathematically, an interesting case is when $\sum_{ij\in E} d_i d_j$ and $\sum_{i\in V} d_i^3$ are of the same order of magnitude. Then the network is assortative but, formally, ρ_n converges to a random variable. In practice this means that ρ_n can result in very different values on two very large graphs constructed by the same mechanism. We will give such an example in Section IV.

III. RANK CORRELATIONS

We propose an alternative measure for the degree-degree dependencies, based on the rank correlations. For two-dimensional data $((X_i, Y_i))_{i=1}^n$, let r_i^X and r_i^Y be the rank of an observation X_i and Y_i, respectively, when the sample values $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$ are arranged in a descending order. The rank correlation measures evaluate statistical dependences on the data $((r_i^X, r_i^Y))_{i=1}^n$, rather than on the original data $((X_i, Y_i))_{i=1}^n$. Rank transformation is convenient, in particular because (r_i^X) and (r_i^Y) are samples from the same uniform distribution, which implies many nice mathematical properties.

The statistical correlation coefficient for the rank is known as Spearman's rho [13]:

$$\rho_n^{\text{rank}} = \frac{\sum_{i=1}^n (r_i^X - (n+1)/2)(r_i^Y - (n+1)/2)}{\sqrt{\sum_{i=1}^n (r_i^X - (n+1)/2)^2 \sum_i^n (r_i^Y - (n+1)/2)^2}}. \tag{III.1}$$

The mathematical properties of the Spearman's rho have been extensively investigated. In particular, if $((X_i, Y_i))_{i=1}^n$ consists of independent realizations of (X, Y), and the joint distribution function of X and Y is differentiable, then ρ_n^{rank} is a consistent statistical estimator, and its standard deviation is of the order $1/\sqrt{n}$ independently of the exact form of the underlying distributions, see e.g. [19].

For a graph G of size n, we propose to compute ρ_n^{rank} using (III.1) as follows. We define the random variables X and Y as the degrees on two ends of a random *undirected* edge in a graph (that is, when rank correlations are computed, ij and ji represent the same edge). For each edge, when the observed degrees are a and b, we assign $[X = a, Y = b]$ or $[X = b, Y = a]$ with probability $1/2$. Many values of X and Y will be the same making their rank ambiguous. We resolve this by we adding independent uniformly distributed random variables on $[0, 1]$ to each value of X and Y. In the setting when the realisations (X_i, Y_i) are independent, this way of resolving ties preserves the original value of the Spearman's rho on the population, see e.g. [20]. We refer to [21] for a general treatment of rank correlations for non-continuous distributions.

In the remainder of the paper we will demonstrate that the measure ρ_n^{rank} gives consistent results for different n, and it is able to reveal strong negative degree-degree correlations in large networks.

IV. RANDOM GRAPH DATA

We consider four random graph models to highlight our results.

The configuration model. The *configuration model* was invented by Bollobás in [22], inspired by [23]. It was popularized by Newman, Strogatz and Watts [24], who realized that it is a useful and simple model for real-world networks. In the configurations model a node i has a given number d_i of half-edges, with $\ell_n = \sum_{i \in V} d_i$ assumed to be even. Each half-edge is connected to a randomly chosen other half-edge to form an edge in the graph. We chose $\gamma = 2$, thus, the maximum degree is of the order $n^{1/2}$, which corresponds to the case of uncorrelated random networks, such that the probability that two vertices are directly connected is close to $d_i d_j / \ell_n$ [25, 26]. Although self-loops and multiple edges can occur these become rare as $n \to \infty$, see e.g. [27] or [28]. In simulations, we collapse multiple edges to a single edge, and remove self-loops. This changes the degree distribution slightly, and intuitively should yield negative dependencies. In Figure 1(a) we observe that, on average, ρ_n and ρ_n^{rank} are indeed negative in smaller networks but then they converge to zero showing that the degrees on two ends of a random edge are uncorrelated.

Configuration model with intermediate vertices. In order to construct a strongly disassortative graph, we first generate a configuration model as described above, and then we replace every edge by two edges that meet at a middle vertex. In this model, there are $n + \ell_n/2$ vertices and $2\ell_n$ edges (recall that ij and ji are two different edges). Now, if E, V, and d_i, $i = 1, \dots, n$ denote, respectively, the edge set, the vertex set, and the degrees of the original configuration model, then in the model with intermediate edges the assortativity coefficient is as follows:

$$\rho_n = \frac{2\sum_{i\in V} 2d_i - \frac{1}{2\ell_n}\Big(\sum_{i\in V} d_i^2 + 2\ell_n\Big)^2}{\sum_{i\in V} d_i^3 + 4\ell_n - \frac{1}{2\ell_n}\Big(\sum_{i\in V} d_i^2 + 2\ell_n\Big)^2}.$$

When $\gamma < 3$ we have $\mu_3 = \infty$, and thus $\rho_n \to 0$ as $n \to \infty$. Furthermore, the lower bound ρ_n^- also converges to zero as n grows. It is clear that this particular random graph, of any size, is equally and strongly disassortative, however, ρ_n fails to capture this. In Figure 1(b) it is clearly seen that both ρ_n and ρ_n^- quickly decrease in magnitude as n grows. It is striking that ρ_n^{rank} shows a totally different and very appropriate behavior. Its values remain around -0.75 identifying the strong negative dependencies, and the dispersion across different realizations of the graph decreases as $n \to \infty$.

Preferential attachment model. We consider the basic version of the undirected preferential attachment model (PAM), where each new vertex adds only one edge to the network, connecting to the existing nodes with probability proportional to their degrees [15]. In this case, it is well known that $\gamma = 2$ (see e.g. [29]). Newman [1] noticed the counterintuitive fact that the Preferential Attachment graph has asymptotically neutral mixing, $\rho_n \to 0$ as $n \to \infty$. This phenomenon has been studied in detail by Dorogovtsev et al. [11], and it can be clearly observed in Figure 1(c). The reason for this behavior is not the genuine neutral mixing in the PAM but rather the unnatural dependence of ρ_n on the graph size. Indeed, we see that PAMs of small sizes have $\rho_n < 0$, and then the magnitude of ρ_n decreases with the graph size. Again, Spearman's rho consistently shows that the degrees are negatively dependent. This can be understood by noting that the majority of edges of vertices with high degrees, which are old vertices, come from vertices which are added late in the graph growth process and thus have small degree. On the other hand, by the growth mechanism of the PAM, vertices with low degree are more likely to be connected to vertices having high degree, which indeed suggests negative degree-degree dependencies.

A collection of complete bipartite graphs. We next present an example where the assortativity coefficient has a nonvanishing dispersion. Take $((X_i, Y_i))_{i=1}^n$ to be a sample of independent realizations of the vector (X, Y). We assume that $X = bU_1 + bU_2$ and $Y = bU_1 + aU_2$, where $b > 0$, $a > 1$, and U_1, U_2 are independent identically distributed (i.i.d.) random variables with power law tail, and tail exponent γ. Then, for $i = 1, \dots, n$, we create a complete bipartite graph of X_i and Y_i vertices, respectively. These n complete bipartite graphs are not connected to one another. We denote such a collection of n bipartite graphs by G_n. This is an extreme scenario of a network consisting of highly connected clusters of different size. Such networks can serve as models for physical human contacts and are used in epidemic modelling [9].

The graph G_n has $|V| = \sum_{i=1}^n (X_i + Y_i)$ vertices and $|E| = 2\sum_{i=1}^n X_i Y_i$ edges. Further,

$$\sum_{i\in V} d_i^p = \sum_{i=1}^n (X_i^p Y_i + Y_i^p X_i), \qquad \sum_{ij\in E} d_i d_j = 2\sum_{i=1}^n (X_i Y_i)^2.$$

Assume that $\mathbb{P}(U_j > x) = c_0 x^{-\gamma}$, where $c_0 > 0$, $x \geq x_0$, and $\gamma \in (3, 4)$, so that $\mathbb{E}[U^3] < \infty$, but $\mathbb{E}[U^4] = \infty$. As a result, $|E|/n \xrightarrow{\mathbb{P}} 2\mathbb{E}[XY] < \infty$ and $\frac{1}{n}\sum_{i\in V} d_i^2 \xrightarrow{\mathbb{P}} \mathbb{E}[XY(X+Y)] < \infty$.

Further,

$$n^{-4/\gamma}b^{-4}\sum_{i=1}^{n}(X_i^3Y_i+Y_i^3X_i)\xrightarrow{d}(a^3+a)Z_1+2Z_2,\qquad n^{-4/\gamma}b^{-4}\sum_{i=1}^{N}(X_iY_i)^2\xrightarrow{d}a^2Z_1+Z_2,$$

where Z_1 and Z_2 and two independent stable distributions with parameter $\gamma/4$. As a result,

$$\rho_n\xrightarrow{d}\frac{2a^2Z_1+2Z_2}{(a+a^3)Z_1+2Z_2},\qquad \text{as } n\to\infty,$$

which is a proper random variable taking values in $(2a/(1+a^2),1)$, see [16] for detailed proof.

Note that in this model there is a genuine dependence between the correlation measure and the graph size. Indeed, if $n=1$ then the assortativity coefficient equals -1 because nodes with larger degrees are connected to nodes with smaller degrees. However, when the graph size grows, the positive linear dependence between X and Y starts dominating, thus, larger graphs of this structure are strongly assortative. While the example we present is quite special, we believe that the effect described is rather general.

In Figure 1(d) we again see that ρ_n^{rank} captures the relation faster and gives consistent results with decreasing dispersion. On a contrary, ρ_n has a persistent dispersion in its values, and we know from the result above that this dispersion will not vanish as $n\to\infty$. In the limit, ρ_n has a non-zero density on $(0.8,1)$. However, the convergence is too slow to observe it at $n=100,000$, because the vanishing terms are of the order $n^{-1/\gamma}$, which is only $n^{-1/3.1}$ in our example.

V. WEB SAMPLES AND SOCIAL NETWORKS

We computed ρ_n, ρ_n^{rank} and ρ_n^- on several Web samples (disassortative networks) and social network samples (assortative networks). We used the compressed graph data from the Laboratory of Web Algorithms (LAW) at the Università degli studi di Milano [30, 31]. We used the bvgraph MATLAB package [32]. The *stanford-cs* database [33] is a 2001 crawl that includes all pages in the cs.stanford.edu domain. In datasets (iv), (vii), (viii) we evaluate ρ_n, ρ_n^{rank} and ρ_n^- over 1000 random edges, and present the average over 10 such evaluations (in 10 samples of 1000 edges, the observed dispersion of the results was small).

The results are presented in Table I. We clearly see that the assortativity coefficient ρ_n and Spearman's ρ_n^{rank} always agree about whether dependencies are positive or negative. They also agree in magnitude of correlations when graph size is small or the lower bound ρ_n^- is sufficiently far from zero. However, ρ_n is not consistent for graphs of similar structure but different sizes. This is especially apparent on the two .uk crawls (iii) and (iv). Here ρ_n is significantly smaller in magnitude on a larger crawl. Intuitively, mixing patterns should not depend on the crawl size. This is indeed confirmed by the value of Spearman's rho, which consistently shows strong negative correlations in both crawls. We could not observe a similar phenomenon so sharply in (vi) and (vii), probably because a larger co-authorship network incorporates articles from different areas of science, and the culture of scientific collaborations can vary greatly from one research field to another.

We also notice that, as predicted by our results, the assortativity coefficient tends to take smaller values than ρ_n^{rank} if ρ_n^- is small in magnitude. This is clearly seen in the data sets (ii), (iv) and (v). Again, (ii) and (iv) are the largest among the analyzed web crawls.

The observed behaviour of the assortativity coefficient is explained by the above stated results that ρ_n is influenced greatly by the large dispersion in the degree values. The latter increases with graph size because of the scale-free phenomenon. As a result, ρ_n becomes

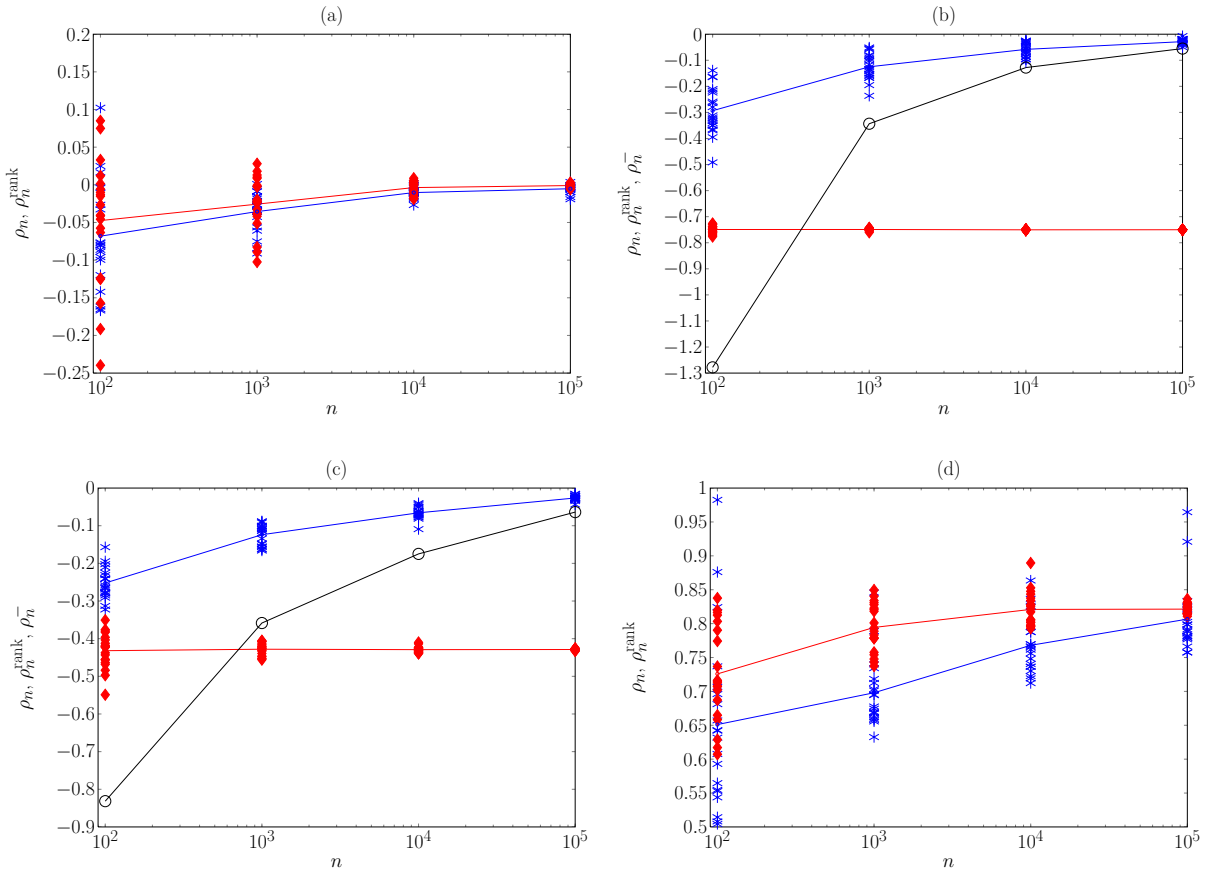


FIG. 1. (Color online) Scatter plots for samples of 20 graphs. For each size we plot the 20 realizations of ρ_n (blue asterisks) and ρ_n^{rank} (red diamonds) in random graphs of different sizes. Solid lines connect the averages of the samples. In (c), (d) the circles connected by the solid line are the averages of ρ_n^- in the samples. (a) Configuration model, $\mathbb{P}(d \geq x) = x^{-2}$, $x \geq 1$. (b) Configuration model with intermediate vertices. (c) Preferential attachment model. (d) A collection of bi-partite graphs, where $b = 1/2$, $a = 2$, and U has a generalized Pareto distribution $\mathbb{P}(U > x) = ((2.1 + x)/3.1)^{-3.1}$, $x > 1$.

nr	Dataset	Description	# nodes	# edges	max degree	ρ_n	ρ_n^{rank}	ρ_n^-
(i)	stanford-cs	web domain	9,914	54,854	340	-.1656	-.1627	-.4648
(ii)	eu-2005	.eu web domain	862,664	5,477,938	68,963	-.0562	-.2525	-.0670
(iii)	uk@100,000	.uk web crawl	100,000	5,559,150	55,252	-.6536	-.5676	-1.117
(iv)	uk@1,000,000	.uk web crawl	1,000,000	77,123,940	403,441	-.0831	-.5620	-.0854
(v)	enron	e-mail exchange	69,244	506,898	1,634	-.1599	-.6827	-.1932
(vi)	dblp-2010	co-authorship	326,186	1,615,400	238	.3018	.2604	-.7736
(vii)	dblp-2011	co-authorship	986,324	6,707,236	979	.0842	.1351	-.2963
(viii)	hollywood-2009	co-starring	1,139,905	113,891,327	11,468	.3446	.4689	-0.6737

TABLE I. (i)–(iv) Web crawls: nodes are web pages, and an (undirected) edge means that there is a hyperlink from one of the two pages to another; (iii),(iv) are breadth-first crawls around one page. (v) e-mail exchange by Enron employees (mostly part of the senior management): node are employees, and an edge means that an e-mail message was sent from one of the two employees to another. (vi), (vii) scientific collaboration networks extracted from the DBLP bibliography service: each vertex represents a scientist and an edge means a co-authorship of at least one article. (viii) vertices are actors, and two actors are connected by an edge if they appeared in the same movie.

smaller in magnitude, which makes it impossible to compare graphs of different sizes. In contrast, the *ranks* of the degrees are drawn from a uniform distribution on $[0, 1]$, scaled by the factor n. Clearly, when a correlation coefficient is computed, the scaling factor cancels, and therefore Spearman's rho provides consistent results in the graphs of different sizes.

VI. DISCUSSION

The assortativity coefficient ρ_n proposed in [1, 2] has been the first dependency measure introduced to describe degree-degree correlations in networks. The assortativity coefficient has provided many interesting insights. It has been successfully used for comparison of dependencies in graphs with the same degree sequences [34, 35], and to generate graphs with given degrees and desired mixing patterns [36]. An important drawback of ρ_n is its dependence on the network size n. It has been noticed by many authors, and shown in this paper for disassortative networks, that ρ_n converges to zero as n grows. In particular, the decay with network size of the assortativity coefficient ρ_n implies that it cannot be used for comparing dependencies in networks of different sizes. Therefore, it prohibits the investigation whether growing networks become more or less assortitative over time.

This paper suggests to use rank-correlation measures such as Spearman's rho. Our experiments convincingly show that Spearman's rho does not suffer from the size-dependence deficiency. In networks of different sizes but similar structure, Spearman's rho yields consistent results, and it is able to reveal strong (positive or negative) correlations in large networks. We conclude that rank correlations are a suitable and informative method for uncovering network mixing patterns.

For the correct interpretation of degree-degree dependencies, it is important to realise that positive or negative correlations can be pre-defined by the degree sequence itself. For instance, there is only one simple graphs with degrees $(3, 1, 1, 1)$, and the result $\rho_4 = -1$ is not informative in this case. It has been discussed in the literature that, conditioned on not having self-loops and multiple edges, random networks with given degrees exhibit disassortative patterns [25, 35, 37], also called *structural* correlations. In order to filter out the structural correlations, one needs to compare the real-world networks to their null-models – graphs with the same degree sequences but *random* connections. This null-model is a uniform simple random graph with the same degree sequence. Here a network is called simple when it has no self-loops nor multiple edges. Such a graph can be obtained by randomly pairing half-edges, as in Section IV, and taking the first realization that is simple. This is especially problematic when $(\max_i d_i)^2 > |E|$, which is the case in many examples, since then one needs a prohibitingly large number of attempts before a simple graph is generated [28, 38].

A widely accepted method for constructing a null-model, is the random rewiring of the connections in a given graph [34, 35]. The disadvantage is the unknown running time before a graph is produced that is close enough to being uniform. Recent work [39] presents a sequential algorithm, where, at each step, the remaining unconnected edges maintain the ability to generate a simple graph. This method always produces the desired outcome but its worst-case running time $O(n^2 \sum_i d_i)$ is infeasible for large networks. The recently introduced grand-canonical model [40] computes the probability of connection between two nodes in a maximum entropy graph with given degree sequence, and enables the evaluation of many characteristics of the graph. To the best of our knowledge, efficient implementation of this method for large networks has not been developed yet.

Constructing a null-model and filtering out the structural correlations in large networks is an interesting and demanding computational task that is beyond the scope of this paper. We believe that structural correlations will affect ρ_n to a larger extent than the rank correlation

ρ_n^{rank} because it is usually the nodes with largest degrees that produce self-loops and multiple edges, and thus the relative contribution of these edges in the cross-products will be larger for ρ_n than for ρ_n^{rank}. This conjecture requires a further investigation.

We conclude by stating that rank correlation measures deserve to become a standard tool in the analysis of complex networks. The use of rank correlation measures has become common ground in the area of statistics for analysing heavy-tailed data. We hope to have provided a sufficient evidence that this method is preferred for analysing network data with heavy-tailed degrees as well.

ACKNOWLEDGMENT

We thank Yana Volkovich for the code generating a Preferential Attachment graph. This article is also the result of joint research in the 3TU Centre of Competence NIRICT (Netherlands Institute for Research on ICT) within the Federation of Three Universities of Technology in The Netherlands. The work of RvdH was supported in part by the Netherlands Organisation for Scientific Research (NWO). The work of NL is partially supported by the EU-FET Open grant NADINE (288956).

---

[1] M. Newman, Physical Review Letters **89**, 208701 (2002).
[2] M. Newman, Physical Review E **67**, 026126 (2003).
[3] J. Doyle, D. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger, PNAS **102**, 14497 (2005).
[4] L. Li, D. Alderson, J. Doyle, and W. Willinger, Internet Mathematics **2**, 431 (2005).
[5] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, ACM SIGCOMM Computer Communication Review **36**, 135 (2006).
[6] R. May, S. Levin, and G. Sugihara, Nature **451**, 893 (2008).
[7] D. Braha and Y. Bar-Yam, Management Science **53**, 1127 (2007).
[8] V. Eguiluz and K. Klemm, Physical Review Letters **89**, 108701 (2002).
[9] S. Eubank, H. Guclu, V. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, Nature **429**, 180 (2004).
[10] S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer, Internet Mathematics **4**, 245 (2007).
[11] S. Dorogovtsev, A. Ferreira, A. Goltsev, and J. Mendes, Physical Review E **81**, 031135 (2010).
[12] M. Raschke, M. Schläpfer, and R. Nibali, Physical Review E **82**, 037102 (2010).
[13] C. Spearman, The American journal of psychology **15**, 72 (1904).
[14] H. Hotelling and M. Pabst, The Annals of Mathematical Statistics **7**, 29 (1936).
[15] R. Albert and A. Barabási, Science **286**, 509 (1999).
[16] N. Litvak and R. van der Hofstad, Arxiv preprint arXiv:1202.3071 (2012), work in progress.
[17] R. Albert and A. Barabási, Reviews of Modern Physics **74**, 47 (2002).
[18] M. Newman, SIAM Review **45**, 167 (2003).
[19] C. Borkowf, Computational statistics & data analysis **39**, 271 (2002).
[20] M. Mesfioui and A. Tajar, Nonparametric Statistics **17**, 541 (2005).
[21] J. Nevslehová, Journal of Multivariate Analysis **98**, 544 (2007).
[22] B. Bollobás, European J. Combin. **1**, 311 (1980).
[23] E. Bender and E. Canfield, Journal of Combinatorial Theory, Series A **24**, 296 (1978).
[24] M. Newman, S. Strogatz, and D. Watts, Physical Review E **64**, 026118 (2001).

[25] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, The European Physical Journal B - Condensed Matter ar 10.1140/epjb/e2004-00038-8.
[26] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, Phys. Rev. E **71**, 027103 (2005).
[27] B. Bollobás, *Random graphs*, Vol. 73 (Cambridge Univ Pr, 2001).
[28] S. Janson, Combinatorics, Probability and Computing **18**, 205 (2009).
[29] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, Random Structures and Algorithms **18**, 279 (2001).
[30] P. Boldi and S. Vigna, in *Proceedings of the 13th International World Wide Web Conference (WWW 2004)* (ACM Press, Manhattan, USA, 2004) pp. 595–601.
[31] P. Boldi, M. Rosa, M. Santini, and S. Vigna, in *Proceedings of the 20th International World Wide Web Conference (WWW 2011)* (ACM Press, 2011).
[32] D. Gleich, A. Gray, C. Greif, and T. Lau, SIAM Journal on Scientific Computing **32**, 349 (2010).
[33] P. Constantine and D. Gleich, in *Proceedings of the 5th Workshop on Algorithms and Models for the Web Graph (WAW2007)*, Lecture Notes in Computer Science, Vol. 4863, edited by A. Bonato and F. C. Graham (Springer, 2007) pp. 82–95.
[34] S. Maslov and K. Sneppen, Science **296**, 910 (2002).
[35] S. Maslov, K. Sneppen, and A. Zaliznyak, Physica A: Statistical Mechanics and its Applications **333**, 529 (2004).
[36] P. Van Mieghem, H. Wang, X. Ge, S. Tang, and F. Kuipers, The European Physical Journal B-Condensed Matter and Complex Systems **76**, 643 (2010).
[37] J. Park and M. Newman, Phys. Rev. E **68**, 026112 (2003).
[38] R. van der Hofstad, "Random graphs and complex networks," (2009).
[39] J. Blitzstein and P. Diaconis, Internet Mathematics **6**, 489 (2011).
[40] T. Squartini and D. Garlaschelli, New Journal of Physics **13**, 083001 (2011).

Degree-degree dependencies in random graphs with heavy-tailed degrees

Remco van der Hofstad*and Nelly Litvak†

August 9, 2013

Abstract

Mixing patterns in large self-organizing networks, such as the Internet, the World Wide Web, social and biological networks are often characterized by degree-degree dependencies between neighbouring nodes. In *assortative* networks, the degree-degree dependencies are positive (nodes with similar degrees tend to connect to each other), while in *disassortative* networks, these dependencies are negative. One of the problems with the commonly used Pearson correlation coefficient, also known as the *assortativity coefficient* is that its magnitude decreases with the network size in disassortative networks. This makes it impossible to compare mixing patterns, for example, in two web crawls of different sizes. As an alternative, we have recently suggested to use rank correlation measures, such as Spearman's rho. Numerical experiments have confirmed that Spearman's rho produces consistent values in graphs of different sizes but similar structure, and it is able to reveal strong (positive or negative) dependencies in large graphs.

In this paper we analytically investigate degree-degree dependencies for scale-free graph sequences. In order to demonstrate the ill behaviour of the Pearson's correlation coefficient, we first study a simple model of two heavy-tailed highly correlated random variables X and Y, and show that the sample correlation coefficient converges in distribution either to a proper random variable on $[-1, 1]$, or to zero, and the limit is non-negative a.s. if $X, Y \geq 0$. We next adapt these results to the degree-degree dependencies in networks as described by the Pearson correlation coefficient, and show that it is non-negative in the large graph limit when the asymptotic degree distribution has an infinite third moment. Furthermore, we provide examples where the Pearson's correlation coefficient converges to zero in a network with strong negative degree-degree dependencies, and another example where this coefficient converges in distribution to a random variable. We suggest the alternative degree-degree dependency measure, based on Spearman's rho, and prove that this statistical estimator converges to an appropriate limit under quite general conditions. These conditions are proved to hold in common network models, such as the configuration model and the preferential attachment model. We conclude that rank correlations provide a suitable and informative method for uncovering network mixing patterns.

Keywords. Dependencies of heavy-tailed random variables, Power-laws, Scale-free graphs, Assortativity, Degree-degree correlations

1 Introduction

In this paper we present an analytical study of degree-degree correlations in graphs with power law degree distribution. In simple words, a random variable X has a power-law distribution with tail exponent $\gamma > 0$ if its tail probability $\mathbb{P}(X > x)$ is roughly proportional to $x^{-\gamma}$, for large enough x. Large self-organizing networks, such as the Internet, the World Wide Web, social and biological networks, usually exhibit high variation in the values of the degrees. Such networks are called *scale free* indicating that there is no typical scale for the degrees, and the high degree vertices are called *hubs*. This phenomenon is often modelled by using power-law degree distributions.

*Eindhoven University of Tedhnology and Eurandom
†University of Twente

Power-law distributions are *heavy tailed* since the tail probability decreases much more slowly than a negative exponential, and thus one observes extremely large values of X much more frequently than in the case of light tails. Statistical analysis of scale-free complex networks has received massive attention in recent literature, see e.g. [33, 40] for excellent surveys. Nevertheless, there still are many fundamental open problems. One of them is how to measure *dependencies* between network parameters.

An important characteristic of networks is the dependency between the degrees of direct neighbours. A network is usually called *assortative* when nodes with similar degrees are often connected, thus, the degree-degree dependencies are positive, while in a *disassortative* network these dependencies are negative. The *degree-degree* dependencies define many of the network's properties. For instance, the negative degree-degree correlations in the Internet graph have a great influence on the robustness to failures [15], efficiency of Internet protocols [29], as well as distances and betweenness [30]. The correlation between in- and out-degree of tasks plays and important role in the dynamics of production and development systems [11]. Mixing patterns affect epidemic spread [17, 18] and Web ranking [19].

Often, degree-degree dependence is characterized by the *assortativity coefficient* of the network, introduced by Newman in [38]. The assortativity coefficient is in fact the Pearson correlation coefficient between the vector of degrees on each side of an edge, as a function of all edges. See [38, Table I] for a list of assortativity coefficients for various real-world networks. The empirical data suggest that social networks tend to be assortative (the assortativity coefficient is positive), while Internet, World Wide Web, and biological networks tend to be disassortative. In [38, Table I], it is striking that, typically, larger disassortative networks have an assortativity coefficient that is closer to 0 and therefore appear to have approximate *uncorrelated* degrees across edges. Similar conclusions can be drawn from [39], see in particular [39, Table II]. This phenomenon arises because Pearson's correlation coefficient in scale-free networks with realistic parameters decreases with the network size, as was pointed out in several recent papers [14, 42, 24]. In this paper, we prove that Pearson's correlation coefficient in scale-free networks shows several types of pathological behavior, in particular, its infinite volume limit, when it exists, is non-negative, independently of the mixing pattern, and in fact this limit can even be *random*.

In [24] we propose an alternative measure for the degree-degree dependencies, based on the *ranks* of degrees. This rank correlation approach is in fact classical in multivariate analysis, falling under the category of 'concordance measures' - dependency measures based on *order* rather than exact values of two stochastic variables. The huge advantage of such dependency measures is that they work well *independently* of the number of finite moments of the degrees, while Pearson's coefficient suffers from a strong dependence on the extreme values of the degrees. Recent applications of rank correlation measures, such as Spearman's rho [44] and the closely related Kendall's tau [27], include the concordance between two rankings for a set of documents in web search. In this application field many other measures for rank distances have been proposed, see e.g. [28] and the references therein.

We show mathematically that statistical estimators for degree-degree dependencies based on rank correlations are *consistent*. That is, for graphs of different sizes but similar structure (e.g. preferential attachment graphs of increasing size), these estimators converge to their 'true' or limiting value that describes the degree-degree dependence in an infinitely large graph (in particular, the variance of the estimator decreases as the size of the graph grows). We also show that Pearson's correlation coefficient does not have this basic property when degree distributions are heavy-tailed. In particular, as explained in more detail in [24], this implies that the assortativity coefficient as suggested in [38] does not allow one to compare the degree-degree dependencies in graphs of different sizes, such as they arise when studying a network at different time stamps, or comparing two different networks, e.g. web crawls of different domains or Wikipedia graphs from different languages. On the other hand, such a comparison *is* possible using Spearman's rho. This paper forms the mathematical justification of our paper [24], where similar results were predicted on a less formal level and confirmed by numerical experiments.

The paper is organized as follows. In Section 2 we start with the analysis of the sample Pearson correlation coefficient and the sample rank correlation, Spearman's rho, for a two-dimensional vector with heavy-tailed marginals. In Section 2.3 we present a simple model with an explicit linear dependence and show that, when the sample size grows to infinity, then Pearson's correlation coefficient does not converge to a constant but rather to a random variable involving stable distributions. We also verify analytically and numerically that the rank correlation provides a consistent statistical estimator for this model. Next, in Section 2.4 we prove that if random variables are heavy-tailed with infinite second moment and non-negative, then the sample Pearson correlation coefficient never converges to a negative value. Thus, such sequence will never be classified as 'disassortative'. This result is extended to sequences of graphs in Section 3, where we also obtain quite general convergence criteria in the infinite volume limit for the Pearson's correlation coefficient and the Spearman's rho. In Section 4 analytical results are provided for Pearson's correlation coefficient and rank correlations in the configuration model and the Preferential Attachment model. We also present an adaptation of the configuration model that has strong negative degree-degree dependencies and prove that Spearman's rho converges to the theoretically justified negative value while Pearson's coefficient converges to zero. Furthermore, we construct an example, where Pearson's correlation coefficient converges to a random variable. Numerical results are presented in Section 5. We close the paper in Section 6 with a discussion on our results and possible extensions thereof.

2 Correlations between random variables

In this section we introduce the dependency measures studied in this paper. We start with a general description of dependency measures for random vectors (X, Y). This will provide the necessary intuition and framework in order to understand what happens when X and Y are the degrees of neighboring nodes in a network graph. We present Pearson's sample correlation coefficient in Section 2.1, and introduce Spearman's rho in Section 2.2. In Section 2.3 we demonstrate an ill behaviour of Pearson's sample coefficient in a simple model with linear dependencies, and in Section 2.4 we show that if X and Y are non-negative then the Pearson's sample coefficient cannot converge to a negative value.

2.1 Sample Pearson's correlation coefficient

The Pearson correlation coefficient ρ for two random variables X and Y with cumulative distribution functions $F_X(\cdot)$ and $F_X(\cdot)$, joint cumulative distribution function $F_{X,Y}(\cdot,\cdot)$, and $\mathrm{Var}(X), \mathrm{Var}(Y) < \infty$ is defined by

$$\rho = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}. \tag{2.1}$$

By Cauchy-Schwarz, $\rho \in [-1, 1]$, and ρ measures the *linear dependence* between the random variables X and Y. We can approximate ρ from a sample by computing the *sample correlation coefficient*

$$\rho_n = \frac{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{S_n(X)S_n(Y)}, \tag{2.2}$$

where

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \qquad \bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i \tag{2.3}$$

denote the *sample averages* of $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$, while

$$S_n^2(X) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2, \qquad S_n^2(Y) = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 \tag{2.4}$$

denote the *sample variances.* For i.i.d. sequences of random vectors $((X_i, Y_i))_{i=1}^n$ under the assumption of finite-variance random variables, i.e., $\mathrm{Var}(X), \mathrm{Var}(Y) < \infty$, it is well known that the estimator ρ_n of ρ is *consistent,* i.e.,

$$\rho_n \xrightarrow{\mathbb{P}} \rho, \tag{2.5}$$

where $\xrightarrow{\mathbb{P}}$ denotes convergence in probability. In practice, however, we tend not to know whether $\mathrm{Var}(X), \mathrm{Var}(Y) < \infty$, since $S_n^2(X) < \infty$ and $S_n^2(Y) < \infty$ clearly hold for any sample, and, therefore, one might be tempted to always use ρ_n. Furthermore, by the Cauchy-Schwarz inequality, $\rho_n \in [-1, 1]$ for every $n \geq 1$, which is part of the problem, because, for any sample, a value in $[-1, 1]$ is produced, and no alarm bells start rinkling when ρ_n is used inappropriately. In this paper we investigate the case $\mathrm{Var}(X), \mathrm{Var}(Y) = \infty$, and show that the use of ρ_n in this case, and in particular in scale-free random graphs, is uninformative. For example, in case of negative correlations ρ_n converges to zero when $n \to \infty$, which makes it impossible to compare the data of different sizes. Moreover, if correlations are positive, ρ_n may even converge to a random variable, thus it can produce very different numbers for two random structures of the same size created by the same mechanism. We provide such examples for linearly dependent random variables in Section 2.3 and for random graphs in Section 4.4.

2.2 Rank correlations

For two-dimensional data $((X_i, Y_i))_{i=1}^n$, let r_i^X and r_i^Y be the rank of an observation X_i and Y_i, respectively, when the sample values $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$ are arranged in a descending order. The idea of rank correlations is in evaluating statistical dependences on the data $((r_i^X, r_i^Y))_{i=1}^n$, rather than on the original data $((X_i, Y_i))_{i=1}^n$. Rank transformation is convenient, in particular because, for continuous random variables, the two marginals of the resulting vector (r_i^X, r_i^Y) are realizations of identical uniform distributions, implying many nice mathematical properties.

The statistical correlation coefficient for the ranks is known as Spearman's rho [44]:

$$\rho_n^{\mathrm{rank}} = \frac{\sum_{i=1}^n (r_i^X - (n+1)/2)(r_i^Y - (n+1)/2)}{\sqrt{\sum_{i=1}^n (r_i^X - (n+1)/2)^2 \sum_i^n (r_i^Y - (n+1)/2)^2}} = \frac{\frac{1}{n}\sum_{i=1}^n r_i^X r_i^Y - ((n+1)/2)^2)}{\frac{1}{12}(n^2-1)}. \tag{2.6}$$

The mathematical properties of Spearman's rho have been extensively investigated in the literature. It is well known that if $((X_i, Y_i))_{i=1}^n$ consists of independent realizations of (X, Y), and the joint distribution cumulative function of X and Y is continuous, then ρ_n^{rank} converges to a number that can be interpreted as its population value, see [26, Chapter 9], [10]:

$$\rho_n^{\mathrm{rank}} \xrightarrow{\mathbb{P}} \rho^{\mathrm{rank}} = 12\mathbb{E}(F_X(X)F_Y(Y)) - 3. \tag{2.7}$$

For completeness, we give a brief explanation of this formula. Observe that $F_X(X)$ is the random variable that takes the value $F_X(x)$ when $X = x$. If X is continuous then $F_X(X)$ has a uniform distribution on $[0, 1]$:

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(F_X(X) \leq F_X(x)). \tag{2.8}$$

Now take $F_X(x) = t$ to obtain $\mathbb{P}(F_X(X) \leq t) = t$, where t can take any value in $[0, 1]$. We note that this derivation holds for *any* continuous random variable X. We will use this many times throughout the paper. In particular, it follows that $\mathbb{E}(F_X(X)) = \mathbb{E}(F_Y(Y)) = 1/2$. Next, note that r_i^X/n is an empirical estimator of $1 - F_X(x_i)$, where x_i is the realized value of X_i. Moreover,

$$\mathbb{E}((1 - F_X(X))(1 - F_Y(Y))) = 1 - \mathbb{E}(F_X(X)) - \mathbb{E}(F_Y(Y)) + \mathbb{E}(F_X(X)F_Y(Y)) = \mathbb{E}(F_X(X)F_Y(Y)).$$

Hence, the right-hand side of (2.6) is a statistical estimator of the last expression in (2.7).

For discrete random variables, the situation is more delicate, as the same values of X and Y may occur more than once. We resolve the ties randomly, using uniformisation as suggested in [31]. Formally, we replace the ranks of $((X_i, Y_i)_{i=1}^n$ by the ranks of the random variables

$$((X_i^*, Y_i^*))_{i=1}^n = ((X_i + U_i, Y_i + U_i'))_{i=1}^n,$$

where $((U_i, U_i'))_{i=1}^n$ is a sequence of $2n$ i.i.d. uniform variables on $(0,1)$. The random variables X_i^* and Y_i^* now *are* continuous. We denote their cumulative distribution functions by F_X^* and F_Y^*. Note that if X takes non-negative integer values then F_X^* can be seen as a linear interpolation of the cumulative probability $\mathbb{P}(X < x)$, $x = 0, 1, 2, \ldots$ because $\mathbb{P}(X = x) = \mathbb{P}(X^* \in [x, x+1))$.

Since (X^*, Y^*) has a continuous distribution, the convergence result in (2.7) remains valid. Moreover, [31, Proposition 3.1] states that the population value ρ^{rank} is the same for (X, Y) en (X^*, Y^*):

$$\mathbb{E}(F_X^*(X^*)F_Y^*(Y^*)) = \mathbb{E}(F_X(X)F_Y(Y)). \tag{2.9}$$

The comparison of different ways for resolving ties, and their effect on the resulting computation is an interesting topic, which is outside the scope of this work. We refer to [36] for a general treatment of rank correlations for non-continuous distributions.

2.3 Linear dependencies

It is well known that ρ in general measures *linear* dependence between two random variables. Therefore, before analyzing the behavior of ρ_n in networks, we wish to illustrate that ρ_n fails to capture the linear dependence between X and Y when the variances of X and Y are infinite, i.e., $\text{Var}(X), \text{Var}(Y) = \infty$, even in a very straightforward case when the linear relation between X and Y is explicitly defined. With this goal in mind, below we analyze the behavior of ρ_n in the following linear model:

$$X \quad = \alpha_1\xi_1 + \cdots + \alpha_m\xi_m, \qquad Y = \beta_1\xi_1 + \cdots + \beta_m\xi_m, \tag{2.10}$$

where ξ_j, $j = 1, \ldots, m$, are independent identically distributed (i.i.d.) non-negative random variables with regularly varying tail, and tail exponent γ. By definition, the non-negative random variable ξ is *regularly varying* with index $\gamma > 0$, if

$$\mathbb{P}(\xi > x) = L(x)x^{-\gamma}, \qquad x \geq 0, \tag{2.11}$$

where $x \mapsto L(x)$ is a slowly varying function, that is, for $u > 0$, $L(ux)/L(x) \to 1$ as $x \to \infty$, for instance, $L(x)$ may be equal to a constant or $\log(x)$. Note that the random variables X and Y have the same distribution when $(\beta_1, \ldots, \beta_m)$ is a permutation of $(\alpha_1, \ldots, \alpha_m)$.

When we take an i.i.d. sample of random variables $((X_i, Y_i))_{i=1}^n$ of random variables with the above linear dependence, then Spearman's rho is consistent by (2.7), with a variance that converges to zero as $1/n$. For the sample correlation coefficient, consistency follows from (2.5) in the case where $\text{Var}(\xi_i) < \infty$, but not when the ξ_i's have infinite variance as we show below in detail. Our main result in this section is the following theorem:

Theorem 2.1 (Weak convergence of the sample Pearson's coefficient)**.** *Let $((X_i, Y_i))_{i=1}^n$ be i.i.d. copies of the random variables (X, Y) in* (2.10)*, and where $(\xi_j)_{j=1}^m$ are i.i.d. random variables satisfying* (2.11) *with $\gamma \in (0, 2)$, so that $\text{Var}(\xi_j) = \infty$. Then,*

$$\rho_n \xrightarrow{d} \rho \equiv \frac{\sum_{j=1}^m \alpha_j\beta_j Z_j}{\sqrt{\sum_{j=1}^m \alpha_j^2 Z_j}\sqrt{\sum_{j=1}^m \beta_j^2 Z_j}}, \tag{2.12}$$

where $(Z_j)_{j=1}^m$ are i.i.d. random variables having stable distributions with parameter $\gamma/2 \in (0, 1)$, and $\xrightarrow{d}$ denotes convergence in distribution. In particular, ρ has a density on $[-1, 1]$. This density is strictly positive on $(-1, 1)$ when there exist k, l such that $\alpha_k\beta_k < 0 < \alpha_l\beta_l$. Furthermore, the density is positive on $(a, 1)$ when $\alpha_k\beta_k \geq 0$ for every k, and on $(-1, -a)$ when $\alpha_k\beta_k \leq 0$ for every k, where

$$a = \inf_{z_1,\ldots,z_m \in \mathbb{R}} \frac{\sum_{j=1}^m |\alpha_j\beta_j| z_j}{\sqrt{\sum_{j=1}^m \alpha_j^2 z_j}\sqrt{\sum_{j=1}^m \beta_j^2 z_j}} \in (0, 1). \tag{2.13}$$

Theorem 2.1 states that the sample correlation coefficient converges in distribution to a proper random variable, contrary to Spearman's rank correlation which converges in probability to a constant. In particular, this implies that when we have two independent samples, the sample correlation coefficient will give two rather distinct values, while Spearman's rank correlation will give two similar values. We prove Theorem 2.1 in the remainder of this section. In its proof, we need the following technical result:

Lemma 2.2 (Asymptotics of sums in stable domain)**.** *Let $(\xi_{i,j})_{i=1,2,\dots,n,j=1,2}$ be i.i.d. random variables satisfying* (2.11) *for some $\gamma \in (0,2)$. Then there exists a sequence a_n with $a_n = n^{2/\gamma}\ell(n)$, where $n \mapsto \ell(n)$ is slowly varying, such that*

$$\frac{1}{a_n}\sum_{i=1}^{n} \xi_{i,1}^2 \xrightarrow{d} Z_1, \qquad \frac{1}{a_n}\sum_{i=1}^{n} \xi_{i,1}\xi_{i,2} \xrightarrow{\mathbb{P}} 0, \tag{2.14}$$

where Z_1 is stable with parameter $\gamma/2$ and $\xrightarrow{\mathbb{P}}$ denotes convergence in probability.

Proof. Let $F(x) = \mathbb{P}(\xi \le x)$ be the cumulative distribution function of ξ. In order to prove the first statement in (2.14) we only need to note that the cumulative distribution function of ξ^2 equals $x \mapsto F(\sqrt{x})$, which, by (2.11), implies that ξ^2 is regularly varying. Thus, the first statement in (2.14) is in fact the classical convergence of infinite variance random variables with slowly varying distribution functions to stable laws (see e.g. [21]), where Z_1 is a stable $\gamma/2$ random variable. In particular, denoting $[1-F](x) = 1 - F(x)$, $x \ge 0$, we can identify $a_n = [1-F]^{-1}(1/n^2)$ [4]. Since $x \mapsto [1-F](x)$ is regularly varying with index γ, $[1-F]^{-1}(1/n)$ is regularly varying with index $1/\gamma$ [4], so that $a_n = [1-F]^{-1}(1/n^2)$ is regularly varying with index $2/\gamma$. To prove the second part of (2.14), we write

$$1 - F(x) = \mathbb{P}(\xi > x) \le c' x^{-\gamma'}, \qquad x \ge 0, \tag{2.15}$$

which is valid for any $\gamma' \in (1,\gamma)$ by (2.11) and Potter's theorem. We next study the cumulative distribution function of $\xi_1\xi_2$ which we denote by H, where ξ_1 and ξ_2 are two independent copies of the random variable ξ. When F satisfies (2.15), then it is not hard to see that there exists a $C > 0$ such that

$$1 - H(u) \le C(1 + \log u)u^{-\gamma'}. \tag{2.16}$$

Indeed, assume that F has a density $f(w) = cw^{-(\gamma'+1)}$, for $w \ge 1$. Then,

$$1 - H(u) = \int_1^\infty f(w)[1-F](u/w)dw.$$

Clearly, $1 - F(w) = c'w^{-\gamma'}$ for $w \ge 1$ and $1 - F(w) = 1$ otherwise. Substitution of this yields

$$1 - H(u) \le cc' \int_1^u w^{-(\gamma'+1)}(u/w)^{-\gamma'}dw + c\int_u^\infty w^{-(\gamma'+1)}\,dw \le C(1+\log u)u^{-\gamma'}.$$

When F satisfies (2.15), then ξ_1 and ξ_2 are stochastically upper bounded by $\hat{\xi}_1$ and $\hat{\xi}_2$ with cumulative distribution function $\hat{F}$ satisfying $1 - \hat{F}(w) = c'w^{-\gamma'} \vee 1$, where $(x \vee y) = \max\{x, y\}$, and the claim in (2.16) follows from the above computation.

By the bound in (2.16), the random variables $\xi_{i,1}\xi_{i,2}$ are stochastically bounded from above by random variables P_i that are in the domain of attraction of a stable γ' random variable. As a result, there exists $b_n = n^{1/\gamma'}\ell'(n)$, where $n \mapsto \ell'(n)$ is slowly varying, such that

$$\frac{1}{b_n}\sum_{i=1}^{n} P_i \xrightarrow{d} W,$$

where W is stable γ'. By choosing $\gamma' > \gamma/2$, we get $b_n/a_n \to 0$, so we obtain the second statement in (2.14). □

Proof of Theorem 2.1. We start by noting that

$$\rho_n = \frac{\frac{1}{n-1}\sum_{i=1}^n (X_iY_i - \bar{X}_n\bar{Y}_n)}{S_n(X)S_n(Y)}, \tag{2.17}$$

and

$$S_n^2(X) = \frac{1}{n-1}\sum_{i=1}^n (X_i^2 - \bar{X}_n^2), \qquad S_n^2(Y) = \frac{1}{n-1}\sum_{i=1}^n (Y_i^2 - \bar{Y}_n^2). \tag{2.18}$$

We continue to identify the asymptotic behavior of

$$\sum_{i=1}^n X_i^2, \qquad \sum_{i=1}^n Y_i^2, \qquad \sum_{i=1}^n X_iY_i.$$

Let $[n]$ denote the set of integers $\{1, 2, \ldots, n\}$. The distribution of $((X_i, Y_i))_{i=1}^n$ is described in terms of an array $(\xi_{i,j})_{i\in[n],j\in[m]}$, which are i.i.d. copies of a random variable ξ. In terms of these random variables, we can identify

$$\sum_{i=1}^n X_iY_i = \sum_{j=1}^m \alpha_j\beta_j\Big(\sum_{i=1}^n \xi_{i,j}^2\Big) + \sum_{j_1\neq j_2=1}^m \alpha_{j_1}\beta_{j_2}\Big(\sum_{i=1}^n \xi_{i,j_1}\xi_{i,j_2}\Big). \tag{2.19}$$

The sums $\sum_{i=1}^n \xi_{i,j}^2$ are i.i.d. for different $j \in \{1, \ldots, m\}$, and by Lemma 2.2, $\sum_{i=1}^n \xi_{i,j_1}\xi_{i,j_2}$ is of a smaller order. Hence, from (2.19) we obtain that

$$\frac{1}{a_n}\sum_{i=1}^n X_iY_i \xrightarrow{d} \sum_{j=1}^m \alpha_j\beta_j Z_j. \tag{2.20}$$

Therefore, by taking $\alpha = \beta$, we also obtain

$$\frac{1}{a_n}\sum_{i=1}^n X_i^2 \xrightarrow{d} \sum_{j=1}^m \alpha_j^2 Z_j, \qquad \frac{1}{a_n}\sum_{i=1}^n Y_i^2 \xrightarrow{d} \sum_{j=1}^m \beta_j^2 Z_j, \tag{2.21}$$

and the convergence holds *simultaneously*. As a result, (2.12) follows. It remains to establish the properties of the limiting random variable ρ in (2.12).

The density of Z_i is strictly positive on $(0, \infty)$. Note that rescaling $z_j = cz_j$ $j = 1, \ldots, m$, in (2.13), does not change the value of a. In particular, we can choose $c = (\max\{z_1, z_2, \ldots, z_m\})^{-1}$. If there exist k and l such that $\alpha_k\beta_k < 0 < \alpha_l\beta_l$ then the density of ρ is strictly positive on $(-1, 1)$. Indeed, with positive probability ρ can be arbitrarily close to -1 if $Z_k = \max\{Z_1, \ldots, Z_m\}$ and Z_j/Z_k, $j \neq k$ are sufficiently small. Similarly, if $Z_l = \max\{Z_1, \ldots, Z_m\}$ then with positive probability, ρ can be arbitrarily close to 1. Now assume that $\alpha_k\beta_k \geq 0$ for every k. In this case, the density of ρ is strictly positive on the support of ρ, which is $(a, 1)$, with a as in (2.13). Analogously, when $\alpha_k\beta_k \leq 0$ then ρ cannot be positive, and has a density on $(-1, -a)$. □

Numerical example. In order to illustrate the result of Theorem 2.1, consider the example with ξ_j's from a Pareto distribution satisfying $\mathbb{P}(\xi > x) = 1/x^{1.1}$, $x \geq 1$, so $L(x) = 1$ and $\gamma = 1.1$ in (2.11). The exponent $\gamma = 1.1$ is as observed for the World Wide Web [12]. In (2.10), we choose $m = 3$ and α_i, β_i, $i = 1, 2, 3$, as specified in Table 1. We generate N data samples $((X_i, Y_i))_{i=1}^n$ and compute ρ_n and ρ_n^{rank} for each of the N samples. Thus, we obtain the vectors $(\rho_{n,j})_{j=1}^N$ and $(\rho_{n,j}^{\text{rank}})_{j=1}^N$ of N independent realizations for ρ_n and ρ_n^{rank}, respectively, where the sub-index $j = 1, \ldots, N$ denotes the jth realization of $((X_i, Y_i))_{i=1}^n$. We then compute

$$\mathbb{E}_N(\rho_n) = \frac{1}{N}\sum_{j=1}^N \rho_{n,j}, \quad \mathbb{E}_N(\rho_n^{\text{rank}}) = \frac{1}{N}\sum_{j=1}^N \rho_{n,j}^{\text{rank}}; \tag{2.22}$$

$$\sigma_N(\rho_n) = \sqrt{\frac{1}{N-1}\sum_{j=1}^N (\rho_{n,j} - \mathbb{E}_N(\rho_n))^2}, \quad \sigma_N(\rho_n^{\text{rank}}) = \sqrt{\frac{1}{N-1}\sum_{j=1}^N (\rho_{n,j}^{\text{rank}} - \mathbb{E}_N(\rho_n^{\text{rank}}))^2}. \tag{2.23}$$

The results are presented in Table 1. We clearly see that ρ_n has a significant standard deviation, of which estimators are similar for different values of n. This means that in the limit as $n \to \infty$, ρ_n is a random variable with a significant spread in its values, as stated in Theorem 2.1. Thus, by evaluating ρ_n for one sample $((X_i, Y_i))_{i=1}^n$ we will obtain a random number, even when n is huge. The convergence to a non-trivial distribution is directly seen in Figure 1 because the plots for the two values of n almost coincide. Note that in all cases, the density is fairly uniform, ensuring a comparable probability for all feasible values and rendering the value obtained in a specific realization even more uninformative.

Model parameters	N	10^3	10^2		
	n	10^2	10^3	10^4	10^5
$\alpha = (1/2, 1/2, 0)$ $\beta = (0, 1/2, 1/2)$	$\mathbb{E}_N(\rho_n)$	0.4395	0.4365	0.4458	0.4067
	$\sigma_N(\rho_n)$	0.3399	0.3143	0.3175	0.3106
	$\mathbb{E}_N(\rho_n^{\text{rank}})$	0.4508	0.4485	0.4504	0.4519
	$\sigma_N(\rho_n^{\text{rank}})$	0.0922	0.0293	0.0091	0.0033
$\alpha = (1/2, 1/3, 1/6)$ $\beta = (1/6, 1/3, 1/2)$	$\mathbb{E}_N(\rho_n)$	0.8251	0.7986	0.8289	0.8070
	$\sigma_N(\rho_n)$	0.1151	0.1125	0.1108	0.1130
	$\mathbb{E}_N(\rho_n^{\text{rank}})$	0.8800	0.8850	0.8858	0.8856
	$\sigma_N(\rho_n^{\text{rank}})$	0.0248	0.0073	0.0023	0.0007
$\alpha = (1/2, -1/3, 1/6)$ $\beta = (1/6, 1/2, -1/3)$	$\mathbb{E}_N(\rho_n)$	-0.3052	-0.3386	-0.3670	-0.3203
	$\sigma_N(\rho_n)$	0.6087	0.5841	0.5592	0.5785
	$\mathbb{E}_N(\rho_n^{\text{rank}})$	-0.3448	-0.3513	-0.3503	-0.3517
	$\sigma_N(\rho_n^{\text{rank}})$	0.1202	0.0393	0.0120	0.0034

Table 1: Estimated mean and standard deviation of ρ_n and ρ_n^{rank} in N samples with linear dependence (2.10), $\mathbb{P}(\xi > x) = x^{-1.1}$, $x \geq 1$.

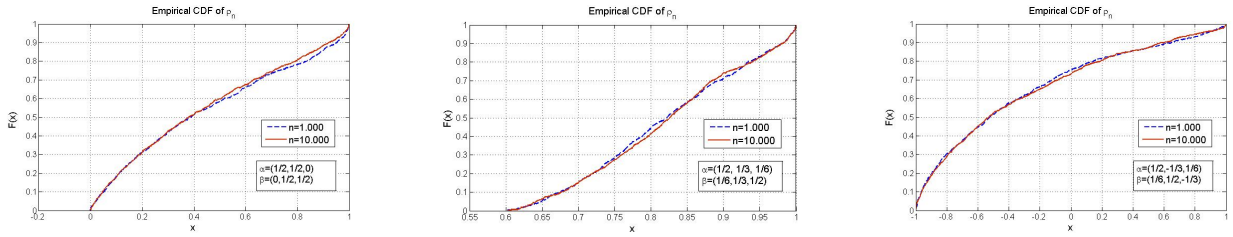


Figure 1: The empirical distribution function $F_N(x) = \mathbb{P}(\rho_n \leq x)$ for the $N = 1.000$ observed values of ρ_n ($n = 1.000$, $n = 10.000$), in the case of linear dependence (2.10).

On the other hand, from Table 1 we clearly see that the behaviour of the rank correlation is exactly as we can expect from a good statistical estimator. The obtained average values are consistent while the standard deviation of ρ_n^{rank} decreases approximately as $1/\sqrt{n}$ as n grows large. Therefore, ρ_n^{rank} converges to a deterministic number.

2.4 Sample Pearson's correlation coefficient for non-negative variables

We proceed by investigating correlations between non-negative heavy-tailed random variables. Our main result in this section shows that the correlation coefficient is asymptotically non-negative:

Theorem 2.3 (Asymptotic non-negativity of the sample Pearson's coefficient for positive r.v.'s)**.** *Let $((X_i, Y_i))_{i=1}^n$ be i.i.d. copies of non-negative random variables (X, Y), where X and Y satisfy*

$$\mathbb{P}(X > x) = L_X(x)x^{-\gamma_X}, \qquad \mathbb{P}(Y > y) = L_Y(y)y^{-\gamma_Y}, \qquad x, y \geq 0, \tag{2.24}$$

with $\gamma_X, \gamma_Y \in (0, 2)$, so that $\text{Var}(X) = \text{Var}(Y) = \infty$. Then, any limit point of the sample Pearson correlation coefficient is non-negative.

N	10^3		10^2		
n	10	10^2	10^3	10^4	10^5
$\mathbb{E}_N(\rho_n)$	-0.4833	-0.1363	-0.0342	-0.0077	-0.0015
$\sigma_N(\rho_n)$	0.1762	0.0821	0.0245	0.0064	0.0011
$\mathbb{E}_N(\rho_n^{\text{rank}})$	-0.6814	-0.4508	-0.4485	-0.4504	-0.4519
$\sigma_N(\rho_n^{\text{rank}})$	0.1580	0.0283	0.0082	0.0024	0.0007

Table 2: The mean and standard deviation of ρ_n and ρ_n^{rank} in N simulations of $((X_i, Y_i))_{i=1}^n$, where $X = 2\xi I$, $Y = 2\xi(1-I)$, I is a Bernoulli(1/2) random variable, $\mathbb{P}(\xi > x) = x^{-1.1}$, $x \geq 1$.

We illustrate Theorem 2.3 with a useful example. Let $(\xi_i)_{i=1}^n$ be a sequence of i.i.d. random variables satisfying (2.11) for some $\gamma \in (0,2)$, and where $\xi \geq 0$ a.s. Let $(X,Y) = (0, 2\xi)$ with probability 1/2 and $(X,Y) = (2\xi, 0)$ with probability 1/2. Then, $XY = 0$ a.s., while $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[\xi]$ and $\text{Var}(X) = \text{Var}(Y) = 2\mathbb{E}[\xi^2] - \mathbb{E}[\xi]^2 = 2\text{Var}(\xi) + \mathbb{E}[\xi]^2$. By Theorem 2.3, $\rho_n \xrightarrow{\mathbb{P}} 0$ when $(\xi_i)_{i=1}^n$ is a sequence of i.i.d. non-negative random variables satisfying (2.11) for some $\gamma \in (0,2)$, which is not appropriate as (X,Y) are highly negatively dependent. When $\gamma > 2$, this anomaly does not arise, since, if $\text{Var}(\xi) < \infty$,

$$\rho_n \xrightarrow{\mathbb{P}} \rho = -\frac{\mathbb{E}[\xi]^2}{2\text{Var}(\xi) + \mathbb{E}[\xi]^2} \in (-1, 0). \tag{2.25}$$

The asymptotics in (2.25) are quite reasonable, since the random variables (X,Y) are highly negatively dependent: When $X > 0$, Y must be equal to 0, and vice versa.

Table 2 shows the empirical mean and standard deviation of the estimators ρ_n and ρ_n^{rank}. Here $\mathbb{P}(\xi > x) = x^{-1.1}$, $x \geq 1$, as in Table 1. As predicted by Theorem 2.3, the sample correlation coefficient (assortativity) converges to zero as n grows large, while ρ_n^{rank} consistently shows a clear negative dependence, and the precision of the estimator improves as $n \to \infty$. This explains why strong disassortativity is not observed in large samples of non-negative power-law data.

We next prove Theorem 2.3:

Proof of Theorem 2.3. Clearly $\sum_{i=1}^n X_i Y_i \geq 0$ when $X_i \geq 0, Y_i \geq 0$, so that

$$\rho_n \geq -\frac{\frac{1}{n-1}\sum_{i=1}^n \bar{X}_n \bar{Y}_n}{S_n(X) S_n(Y)} = -\frac{n}{n-1}\frac{\bar{X}_n}{S_n(X)}\frac{\bar{Y}_n}{S_n(Y)}.$$

It remains to show that if $\text{Var}(X) = \infty$, then $\bar{X}_n / S_n(X) \xrightarrow{\mathbb{P}} 0$. Indeed, if $\gamma \in (1,2)$ then $\bar{X}_n \xrightarrow{\mathbb{P}} \mathbb{E}[X] < \infty$ by the strong law of large numbers. When $\gamma \in (0,1]$, instead, then X is in the domain of attraction of a γ stable random variable, hence $\bar{X}_n$, loosely speaking, it scales as $n^{1/\gamma_X - 1}$. Further, from (2.24) and Lemma 2.2 it follows that $S_n(X)$ scales as $n^{2/\gamma_X - 1}$, in particular, $\bar{X}_n / S_n(X) \xrightarrow{\mathbb{P}} 0$ for all $\gamma \in (0,2)$. □

3 Applications to networks

In real-world networks it is particularly important to measure degree-degree dependencies for neighboring vertices. We refer to [37] for an extensive introduction to networks, their empirical properties and models for them. In Section 3.1 below, we start with the formal definition of Pearson's correlation coefficient (which was termed the *assortativity coefficient* in [38]), and Spearman's rho in the network context. Next, in Section 3.2 we show that all limit points of Pearson's coefficients for sequences of growing scale-free random graphs with power-law exponent $\gamma < 3$ are non-negative, a result that is similar in spirit to Theorem 2.3. In Section 3.3, we state general convergence conditions for both Pearson's correlation coefficient as well as Spearman's rho.

3.1 Definitions and notations

We start by introducing some notation. Let $G = (V, E)$ be an undirected random graph. For a *directed* edge $e = (u, v)$, we write $\underline{e} = u, \overline{e} = v$ and we denote the set of directed edges in E by E' (so that $|E'| = 2|E|$), and D_v is the degree of vertex $v \in V$. In general, D_v is a random variable.

The assortativity coefficient of G is equal to (see, e.g., [38, (4)])

$$\rho(G) = \frac{\frac{1}{|E'|}\sum_{(u,v)\in E'} D_u D_v - \Big(\frac{1}{|E'|}\sum_{(u,v)\in E'} \frac{1}{2}(D_u + D_v)\Big)^2}{\frac{1}{|E'|}\sum_{(u,v)\in E'} \frac{1}{2}(D_u^2 + D_v^2) - \Big(\frac{1}{|E'|}\sum_{(u,v)\in E'} \frac{1}{2}(D_u + D_v)\Big)^2}. \tag{3.1}$$

Note that the assortativity coefficient in (3.1) is equal to the sample correlation coefficient, where $((D_u, D_v))_{(u,v)\in E'}$ represent a sequence of non-negative random variables, as studied in Theorem 2.3. However, $((D_u, D_v))_{(u,v)\in E'}$ are not independent, so that we may not immediately apply the previous theory. Theorem 3.1 below is the analogue of Theorem 2.3 in the network context, and we give a formal proof of it below.

Let us now introduce Spearman's rho in G that we denote by $\rho^{\mathrm{rank}}(G)$. In accordance to the original definition of Spearman's rho, $\rho^{\mathrm{rank}}(G)$ is the correlation coefficient of the sequence of random variables $(R_{\underline{e}}, R_{\overline{e}})$, where e is a uniformly chosen directed edge (u, v) from E'_n. We let $R_{\underline{e}}$ and $R_{\overline{e}}$ be the *rank* of respectively $D_{\underline{e}} + U_e$ and $D_{\overline{e}} + U'_e$ in the sequences $(D_{\underline{e}} + U_e)_{e\in E'_n}$ and $(D_{\overline{e}} + U'_e)_{e\in E'_n}$. Here, as discussed on page 4, $(U_e)_{e\in E'_n}$ and $(U'_e)_{e\in E'_n}$ are i.i.d. sequences of uniform $(0, 1)$ random variables. Then, Spearman's rank correlation coefficient is defined as follows:

$$\rho^{\mathrm{rank}}(G) = \frac{\frac{1}{|E'|}\sum_{e\in E'} R_{\underline{e}} R_{\overline{e}} - (|E'| + 1)^2/4}{(|E'|^2 - 1)/12}. \tag{3.2}$$

3.2 No disassortative scale-free random graph sequences

We compute that

$$\frac{1}{|E'|}\sum_{(u,v)\in E'} \frac{1}{2}(D_u + D_v) = \frac{1}{|E'|}\sum_{v\in V} D_v^2, \qquad \frac{1}{|E'|}\sum_{(u,v)\in E'} \frac{1}{2}(D_u^2 + D_v^2) = \frac{1}{|E'|}\sum_{v\in V} D_v^3. \tag{3.3}$$

Thus, $\rho(G)$ can be written as

$$\rho(G) = \frac{\sum_{(u,v)\in E'} D_u D_v - \frac{1}{|E'|}\Big(\sum_{v\in V} D_v^2\Big)^2}{\sum_{v\in V} D_v^3 - \frac{1}{|E'|}\Big(\sum_{v\in V} D_v^2\Big)^2}. \tag{3.4}$$

Consider a sequence of graphs $(G_n)_{n\geq 1}$, where $G_n = (V_n, E_n)$ and n denotes the number of vertices $n = |V_n|$ in the graph. Since many real-world networks are quite large, we are interested in the behavior of $\rho(G_n)$ as $n \to \infty$. Note that this discussion applies both to sequences of real-world networks of increasing size, as well as to graph sequences of random graphs. We start by generalizing Theorem 2.3 to this setting:

Theorem 3.1 (Asymptotic non-negativity of Pearson's coefficient in scale-free graphs)**.** *Let $(G_n)_{n\geq 1}$ be a sequence of graphs of size n satisfying that there exist $\gamma \in (1, 3)$ and $0 < c < C < \infty$ such that $cn \leq |E| \leq Cn$, $cn^{1/\gamma} \leq \max_{v\in V_n} D_v \leq Cn^{1/\gamma}$ and $cn^{(2/\gamma)\vee 1} \leq \sum_{v\in V_n} D_v^2 \leq Cn^{(2/\gamma)\vee 1}$. Then, any limit point of Pearson's correlation coefficient $\rho(G_n)$ is non-negative.*

In the next section, we give several examples where Theorem 3.1 applies and yields results that are not sensible. The powerful feature of Theorem 3.1 is that it applies to *all* graphs, not just realizations of certain random graphs.

Proof. We note that $D_v \geq 0$ for every $v \in V$, so that, from (3.4)

$$\rho(G_n) \geq \rho^-(G_n) \equiv -\frac{\frac{1}{|E'|}\Big(\sum_{v\in V} D_v^2\Big)^2}{\sum_{v\in V} D_v^3 - \frac{1}{|E'|}\Big(\sum_{v\in V} D_v^2\Big)^2}. \tag{3.5}$$

By assumption, $\sum_{v\in V} D_v^3 \geq (\max_{v\in[n]} D_v)^3 \geq c^3 n^{3/\gamma}$, whereas $\frac{1}{|E'|}\Big(\sum_{v\in V} D_v^2\Big)^2 \leq (C^2/c)n^{2(2/\gamma\vee 1)-1} = (C^2/c)n^{[(4/\gamma-1)\vee 1]}$. Since $\gamma \in (1,3)$ we have $(4/\gamma - 1) \vee 1 < 3/\gamma$, so that

$$\frac{\sum_{v\in V} D_v^3}{\frac{1}{|E'|}\Big(\sum_{v\in V} D_v^2\Big)^2} \to \infty.$$

Hence, $\rho^-(G_n) \to 0$ as $n \to \infty$. This proves the claim. □

In the literature, many examples are reported of real-world networks where the degree distribution closely follows a power law with γ in $(1,3)$, see e.g., [1, Table I] or [40, Table I]. Let D be such a power-law random variable, and denote $\mu_p = \mathbb{E}[D^p]$ for $p \in (0,\gamma)$. In that case one can expect that

$$|E'| = \sum_{v\in V} D_v \sim \mu_1 n,$$

while $\max_{v\in V} D_v \sim n^{1/\gamma}$, and

$$\frac{1}{n}\sum_{v\in V} D_v^p \sim \begin{cases} \mu_p & \text{when } \gamma > p, \\ C_p n^{p/\gamma - 1} & \text{when } \gamma < p. \end{cases} \tag{3.6}$$

Of course, the convergence in (3.6) depends sensitively on the occurrence of large degrees. However, intuitively it can be explained as follows. When

$$\frac{1}{n}\sum_{v\in V} \mathbb{1}_{\{D_v \geq k\}} = C' k^{-\gamma}(1 + o(1))$$

for all k for which $k^{-\gamma} \gg 1/n$ so that $k \ll n^{1/\gamma}$, then

$$\frac{1}{n}\sum_{v\in V} D_v^p = \sum_{k\geq 1}(k^p - (k-1)^p)\frac{1}{n}\sum_{v\in V} \mathbb{1}_{\{D_v\geq k\}} \approx C'' \sum_{k=1}^{n^{1/\gamma}} k^{p-1-\gamma} = C_p n^{p/\gamma - 1},$$

where C'' and C_p are appropriately chosen constants. In particular, the conditions of Theorem 3.1 hold and $\rho^-(G_n) \to 0$ when $\gamma < 3$. Thus, the asymptotic degree-degree correlation of the graph sequence $(G_n)_{n\geq 1}$ is non-negative. As a result, when the power-law exponent satisfies $\gamma < 3$ there exist no scale-free graph sequences that will be identified as disassortative by Pearson's coefficient. We next investigate a general theorem that allows us to identify the limit of Spearman's rho and Pearson's coefficient for many random graph models.

3.3 Convergence conditions for degree-degree dependency measures

Let $(G_n)_{n\geq 1}$ be again a sequence of graphs of size n, where $G_n = (V_n, E_n)$, $|V_n| = n$. We write $\mathbb{E}_n$ for the conditional expectation given the graph G_n (which in itself is random, so that we are *not* taking the expectation w.r.t. G_n). Consider a random vector $(X,Y) = (D_{\underline{e}}, D_{\overline{e}})$ where e is chosen uniformly at random from E'. Recall that for a discrete random variable X, F_X denotes its cumulative distribution function, and F_X^* denotes the cumulative distribution function of $X^* = X + U$, where U is an independent uniform random variable on $(0,1)$. Then $F_X^*(X^*)$ has a uniform distribution on $(0,1)$, see (2.8). Our main result to identify the limits of Spearman's rho as given by (3.2) and Pearson's coefficient is the following theorem:

Theorem 3.2 (Convergence criteria for degree-degree dependency measures)**.** *Let $(G_n)_{n\geq 1}$ be a sequence of random graphs of size n, where $G_n = (V_n, E_n)$, $|V_n| = n$. Let (X_n, Y_n) be the degrees on both sides of a uniform directed edge $e \in E_n'$. Suppose that for every bounded continuous $h\colon \mathbb{R}^2 \to \mathbb{R}$,*

$$\mathbb{E}_n[h(X_n, Y_n)] \xrightarrow{\mathbb{P}} \mathbb{E}[h(X, Y)], \tag{3.7}$$

where the r.h.s. is non-random. Then
(a)

$$\rho^{\mathrm{rank}}(G_n) \xrightarrow{\mathbb{P}} \rho^{\mathrm{rank}} = 12\mathbb{E}(F_X^*(X^*)F_X^*(Y^*)) - 3 = 12\mathbb{E}(F_X(X)F_X(Y)) - 3, \tag{3.8}$$

where $X^* = X + U$, $Y^* = Y + U'$, U and U' are independent random variables on $(0, 1)$, also independent of X and Y, and $F^*{}_X(\cdot)$ is the cumulative distribution function of X^*;
(b) when we further suppose that $\mathbb{E}_n[X_n^2] \xrightarrow{\mathbb{P}} \mathbb{E}[X^2] < \infty$, and $\mathrm{Var}(X) > 0$, then also

$$\rho(G_n) \xrightarrow{\mathbb{P}} \rho = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}. \tag{3.9}$$

We remark that when G_n is a random graph, then $\rho^{\mathrm{rank}}(G_n)$ and $\rho(G_n)$ are random variables. Equation (3.7) implies that the *distribution* of the degrees on either side of an edge converges in probability to a deterministic limit, which can be interpreted as the statement that the degree distribution converges to a deterministic limit. The limits of $\rho^{\mathrm{rank}}(G_n)$ and $\rho(G_n)$ only depend on the limiting degree distribution, where $\rho^{\mathrm{rank}}(G_n)$ *always* converges, while $\rho(G_n)$ can only be proved to converge when its limit is well defined. We further note that (3.7) is equivalent to showing that

$$\#\{e = (u, v) \in E_n' \colon (D_u, D_v) = (k, l)\}/|E_n'| \xrightarrow{\mathbb{P}} \mathbb{P}(X = k, Y = l). \tag{3.10}$$

Condition (3.10) will be simpler to verify in practice. We emphasize that we study *undirected* graphs but we work with *directed* edges $e = (u, v)$, which we vary over the whole set of edges, in such a way that (u, v) and (v, u) contribute as different edges. In particular, the marginal distributions of X_n and Y_n and consequently of X and Y, are the same. We next prove Theorem 3.2:

Proof. We start with part (a). The sequence $(R_{\underline{e}}/|E_n'|, R_{\overline{e}}/|E_n'|)$ is a bounded sequence of two-dimensional random variables. Let $F_{n,X}$ denote the empirical cumulative distribution function of $(D_{\underline{e}})_{e\in E_n'}$ (which equals that of $(D_{\overline{e}})_{e\in E_n'}$), and let $F_{n,X}^*$ denote the empirical cumulative distribution functions of $(D_{\underline{e}} + U_e)_{e\in E_n'}$ (which equals that of $(D_{\overline{e}} + U_e')_{e\in E_n'}$), where $(U_e)_{e\in E_n'}$, $(U_e')_{e\in E_n'}$ are independent sequences of i.i.d uniform $(0, 1)$ random variables. Then, we can rewrite, with $\ell_n = |E_n'|$,

$$(R_{\underline{e}}, R_{\overline{e}}) = \big((\lceil \ell_n F_{n,X}^*(D_{\underline{e}} + U_e)\rceil, \lceil \ell_n F_{n,X}^*(D_{\overline{e}} + U_e')\rceil\big). \tag{3.11}$$

In particular,

$$(R_{\underline{e}}/\ell_n, R_{\overline{e}}/\ell_n) = \big(\lceil \ell_n F_{n,X}^*(D_{\underline{e}} + U_e)\rceil/\ell_n, \lceil \ell_n F_{n,X}^*(D_{\overline{e}} + U_e')\rceil/\ell_n\big). \tag{3.12}$$

Thus,

$$(R_{\underline{e}}/\ell_n, R_{\overline{e}}/\ell_n) = \big(F_{n,X}^*(D_{\underline{e}} + U_e), F_{n,X}^*(D_{\overline{e}} + U_e')\big) + O(1/\ell_n). \tag{3.13}$$

By (3.7), the fact that $X_n \xrightarrow{d} X$ and the fact that F_X^* is continuous, $F_{n,X}^*(x) \xrightarrow{\mathbb{P}} F_X^*(x)$ for every $x \geq 0$. Moreover, we claim that this convergence holds uniformly in x, i.e., $\sup_{x\in\mathbb{R}} |F_{n,X}^*(x) - F_X^*(x)| \xrightarrow{\mathbb{P}} 0$. To see this, note that (3.7) implies that the distribution functions of X_n and Y_n converge to those of X and Y. Since all these random variables take on only integer values, this convergence is *uniform*, i.e., $\sup_{k\geq 0} |F_{n,X}(k) - F_X(k)| \xrightarrow{\mathbb{P}} 0$. We obtain $F_{n,X}^*$ by linearly interpolating between $F_{n,X}(k-1)$ and $F_{n,X}(k)$ for every k, so also $F_{n,X}^*$ converges uniformly, as we claimed.

By this uniform convergence, for every bounded continuous function $g\colon [0,1]^2 \to \mathbb{R}$,

$$\begin{aligned}\mathbb{E}_n[g(R_{\underline{e}}/\ell_n, R_{\overline{e}}/\ell_n)] &= \mathbb{E}_n[g(F^*_{n,X}(D_{\underline{e}}+U_e), F^*_{n,X}(D_{\overline{e}}+U'_e))] \\ &= \mathbb{E}_n[g(F^*_X(D_{\underline{e}}+U_e), F^*_X(D_{\overline{e}}+U'_e))] + o_{\mathbb{P}}(1) \\ &= \mathbb{E}_n[g(F^*_X(X_n+U), F^*_X(Y_n+U'))] + o_{\mathbb{P}}(1) \\ &\xrightarrow{\mathbb{P}} \mathbb{E}[g(F^*_X(X+U), F^*_X(Y+U'))] = \mathbb{E}[g(F^*_X(X^*), F^*_X(Y^*))],\end{aligned} \tag{3.14}$$

again by (3.7) and the fact that $(x,y) \mapsto \mathbb{E}[g(F^*_X(x+U), F^*_X(y+U'))]$ is continuous and bounded. Applying this to $g(x,y) = xy$, $g(x,y) = x^2$ and $g(x,y) = y^2$ yields the required convergence. Moreover, since $F^*_X(X^*)$ and $F^*_X(Y^*)$ are uniform random variables, $\mathrm{Var}(F^*_X(X^*)) = \mathrm{Var}(F^*_X(Y^*)) = 1/12$. This completes the proof of convergence and the first equality in (a). The second equality is just [31, Proposition 3.1], see (2.9).

For part (b), we note that

$$\rho(G_n) = \frac{\mathrm{Cov}_n(X_n, Y_n)}{\mathrm{Var}_n(X_n)}. \tag{3.15}$$

Since $\mathbb{E}_n[X_n^2] \xrightarrow{\mathbb{P}} \mathbb{E}[X^2] < \infty$, also $\mathbb{E}_n[X_n] \xrightarrow{\mathbb{P}} \mathbb{E}[X] < \infty$, so that $\mathrm{Var}_n(X_n) \xrightarrow{\mathbb{P}} \mathrm{Var}(X)$. Since these limits are positive, by Slutzky's theorem,

$$\rho(G_n) = \frac{\mathrm{Cov}_n(X_n, Y_n)}{\mathrm{Var}(X)}(1 + o_{\mathbb{P}}(1)). \tag{3.16}$$

Furthermore, the random variables $(X_nY_n)_{n\geq 1}$ converge in distribution, and are uniformly integrable (since both $(X_n^2)_{n\geq 1}$ and $(Y_n^2)_{n\geq 1}$ are, which again follows from the fact that $\mathbb{E}_n[X_n^2] \xrightarrow{\mathbb{P}} \mathbb{E}[X^2] < \infty$ and the fact that X_n and Y_n have the same marginals). Therefore, also $\mathbb{E}_n[X_nY_n] \xrightarrow{\mathbb{P}} \mathbb{E}[XY]$, so that the convergence follows. □

4 Random graph examples

In this section we consider four random graph models to highlight our result: the configuration model, the configuration model with intermediate vertices, the preferential attachment model and a model of complete bipartite random graphs. In Section 5, we present the numerical results for these models.

4.1 The configuration model

The *configuration model* (CM) was invented by Bollobás in [7], inspired by [3]. Its connectivity structure was first studied by Molloy and Reed [34, 35]. It was popularized by Newman, Srogatz and Watts [41], who realized that it is a useful and simple model for real-world networks.

Given a *degree sequence*, namely a sequence of n positive integers $\boldsymbol{d} = (d_1, d_2, \ldots, d_n)$ with $\ell_n = \sum_{i\in[n]} d_i$ assumed to be even, the configuration model (CM) on n vertices and degree sequence $\boldsymbol{d}$ is constructed as follows. Start with n vertices, labelled $1, 2, \ldots, n$, and d_v half-edges adjacent to vertex v. The graph is constructed by randomly pairing each half-edge to some other half-edge to form an edge. Number the half-edges from 1 to ℓ_n in some arbitrary order. Then, at each step, two half-edges that are not already paired are chosen uniformly at random among all the unpaired half-edges and are paired to form a single edge in the graph. These half-edges are removed from the list of unpaired half-edges. We continue with this procedure of choosing and pairing two unpaired half-edges until all the half-edges are paired. In the resulting graph $G_n = (V_n, E_n)$ we have $|V_n| = n$, $\ell_n = 2|E_n|$. Although self-loops and double edges may occur, these become rare as $n \to \infty$ (see e.g. [8] or [25] for more precise results in this direction). In the analysis we keep the self-loops and multiple edges, so that $\ell_n = |E'_n|$. In the numerical simulation we also consider the case where

the self-loops are removed, and we collapse multiple edges to a single edge. As we will see in the simulations, these two cases are qualitatively similar.

We investigate the CM where the degrees are i.i.d. random variables, and note that the probability that two vertices u and v are directly connected is close to $d_u d_v/\ell_n$. Since this is of product form in u and v, the degrees at either end of an edge are close to being independent, and in fact are asymptotically independent. Therefore, one expects the assortativity coefficient of the configuration model to converge to 0 in probability, irrespective of the degree distribution.

We now make this argument precise. We make the following assumptions on our degree sequence $(d_v)_{v\in V_n}$:

Condition 4.1 (Degree regularity)**.**
(a) There exists a probability distribution $(p_k)_{k\geq 0}$ such that $n_k/n \to p_k$ for every $k \geq 1$, where $n_k = \#\{v\colon d_v = k\}$ denotes the number of vertices of degree k.
(b) $\mathbb{E}[D_{(n)}] \to \mathbb{E}[D]$, where $\mathbb{P}(D_{(n)} = k) = n_k/n$ and $\mathbb{P}(D = k) = p_k$.

See [23, Chapter 7] for an extensive discussion of the CM under Condition 4.1.

Theorem 4.2 (Convergence of the degree-degree dependency measures for CM)**.** *Let $(G_n)_{n\geq 1}$ be a sequence of configuration models of size n, for which the degree sequence $(d_v)_{v\in V_n}$ satisfies Condition 4.1. Then*

$$\rho^{\mathrm{rank}}(G_n) \xrightarrow{\mathbb{P}} 0,$$

and

$$\rho(G_n) \xrightarrow{\mathbb{P}} 0.$$

Proof. We apply Theorem 3.2, for which we start by investigating (3.10). We note that a uniform edge can be constructed by taking two half-edges uniformly at random. Indeed, we can first draw the first half edge uniformly at random, and this will be paired to another half edge uniformly at random by construction of the CM. We perform a second moment argument on $N_{k,l} = \#\{e = (u, v) \in E_n'\colon (d_u, d_v) = (k, l)\}$, and will prove that

$$N_{k,l}/\ell_n \xrightarrow{\mathbb{P}} \frac{kp_k}{\mathbb{E}[D]}\frac{lp_l}{\mathbb{E}[D]},$$

For this, it suffices to prove that

$$\mathbb{E}[N_{k,l}]/\ell_n \to \frac{kp_k}{\mathbb{E}[D]}\frac{lp_l}{\mathbb{E}[D]}, \qquad \mathbb{E}[N_{k,l}^2]/\ell_n^2 \to \Big(\frac{kp_k}{\mathbb{E}[D]}\frac{lp_l}{\mathbb{E}[D]}\Big)^2,$$

since then $\mathrm{Var}(N_{k,l}/\ell_n) = o(1)$.

We note that

$$\mathbb{E}[N_{k,l}] = \frac{kln_k n_l}{\ell_n - 1},$$

where $\ell_n = \sum_{v\in V_n} d_v = 2|E_n|$ and $n_k = \#\{v\colon d_v = k\}$ is the number of vertices with degree k. Therefore, also using that $\ell_n = n\mathbb{E}[D_{(n)}]$, Condition 4.1 implies that

$$\mathbb{E}[N_{k,l}]/\ell_n \to \frac{kp_k}{\mathbb{E}[D]}\frac{lp_l}{\mathbb{E}[D]}.$$

Further,

$$\mathbb{E}[N_{k,l}^2]/\ell_n^2 = \frac{1}{\ell_n^2}\sum_{(u_1,v_1),(u_2,v_2)} \mathbb{P}(d_{u_1} = k, d_{v_1} = l, d_{u_2} = k, d_{v_2} = l).$$

There are four different cases, depending on $a = \#\{u_1, u_2, v_1, v_2\}$. When $a = 4$, the contribution is

$$\frac{k^2 n_k(n_k - 1)l^2 n_l(n_l - 1)}{\ell_n^2(\ell_n - 1)(\ell_n - 3)} = \frac{(kn_k l n_l)^2}{\ell_n^4}(1 + O(1/n)) \to \Big(\frac{kp_k}{\mathbb{E}[D]}\frac{lp_l}{\mathbb{E}[D]}\Big)^2.$$

Therefore, we are left to show that the contributions due to $a \leq 3$ vanish.

When $a = 3$, either one of the edges (u_1, v_1) and (u_2, v_2) is a self-loop, while the other joins two other vertices (which only contributes when $k = l$), or both edges start in the same vertex v, so that this contribution is at most

$$\frac{k^2 n_k (n_k - 1) l^2 n_l}{\ell_n^2 (\ell_n - 1)(\ell_n - 3)} = O(1/n) = o(1).$$

When $a = 2$, similar computations show that the contribution is at most $O(1/n^2)$. When $a = 1$, the edges (u_1, v_1) and (u_2, v_2) are self-loops from the same vertex v, so that this contributes only when $k = l$, and then at most

$$\frac{k(k-1)(k-2)(k-3) n_k}{\ell_n^2 (\ell_n - 1)(\ell_n - 3)} = O(1/n^3) = o(1).$$

We conclude that (3.10) holds with

$$\mathbb{P}(X = k, Y = l) = \frac{k p_k}{\mathbb{E}[D]} \frac{l p_l}{\mathbb{E}[D]}.$$

In particular, X and Y are independent, so that $\rho^{\mathrm{rank}} = 0$. This proves the first part of Theorem 4.2.

For the second part, we note that when the degrees $(d_v)_{v \in V_n}$ are *fixed*, the only random part in $\rho(G_n)$ is

$$M_n = \frac{1}{\ell_n} \sum_{e \in E_n'} d_{\underline{e}} d_{\overline{e}}.$$

We perform a second moment method on this quantity. We use that an edge e is a pair of two specified half-edges incident to two specific vertices. Thus, we can denote e by $\underline{e} = (u, s), \overline{e} = (v, t)$, where u, v are the vertices to which the specific half-edges are incident, while $s \in \{1, \ldots, d_u\}$ is the label of the half-edge incident to vertex u and $t \in \{1, \ldots, d_v\}$ is the label of the half-edge incident to vertex v, that are paired together. The probability of pairing them together equals $1/(\ell_n - 1)$. Therefore,

$$\mathbb{E}[M_n] = \frac{1}{\ell_n} \sum_{u,v,s,t} \frac{d_u d_v}{\ell_n - 1} = \sum_{u,v \in V_n} d_u^2 d_v^2 / \ell_n (\ell_n - 1) = \sum_{u,v \in V_n} d_u^2 d_v^2 / \ell_n^2 (1 + O(1/n)),$$

where we note that we count multiple edges as frequently as they occur. Further, and in a similar way,

$$\mathbb{E}[M_n^2] = (1 + o(1)) \sum_{u,v,u',v' \in V_n} d_u^2 d_{u'}^2 d_v^2 d_{v'}^2 / \ell_n^4,$$

so that

$$\frac{M_n}{\Big(\sum_{v \in V_n} d_v^2 / \ell_n \Big)^2} \xrightarrow{\mathbb{P}} 1.$$

In particular,

$$\rho(G_n) = \frac{M_n - \Big(\sum_{u,v \in V_n} d_u^2 / \ell_n \Big)^2}{\sum_{u \in V_n} d_u^3 / \ell_n - \Big(\sum_{u \in V_n} d_u^2 / \ell_n \Big)^2} \xrightarrow{\mathbb{P}} 0,$$

both when $\sum_{u \in V_n} d_u^3 / \ell_n \gg \Big(\sum_{u \in V_n} d_u^2 / \ell_n \Big)^2$, as well as when $\sum_{u \in V_n} d_u^3 / \ell_n = \Theta\Big(\sum_{u \in V_n} d_u^2 / \ell_n \Big)^2$. □

4.2 Configuration model with intermediate vertices

We now give an example of a strongly disassortative graph to demonstrate that $\rho(G_n)$ fails to capture obvious negative degree-degree dependencies when the degree distribution is heavy tailed. In order to do that we adapt the configuration model slightly, by replacing every edge by two edges that meet at a middle vertex. Denote this graph by $\bar{G}_n = (\bar{V}_n, \bar{E}_n)$, while the configuration model is $G_n = (V_n, E_n)$. In this model, there are $n + \ell_n/2$ vertices and $|\bar{E}'_n| = 2\ell_n$ directed edges. For $(u, v) \in \bar{E}'_n$, the degree of either vertex u or vertex v equals 2, and the degree of the other vertex in the edge is equal to d_s, where s is the unique vertex in the original configuration model that corresponds to u or v.

Theorem 4.3 (Convergence of degree-degree dependency measures for CM with intermediate vertices)**.** *Let $(\bar{G}_n)_{n\geq 1}$ be a sequence of configuration models with intermediate vertices, where the degree sequence $(d_v)_{v\in V_n}$ satisfies Condition 4.1. Then*

$$\rho^{\mathrm{rank}}(\bar{G}_n) \xrightarrow{\mathbb{P}} 12\mathbb{E}(F_X(X)F_X(Y)) - 3 = -\frac{3}{4} + 3\left(\tilde{p}_1 + \frac{1}{2}\tilde{p}_2\right)\left(1 - \tilde{p}_1 - \frac{1}{2}\tilde{p}_2\right), \tag{4.1}$$

where $(X, Y) = (2I + (1 - I)\tilde{D}_1, 2(1 - I) + I\tilde{D}_2)$ with $\tilde{D}_1, \tilde{D}_2$ i.i.d. random variables with $\mathbb{P}(\tilde{D} = k) = kp_k/\mathbb{E}[D] := \tilde{p}_k$ and I an independent Bernoulli(1/2) random variable. Further,

$$\rho(G_n) \xrightarrow{\mathbb{P}} \begin{cases} \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} & \text{if } \mathbb{E}[D^3_{(n)}] \to \mathbb{E}[D^3] < \infty; \\ 0 & \text{if } \mathbb{E}[D^3_{(n)}] \to \infty, \end{cases}$$

and, for $\mathbb{E}[D^3_{(n)}] \to \mathbb{E}[D^3] < \infty$, and writing $\mu_p = \mathbb{E}[D^p]$,

$$\frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = \frac{2\mu_2/\mu_1 - (1 + \mu_2/(2\mu_1))^2}{(2 + \mu_3/(2\mu_1)) - (1 + \mu_2/(2\mu_1))^2} < 0.$$

The fact that the degree-degree correlation is negative is quite reasonable, since in this model, vertices of high degree are label only connected to vertices of degree 2, so that there is a negative dependence between the degrees at either end of an edge. When $\mathbb{E}[D^3_{(n)}] \to \infty$, on the other hand, $\rho(\bar{G}_n) \xrightarrow{\mathbb{P}} 0$, which is inappropriate, as the negative dependence of the degrees persists.

Proof. The first part follows directly from Theorem 3.2, since the collection of values $(\bar{d}_{\underline{e}}, \bar{d}_{\overline{e}})_{e\in\bar{E}'_n}$ only depends on the degrees $(d_v)_{v\in V_n}$ and

$$\#\{e\colon \bar{d}_{\underline{e}} = l, \bar{d}_{\overline{e}} = k\}/|\bar{E}'_n| = (kn_k\delta_{2,l} + ln_l\delta_{2,k} - 2n_2\mathbb{1}_{\{k=l=2\}})/(2\ell_n),$$

which converges to $\mathbb{P}(X = k, Y = 2)$. Now, consider the possible values of X, and notice that

$$\mathbb{P}(X = 1) = \tilde{p}_1/2, \tag{4.2}$$
$$\mathbb{P}(X = 2) = 1/2 + \tilde{p}_2/2, \tag{4.3}$$
$$\mathbb{P}(X \geq 3) = 1/2 - \tilde{p}_1/2 - \tilde{p}_2/2. \tag{4.4}$$

Then we obtain

$$F^*_X(x + U) = \begin{cases} \frac{1}{2}\tilde{p}_1 U, & \text{if } x = 1, \\ \frac{\tilde{p}_1}{2} + \left(\frac{\tilde{p}_2}{2} + \frac{1}{2}\right)U, & \text{if } x = 2, \\ \frac{1}{2} + \sum_{k=1}^{x-1}\frac{\tilde{p}_k}{2} + \frac{\tilde{p}_x}{2}U, & \text{if } x \geq 3. \end{cases} \tag{4.5}$$

Since either X or Y equals 2 and corresponds to the intermediate node, we further condition on $\tilde{D}$:

$$\begin{aligned} \mathbb{E}(F^*_X(X^*)F^*_X(Y^*)) &= \mathbb{E}(F^*_X(\tilde{D} + U)F^*_X(2 + U')) \\ &= \mathbb{E}(F^*_X(2 + U')) \\ &\times \left[(\mathbb{E}(F^*_X(1 + U))\mathbb{P}(\tilde{D} = 1) + \mathbb{E}(F^*_X(2 + U))\mathbb{P}(\tilde{D} = 2) + \mathbb{E}(\tilde{D} + U|\tilde{D} \geq 3)\mathbb{P}(\tilde{D} \geq 3)\right]. \end{aligned} \tag{4.6}$$

Now, using (4.5) and substituting (4.2–4.4), from the last expression we readily obtain

$$\begin{aligned}\mathbb{E}(F_X^*(X^*)F_X^*(Y^*)) &= \left(\frac{\tilde{p}_1}{2}+\frac{\tilde{p}_2}{4}+\frac{1}{4}\right)\\ &\qquad\times\left[\frac{1}{4}(\tilde{p}_1)^2+\Big(\frac{\tilde{p}_1}{2}+\frac{\tilde{p}_2}{4}+\frac{1}{4}\Big)\tilde{p}_2+\Big(\frac{\tilde{p}_1}{4}+\frac{\tilde{p}_2}{4}+\frac{3}{4}\Big)(1-\tilde{p}_1-\tilde{p}_2)\right]\\ &= \frac{3}{16}+\frac{1}{4}\left(\tilde{p}_1+\frac{1}{2}\tilde{p}_2\right)\left(1-\tilde{p}_1-\frac{1}{2}\tilde{p}_2\right).\end{aligned}$$

Substituting this in (3.8) and again using (2.9) we obtain (4.1).

For the second part, we compute

$$\frac{1}{|\bar{E}_n'|}\sum_{(u,v)\in\bar{E}_n'}\bar{d}_u\bar{d}_v=\frac{2}{\ell_n}\sum_{v\in V_n}d_v^2,$$

and for $p\geq 2$,

$$\frac{1}{|\bar{E}_n'|}\sum_{s\in\bar{V}_n}\bar{d}_s^p=\frac{1}{2\ell_n}2^p(\ell_n/2)+\frac{1}{2\ell_n}\sum_{v\in V_n}d_v^p=2^{p-2}+\frac{1}{2\ell_n}\sum_{v\in V_n}d_v^p,$$

As a result, when $\mathbb{E}[D_{(n)}^3]\to\mathbb{E}[D^3]<\infty$, we have

$$\rho(\bar{G}_n)\xrightarrow{\mathbb{P}}\frac{2\mu_2/\mu_1-(1+\mu_2/(2\mu_1))^2}{(2+\mu_3/(2\mu_1))-(1+\mu_2/(2\mu_1))^2}<0,$$

where $\mu_p=\mathbb{E}[D^p]$. □

4.3 Preferential attachment model

We discuss the general Preferential Attachment model (PAM), as formulated, for example, in [23, Chapter 8] or [16, Chapter 4]. The PAM is a *dynamical* random graph model, and thus models a growing network. It is defined in terms of two parameters, m, which denotes the number of edges of newly added vertices, and $\delta>-m$, which quantifies the tendency to attach to vertices that already have a high degree. We start by defining the model for $m=1$.

We start with one vertex having one self-loop. Suppose we have the graph of size t, which we denote by $G_t^{(1)}$. Let i label the vertex that appeared at time $i=1,2,\ldots$. Then, $G_{t+1}^{(1)}$ is constructed by adding one extra vertex that has one edge, which forms a self-loop with probability $(1+\delta)/((2+\delta)t+1+\delta)$ and, conditionally on $G_t^{(1)}$, attaches to a vertex $v\in[t]$ with probability $(D_i(t)+\delta)/((2+\delta)t+1+\delta)$, where $D_i(t)$ is the random degree of vertex i in $G_t^{(1)}$. As a result, vertices with high degree have a higher probability to be attached to, which explains the name *preferential attachment model*.

The model with $m\geq 2$ is obtained from the model with $m=1$ as follows. Collapse vertices $m(s-1)+1,\ldots,ms$, and all of their edges, in $(G_t^{(1)})_{t\geq 1}$ with δ replaced by $\delta'=\delta/m$ to form vertex s in $(G_t^{(m)})_{t\geq 1}$ with parameter δ. It is well known (see e.g., [9] where this was first derived for $\delta=0$ and [23, Theorem 8.3] as well as the references in [23] for a more detailed literature overview) that the resulting graph has an asymptotic degree sequence p_k, i.e.,

$$N_k(t)/t=\#\{i\in[t]\colon D_i(t)=k\}/t\xrightarrow{\mathbb{P}}p_k,\tag{4.7}$$

where, for $k\geq m$,

$$p_k=(2+\delta/m)\frac{\Gamma(k+\delta)\Gamma(m+2+\delta+\delta/m)}{\Gamma(m+\delta)\Gamma(k+3+\delta+\delta/m)}.\tag{4.8}$$

In particular, the PAM is scale free with power-law exponent $\gamma=2+\delta/m$. See [23, Section 8.2] for more details on the scale-free behavior of the PAM. The next theorem investigates the behaviour of Pearson's correlation coefficient as well as Spearman's rho for the PAM:

Theorem 4.4 (Convergence of degree-degree dependency measures for PAM)**.** *Let $(G_t^{(m)})_{t\geq 1}$ be the PAM. Then*

$$\rho^{\mathrm{rank}}(G_t^{(m)}) \xrightarrow{\mathbb{P}} \rho^{\mathrm{rank}}, \tag{4.9}$$

while

$$\rho(G_t^{(m)}) \xrightarrow{\mathbb{P}} \begin{cases} 0 & \text{if } \delta \leq m, \\ \rho & \text{if } \delta > m, \end{cases} \tag{4.10}$$

where, abbreviating $a = \delta/m$,

$$\rho = \frac{(m-1)(a-1)[2(1+m)+a(1+3m)]}{(1+m)[2(1+m)+a(5+7m)+a^2(1+7m)]}. \tag{4.11}$$

The value of ρ in (4.11) was predicted in [14], and we make this analysis mathematically rigorous. The remainder of the section is the proof of Theorem 4.4. It involves intermediate technical results formulated as Lemma's 4.5–4.9 below.

For the PAM, it will be convenient to direct the edges from young to old, so that there are mt directed edges. Let $N_{k,l}(t)$ denote the number of directed edges e for which $D_{\underline{e}}(t) = k$, $D_{\overline{e}}(t) = l$. We will prove that there exists a probability distribution $(q_{k,l})_{k,l\geq m}$ such that

$$N_{k,l}(t)/(mt) \xrightarrow{\mathbb{P}} q_{k,l}. \tag{4.12}$$

Since a uniform directed edge oriented from young to old can be obtained by taking a uniform vertex and then a uniform edge coming out of this vertex, this proves (3.10) with

$$p_{kl} = \mathbb{P}(X = k, Y = l) = \tfrac{1}{2}(q_{k,l} + q_{l,k}). \tag{4.13}$$

In particular, by Theorem 3.2(a), this proves (4.9) in Theorem 4.4. We follow the proof of [23, Theorem 8.2], which, in turn, is strongly inspired by the proof in [9].

Proofs for convergence of the degree sequence typically consist of two key steps. The first is a martingale concentration argument in Lemma 4.5.

Lemma 4.5 (Convergence of degree-degree counts)**.** *For every k, l, there exists a $C > 0$ such that,*

$$\mathbb{P}\Big(\max_{k,l} |N_{kl}(t) - \mathbb{E}[N_{kl}(t)]| \geq C\sqrt{t\log t}\Big) = o(1). \tag{4.14}$$

Proof. The proof for the degree distribution in [23] applies almost verbatim (see, in particular, [23, Proposition 8.4] and its proof). Indeed, the proof relies on a martingale argument. Define the Doob-martingale, for $t = 0, \ldots, n$,

$$M_n = \mathbb{E}[N_{kl}(t) \mid G_n^{(m)}].$$

The crucial observation is that $(M_n)_{n=0}^t$ is a martingale with $M_t = N_{kl}(t)$ and $M_0 = \mathbb{E}[N_{kl}(t)]$ that satisfies

$$|M_n - M_{n-1}| \leq 4m. \tag{4.15}$$

We prove (4.15) below. The Azuma-Hoeffding inequality [2, 22] then proves (4.14) for any $C > 4[4m]^2$. Indeed,

$$\mathbb{P}\Big(|N_{kl}(t) - \mathbb{E}[N_{kl}(t)]| \geq A\Big) = \mathbb{P}\Big(|M_t - M_0| \geq A\Big) \leq \mathrm{e}^{-A^2/(2t[4m]^2)}.$$

Taking $A = C\sqrt{t\log t}$ with $C^2 > 4[4m]^2$ proves that

$$\mathbb{P}\Big(|N_{kl}(t) - \mathbb{E}[N_{kl}(t)]| \geq C\sqrt{t\log t}\Big) = o(1/t^2),$$

so that even

$$\mathbb{P}\Big(\max_{k,l}|N_{kl}(t)-\mathbb{E}[N_{kl}(t)]|\geq C\sqrt{t\log t}\Big)$$
$$\leq (mt)^2\max_{k,l}\mathbb{P}\Big(\max_{k,l}|N_{kl}(t)-\mathbb{E}[N_{kl}(t)]|\geq C\sqrt{t\log t}\Big)=o(1).$$

This completes the proof of Lemma 4.5 assuming (4.15).

We complete the proof by deriving (4.15). For this, it will be convenient to introduce some further notation. Let $e\in[mt]$ label the edges. Let $v_e=\lceil e/m\rceil$ denote the vertex from which the eth edge emanates, and V_e (which is a random variable) the vertex to which the eth edge points. Then,

$$N_{k,l}(t)=\sum_{e\in[mt]}\mathbb{1}_{\{D_{v_e}(t)=k,D_{V_e}(t)=l\}}.$$

As a result,

$$M_n-M_{n-1}=\sum_{e\in[mt]}\Big[\mathbb{P}(D_{v_e}(t)=k,D_{V_e}(t)=l\mid G_n)-\mathbb{P}(D_{v_e}(t)=k,D_{V_e}(t)=l\mid G_{n-1})\Big],$$

where we abbreviate $G_n=G_n^{(m)}$. We let $(G'_l)_{l\geq 0}$ denote the PAM with $G'_{n-1}=G_{n-1}$, while the evolution of $(G'_l)_{l\geq 0}$ after time $n-1$ is the same in distribution as that of $(G_l)_{l\geq 0}$, but conditionally independent of it given $G_{n-1}=G'_{n-1}$. Let $D'_i(t)$ denote the degree of vertex i in G'_t. Then,

$$\begin{aligned}\mathbb{P}(D_{v_e}(t)=k,D_{V_e}(t)=l\mid G_{n-1})&=\mathbb{P}(D'_{v_e}(t)=k,D'_{V_e}(t)=l\mid G_{n-1})\\&=\mathbb{P}(D'_{v_e}(t)=k,D'_{V_e}(t)=l\mid G_{n-1},G_n\setminus G_{n-1}),\end{aligned}$$

where $G_n\setminus G_{n-1}$ is shorthand for the edges of G_n that are not in G_{n-1}. The last step is due to the conditional independence of the evolution after time $n-1$ in $(G'_t)_{t\geq 0}$. Thus,

$$\mathbb{P}(D_{v_e}(t)=k,D_{V_e}(t)=l\mid G_{n-1})=\mathbb{P}(D'_{v_e}(t)=k,D'_{V_e}(t)=l\mid G_n).$$

We conclude that

$$M_n-M_{n-1}=\sum_{e\in[mt]}\Big[\mathbb{P}(D_{v_e}(t)=k,D_{V_e}(t)=l\mid G_n)-\mathbb{P}(D'_{v_e}(t)=k,D'_{V_e}(t)=l\mid G_n)\Big].$$

When $V_e>n$, clearly $\mathbb{P}(D_{v_e}(t)=k,D_{V_e}(t)=l\mid G_n)=\mathbb{P}(D'_{v_e}(t)=k,D'_{V_e}(t)=l\mid G_n)$, as the degrees of vertices i with $i>n$ are independent of G_n. Thus, we can restrict to $V_e\leq n$. Further, when $v_e>n$, then $D_{v_e}(t)$ is independent of G_n, so that

$$\begin{aligned}&\mathbb{P}(D_{v_e}(t)=k,D_{V_e}(t)=l\mid G_n)-\mathbb{P}(D'_{v_e}(t)=k,D'_{V_e}(t)=l\mid G_n)\\&=\mathbb{P}(D_{v_e}(t)=k)\Big[\mathbb{P}(D_{V_e}(t)=l\mid G_n)-\mathbb{P}(D'_{V_e}(t)=l\mid G_n)\Big].\end{aligned}$$

Note that $D_{V_e}(n-1)=D'_{V_e}(n-1)$ a.s., $\mathbb{P}(D_{V_e}(t)=l\mid G_n,D_{V_e}(n)=j)=\mathbb{P}(D_{V_e}(t)=l\mid D_{V_e}(n)=j)$, and

$$\mathbb{P}(D'_{V_e}(t)=l\mid G_n,D'_{V_e}(n)=j)=\mathbb{P}(D'_{V_e}(t)=l\mid D'_{V_e}(n)=j)=\mathbb{P}(D_{V_e}(t)=l\mid D_{V_e}(n)=j).$$

Thus, using that

$$\begin{aligned}\mathbb{P}(D_{V_e}(t)=l\mid G_n)&=\mathbb{E}[\mathbb{P}(D'_{V_e}(t)=l\mid D_{V_e}(n))\mid G_n],\\\mathbb{P}(D'_{V_e}(t)=l\mid G_n)&=\mathbb{E}[\mathbb{P}(D'_{V_e}(t)=l\mid D'_{V_e}(n))\mid G_n],\end{aligned}$$

we obtain at

$$|\mathbb{P}(D'_{V_e}(t)=l\mid D_{V_e}(n))-\mathbb{P}(D'_{V_e}(t)=l\mid D'_{V_e}(n))|\leq\mathbb{1}_{\{D_{V_e}(n)\neq D'_{V_e}(n)\}}.$$

Taking expectations yields

$$\Big|\mathbb{P}(D_{v_e}(t)=k, D_{V_e}(t)=l \mid G_n) - \mathbb{P}(D'_{v_e}(t)=k, D'_{V_e}(t)=l \mid G_n)\Big| \le \mathbb{P}(D_{V_e}(n) \neq D'_{V_e}(n) \mid G_n).$$

In a similar way, we see that for $v_e \le n$,

$$\begin{aligned}&|\mathbb{P}(D_{v_e}(t)=k, D_{V_e}(t)=l \mid G_n) - \mathbb{P}(D'_{v_e}(t)=k, D'_{V_e}(t)=l \mid G_n)|\\ &\quad\le \mathbb{P}(D_{V_e}(n) \neq D'_{V_e}(n) \mid G_n) + \mathbb{P}(D_{v_e}(n) \neq D'_{v_e}(n) \mid G_n).\end{aligned}$$

We conclude that

$$|M_n - M_{n-1}| \le \sum_{e\in[mt]} \big[\mathbb{P}(D_{V_e}(n) \neq D'_{V_e}(n) \mid G_n) + \mathbb{P}(D_{v_e}(n) \neq D'_{v_e}(n) \mid G_n)\big] \le 4m. \qquad \square$$

We continue with the proof of (4.12). The second key step the proof of (4.12) is to prove that, for each k, l,

$$\lim_{t\to\infty} \mathbb{E}[N_{kl}(t)]/(mt) = q_{k,l}. \tag{4.16}$$

We sum over the vertex s that has degree l at time t, and condition on the degree $r \ge m$ of the vertex to which the edge of vertex s is attached. This yields

$$\mathbb{E}[N_{kl}(t)] = m\sum_{s=1}^{t}\sum_{r\ge m} \frac{(r+\delta)}{(2m+\delta)s}\mathbb{E}[N_r(s)]\Big[\mathbb{P}\big(B_{r+1}[s+1,t]=k, B_m[s+1,t]=l\big) + O(1/s)\Big], \tag{4.17}$$

where $B_m[s+1,t]$ is m plus the number of edges attached to vertex s between time $s+1$ and t, while $B_{r+1}[s+1,t]$ is r plus the number of further edges attached to the vertex of degree r to which the edge of vertex s is attached. The $O(1/s)$ term is due to contributions where at least *two* edges of vertex s are attached to the same vertex of degree r, and also due to the fact that the probability of attaching the jth edge of vertex s to a vertex of degree r at time s is actually equal to $\frac{(r+\delta)}{(2m+\delta)s+(j-1)(2+\delta/m)+1+\delta/m}$, which is $\frac{(r+\delta)}{(2m+\delta)s}(1+O(1/s))$. Further,

$$\mathbb{P}\big(B_{r+1}[s+1,t]=k, B_m[s+1,t]=l\big) = \mathbb{P}(B_{r+1}[s+1,t]=k)\mathbb{P}(B_m[s+1,t]=l) + O(1/t),$$

since the dependence between the two probabilities is entirely due to the fact that edges that contribute to $B_{r+1}[s+1,t]$ cannot contribute to $B_m[s+1,t]$. Indeed, $(B_{r+1}[s+1,t], B_m[s+1,t])$ is equal in distribution to the number of balls in two urns at time $m(t-s)$, where we start with $r+1$ and m balls at time 0, and in each draw, we draw a ball in each of the urns with probability equal to the number of balls plus δ and then replace it with two balls. Knowing how many balls are put into the first urn only gives us information about how many balls cannot be put into the second urn, so the balls in the different urns are close to independent. We study these probabilities now:

Lemma 4.6 (Growth of degrees in PAM)**.** *For all $k \ge r \ge m$ and $a \in (0,1)$,*

$$\lim_{s\to\infty} \mathbb{P}(B_r[as,s]=k) = P_k(a;r),$$

where, for each $r \ge m$ and $a \in (0,1)$, $(P_k(a;r))_{k\ge r}$ is a probability measure.

Proof. We note that $(B_r[s,ts])_{t\ge1} \xrightarrow{d} (Z_t)_{t\ge1}$, as $s\to\infty$, where $(Z_t)_{t\ge0}$ is a pure birth process, which increases by 1 at rate $m(Z_t+\delta)/((2m+\delta)t)$ at time t. Indeed, when $B_r[s,ts]=k$, then each of the m edges of vertex $st+1$ has probability $(k+\delta)/[(2m+\delta)(st)] + O(1/s^2)$ of being attached to the vertex that has degree k at time ts, and thus of increasing $B_r[s,ts]$ to $k+1$. Thus, within a short time interval $[t,t+dt]$ and conditionally on $B_r[s,ts]=k$, the probability that $B_r[s,(t+dt)s]=k+1$ is equal to

$$sdt\Big[m(k+\delta)/[(2m+\delta)(st)] + O(1/s^2) + o(1)\Big] \to dt\frac{m(k+\delta)}{(2m+\delta)t} + o(dt),$$

as $s \to \infty$. This is the birth rate of the pure birth process $(Z_t)_{t\geq 1}$.

We next study the limiting birth process, for which is it useful to make a time change. With $b_t = Z_{\mathrm{e}^{(2+\delta/m)t}}$, $(b_t)_{t\geq 0}$ is a birth process that grows at rate $b_t + \delta$ at time t. Define

$$f_{r,k}(t) = \mathbb{P}(b_t = k \mid b_0 = r).$$

Then,

$$\frac{\partial}{\partial t} f_{r,k}(t) = -(k+\delta) f_{r,k}(t) + (k-1+\delta) f_{r,k-1}(t).$$

This set of differential equations is solved by $f_{r,r}(t) = \mathrm{e}^{-(r+\delta)t}$ and, for $k \geq r+1$,

$$f_{r,k}(t) = (k-1+\delta)\mathrm{e}^{-(k+\delta)t} \int_0^t \mathrm{e}^{(k+\delta)s} f_{r,k-1}(s) ds.$$

This can be solved by, for $k \geq r+1$,

$$f_{r,k}(t) = \mathbb{P}(b_t = i \mid b_0 = r) = \frac{\Gamma(k+\delta)}{\Gamma(r+\delta)} \mathrm{e}^{-(k+\delta)t} \sum_{j=0}^{k-r} \alpha_{j,k} \mathrm{e}^{jt},$$

where $\alpha_{0,k} = -\sum_{j=0}^{k-1} \alpha_{j,k-1}/(j+1)$, while, for $j \geq 1$,

$$\alpha_{j,k} = \alpha_{j-1,k-1}/j.$$

As a result, for all $a \in (0,1)$,

$$\lim_{t\to\infty} \mathbb{P}(B_r[at,t] = k) = \mathbb{P}(Z_{1/a} = k \mid Z_1 = r) = f_{r,k}((2+\delta/m)^{-1}\log(1/a)).$$

Note that $P_r(a;r)$ is the probability that the birth process has no births. We thus compute that $P_r(a;r) = f_{r,r}((2+\delta/m)^{-1}\log(1/a)) = a^{(r+\delta)/(2+\delta/m)}$ for $k = r$, while

$$P_k(a;r) = f_{r,k}((2+\delta/m)^{-1}\log(1/a)) = \frac{\Gamma(k+\delta)}{\Gamma(r+\delta)} a^{(k+\delta)/(2+\delta/m)} \sum_{j=0}^{k-r} \alpha_{j,k} a^{-j/(2+\delta/m)}. \qquad \square$$

We continue from (4.17), and rewrite it as

$$\mathbb{E}[N_{kl}(t)]/(mt) = \sum_{r\geq m} \mathbb{E}\Big[\frac{(r+\delta)}{(2m+\delta)Ut} \mathbb{E}[N_r(Ut)] \mathbb{P}(B_{r+1}[Ut,t] = k \mid U) \mathbb{P}(B_m[Ut,t] = l \mid U)\Big] + O(\log t/t), \tag{4.18}$$

where U has a uniform distribution, we interpret $Ut = \lceil Ut \rceil$, and the outer expectation is over U only. Using that $\mathbb{E}[N_r(s)]/s = p_r + O(1/s)$ (see [23, Proposition 8.4]), we further arrive at

$$\mathbb{E}[N_{kl}(t)]/(mt) = \sum_{r\geq m} \frac{r+\delta}{2m+\delta} p_r \mathbb{E}\Big[\mathbb{P}(B_{r+1}[Ut,t] = k \mid U) \mathbb{P}(B_m[Ut,t] = l \mid U)\Big] + o(1). \tag{4.19}$$

By Lemma 4.6, this converges to

$$\mathbb{E}[N_{kl}(t)]/(mt) \to q_{k,l} \equiv \sum_{r\geq m} \frac{r+\delta}{2m+\delta} p_r \mathbb{E}[P_k(U;r) P_l(U;m)]. \tag{4.20}$$

This proves (4.16), and thus, by Theorem 3.2(a), proves the convergence of the rank correlation in (4.9) in Theorem 4.4.

For the convergence of the correlation coefficient in (4.10) in Theorem 4.4, we aim to use Theorem 3.2(b) and thus start by investigating the convergence of moments of X_n. By (3.3), and letting $\mathbb{E}_n$ denote the conditional expectation given G_n,

$$\mathbb{E}_n[X_n^2] = \frac{1}{n}\sum_{i\in[n]} D_i(n)^3.$$

Thus, we are lead to studying sums of powers of degrees. To analyze the limit of sums of powers of degrees, we rely on the following lemma:

Lemma 4.7 (Sum of powers of degrees in PAM)**.** *For all $p < \gamma = 2 + \delta/m$,*

$$\frac{1}{n}\sum_{i\in[n]} D_i(n)^p \xrightarrow{\mathbb{P}} \mu_p = \sum_{k\geq m} k^p p_k < \infty.$$

Proof. We note that $\sum_{i\in[n]} D_i(n)^p = \sum_{k\geq m} k^p N_k(n)$. Under the conditions stated, for every $k_n \to \infty$,

$$\sum_{k\geq m} k^p N_k(n) = \sum_{m\leq k\leq k_n} k^p N_k(n) + o_{\mathbb{P}}(n).$$

This follows since, for any $\varepsilon > 0$, $k > k_n$ implies that $k^\varepsilon/k_n^\varepsilon > 1$, so that

$$\sum_{k>k_n} k^p N_k(n) \leq k_n^{-\varepsilon}\sum_{k\geq m} k^{p+\varepsilon} N_k(n) = k_n^{-\varepsilon}\frac{1}{n}\sum_{i\in[n]} D_i(n)^{p+\varepsilon}.$$

By the analysis in [23, Section 8.1 and 8.6], when $p + \varepsilon < \gamma + 1 = 3 + \delta/m$,

$$\limsup_{n\to\infty} \mathbb{E}\Big[\frac{1}{n}\sum_{i\in[n]} D_i(n)^{p+\varepsilon}\Big] < \infty.$$

Therefore, by the Markov inequality, $\sum_{k>k_n} k^p N_k(n) = o_{\mathbb{P}}(n)$.

Now, since $\max_k |N_k(n) - p_k| \leq \sqrt{Cn\log n}$ whp by [23, Proposition 8.4],

$$\sum_{m\leq k\leq k_n} k^p N_k(n) = t\sum_{m\leq k\leq k_n} k^p p_k + O_{\mathbb{P}}(k_n^{p+1}\sqrt{n\log n}).$$

This proves the claim. □

It follows from Lemma 4.7 that for $3 < \gamma = 2 + \delta/m$,

$$\mathbb{E}_n[X_n^2] = \frac{1}{n}\sum_{i\in[n]} D_i(n)^3 = B(1 + o_{\mathbb{P}}(1)).$$

where B is a constant. As a result,

$$\rho(G_n) \xrightarrow{a.s.} \rho = \mathrm{Cov}(X,Y)/\mathrm{Var}(X) = \frac{\sum_{k,l} klq_{k,l} - \mathbb{E}[X]^2}{\mathbb{E}[X^2] - \mathbb{E}[X]^2}. \tag{4.21}$$

This proves (4.10) in Theorem 4.4 when $\delta > m$. For $\gamma < 3$, instead, $D_1(n)/n^{1/\gamma} \xrightarrow{a.s.} \xi$, for some strictly positive random variable ξ (see e.g., [23, Sections 8.1 and 8.6]). Therefore, $\mathbb{E}_n[X_n^2] \geq \xi^3 n^{3/\gamma-1}(1 + o(1))$. Further, the majority of edges of high degree vertices is young, so that

$$\mathbb{E}_n[X_nY_n] = o_{\mathbb{P}}(n^{3/\gamma-1}). \tag{4.22}$$

Indeed, fix T_n such that $T_n \to \infty$ and $T_n = o(n)$. There are at most mT_n edges between vertices with index at most T_n, and, since the maximal degree is $O_{\mathbb{P}}(n^{1/\gamma})$, these contribute at most $O_{\mathbb{P}}(n^{2/\gamma-1}T_n)$.

For the other edges, one of the vertices involved was born after time T_n. Since $\max_{i\geq T_n} D_i(n) = o_{\mathbb{P}}(n^{1/\gamma})$, the contribution of these edges is at most

$$o_{\mathbb{P}}(n^{1/\gamma})\mathbb{E}_n[X_n + Y_n].$$

In turn, $\mathbb{E}_n[X_n + Y_n] = O_{\mathbb{P}}(n^{(2/\gamma-1)\wedge 1})$, which completes the proof of (4.22). This implies that $\rho(G_n) \xrightarrow{\mathbb{P}} 0$, which proves (4.10) in Theorem 4.4 when $\delta < m$. For $\delta = m$, so that $\gamma = 3$, $\sum_{i\in[n]} D_i(n)^3 = \Theta_{\mathbb{P}}(n\log n)(1 + o_{\mathbb{P}}(1))$. As a result, also in this case $\rho(G_n) \xrightarrow{a.s.} 0$ for $\delta \leq m$. □

We continue with the proof of (4.11) in Theorem 4.4. To compute expectations involving X, we often rely on the following lemma:

Lemma 4.8 (Degree on one side of uniform edge)**.** *For every function $f\colon \mathbb{N} \to \mathbb{R}$,*

$$\mathbb{E}[f(X)] = \sum_{k\geq m} f(k)\frac{kp_k}{2m}.$$

Proof. Let f be bounded, and let X_n be the degree at the bottom of a uniform edge. Then,

$$\mathbb{E}[f(X_n) \mid G_n^{(m)}] = \frac{1}{|E_n'|}\sum_{e\in E_n'} f(D_{\underline{e}}(n)) = \frac{1}{2mn}\sum_{v\in[n]} f(D_v(n))D_v(n) = \frac{1}{2m}\sum_{k\geq m} f(k)kN_k(n)/n.$$

Taking the limit of $n \to \infty$ and using that $N_k(n)/n \xrightarrow{\mathbb{P}} p_k$, as well as $X_n \xrightarrow{d} X$ proves the claim. □

Lemma 4.8 allows us to identify the r.h.s. of (4.21) as

$$\rho = \mathrm{Cov}(X, Y)/\mathrm{Var}(X) = \frac{(2m)^2\sum_{k,l} klq_{k,l} - \lambda_2^2}{2m\lambda_3 - \lambda_2^2},$$

where $\lambda_a = \sum_{k\geq m} k^a p_k$. To identify the limit, we follow [14]. Recall the definition of p_{kl} in (4.13).

Lemma 4.9 (Asymptotic degree-degree distribution for PAM)**.** *For all $k, l \geq m$,*

$$\begin{aligned} p_{kl} &= \mathbb{P}(X = k, Y = l) \qquad (4.23)\\ &= (2+\delta/m)\frac{\Gamma(m+2+\delta+\delta/m)}{\Gamma(m+\delta)^2}\frac{\Gamma(l+\delta)\Gamma(k+\delta)}{\Gamma(k+2+\delta)\Gamma(l+k+2+2\delta+\delta/m)}\\ &\quad\times\Big[\sum_{j=m+1}^{k}\binom{k+l-j-m}{l-m}\binom{j+k+2+2\delta+\delta/m}{k+1+\delta} + \sum_{j=m+1}^{l}\binom{k+l-j-m}{k-m}\binom{j+l+2+2\delta+\delta/m}{l+1+\delta}\Big]. \end{aligned}$$

Consequently, (4.11) *follows.*

Proof. To compute $\mathbb{P}(X = k, Y = l)$, we let $M_{kl}(t)$ denote the number of edges at time t where one side has degree k and the other side degree l, so that

$$p_{kl} = \lim_{t\to\infty} \mathbb{E}[M_{kl}(t)]/(2mt).$$

We note that $M_{kl}(t)$ satisfies the recursion relation

$$\begin{aligned} \mathbb{E}[M_{kl}(t+1)] - \mathbb{E}[M_{kl}(t)] &= m\frac{(k\vee l) - 1 + \delta}{(2m+\delta)t}\mathbb{E}[N_{k\vee l-1}(t)]\mathbb{1}_{\{k\wedge l = m\}}\\ &\quad + m\frac{k-1+\delta}{(2m+\delta)t}\mathbb{E}[M_{k-1,l}(t)] + m\frac{l-1+\delta}{(2m+\delta)t}\mathbb{E}[M_{k,l-1}(t)]\\ &\quad - m\frac{k+\delta}{(2m+\delta)t}\mathbb{E}[M_{k,l}(t)] - m\frac{l+\delta}{(2m+\delta)t}\mathbb{E}[M_{k,l}(t)] + O(1/t^2). \end{aligned}$$

It is not clear that the left-hand side converges since we only know that $\mathbb{E}[M_{k,l}(t)]/(2mt) \to p_{kl}$, and we will show this now. Indeed, since $\mathbb{E}[M_{k,l}(t)]/(2mt) \to p_{kl}$ and $\mathbb{E}[N_k(t)]/t \to p_k$, we arrive at the claim that, for all k, l with $k \vee l \geq m+1$,

$$\begin{aligned}&\lim_{t\to\infty} \mathbb{E}[M_{kl}(t+1)] - \mathbb{E}[M_{kl}(t)]\\ &\quad = 2m^2 \frac{(k\vee l)-1+\delta}{2m+\delta} p_{k-1} \mathbb{1}_{\{k\wedge l=m\}} + 2m^2 \frac{k-1+\delta}{2m+\delta} p_{k-1,l} + 2m^2 \frac{l-1+\delta}{2m+\delta} p_{k,l-1} - 2m^2 \frac{k+l+2\delta}{2m+\delta} p_{k,l}.\end{aligned}$$

Since $\lim_{t\to\infty} \mathbb{E}[M_{kl}(t)]/(2mt) = p_{kl}$, we must therefore have that $\lim_{t\to\infty} \mathbb{E}[M_{kl}(t+1)] - \mathbb{E}[M_{kl}(t)] = 2mp_{kl}$, so that

$$p_{kl} = m \frac{(k\vee l)-1+\delta}{2m+\delta} p_{k\vee l-1} \mathbb{1}_{\{k\wedge l=m\}} + m \frac{k-1+\delta}{2m+\delta} p_{k-1,l} + m \frac{l-1+\delta}{2m+\delta} p_{k,l-1} - m \frac{k+l+2\delta}{2m+\delta} p_{k,l},$$

and

$$(k+l+2+2\delta+\delta/m)p_{kl} = ((k\vee l)-1+\delta)p_{k\vee l-1}\mathbb{1}_{\{k\wedge l=m\}} + (k-1+\delta)p_{k-1,l} + (l-1+\delta)p_{k,l-1}. \quad (4.24)$$

This is equivalent to [14, (12)]. This can be worked out to yield

$$\begin{aligned}p_{kl} = &\sum_{j=m+1}^{k} \binom{k+l-j-m}{k-j} \frac{\Gamma(k+\delta)}{\Gamma(j-1+\delta)} \frac{\Gamma(l+\delta)}{\Gamma(m+\delta)} \frac{\Gamma(j+k+2+2\delta+\delta/m)}{\Gamma(l+k+3+2\delta+\delta/m)} p_{j-1}\\ &+ \sum_{j=m+1}^{l} \binom{k+l-j-m}{l-j} \frac{\Gamma(k+\delta)}{\Gamma(j-1+\delta)} \frac{\Gamma(l+\delta)}{\Gamma(m+\delta)} \frac{\Gamma(j+l+2+2\delta+\delta/m)}{\Gamma(l+k+3+2\delta+\delta/m)} p_{j-1}.\end{aligned}$$

Substituting (4.8), we arrive at (4.23).

The computation to go from (4.24) to (4.11) is performed in [14, (12)], and applies verbatim. □

4.4 Asymptotically random Pearson's coefficient: collection of complete bipartite graphs

In this section, we present an example where $\rho(G_n)$ in (3.4) converges to a random variable when the number of vertices tends to infinity. For $|V_n| = n$, under the assumptions of Theorem 3.1, we have

$$\sum_{(u,v)\in E_n'} D_u D_v \leq \max_{v\in V_n} d_v \sum_{(u,v)\in E_n'} D_u = \max_{v\in V_n} D_v \Big(\sum_{v\in V_n} D_v^2 \Big) \leq C^2 n^{1/\gamma + (2/\gamma \vee 1)}, \quad (4.25)$$

$$\sum_{(u,v)\in E_n'} D_u D_v \geq \max_{v\in V_n} D_v \geq c n^{1/\gamma}, \quad (4.26)$$

$$\sum_{(u,v)\in E_n'} D_u D_v \geq \sum_{v\in V_n} D_v^2 \geq c n^{2/\gamma \vee 1}. \quad (4.27)$$

Further, from the proof of Theorem 3.1, we know that

$$\sum_{v\in V_n} D_v^3 \geq (\max_{v\in V_n} D_v)^3 \geq c^3 n^{3/\gamma}, \quad (4.28)$$

and

$$\frac{1}{|E_n'|} \Big(\sum_{v\in V_n} D_v^2 \Big)^2 \leq (C^2/c) n^{(4/\gamma - 1)\vee 1}, \quad (4.29)$$

where we see that (4.29) is vanishing compared to (4.28). The convergence of (3.4) to a random variable can only take place if the crossproducts on the left-hand side of (4.25 – 4.27) are of the same

order of magnitude as the left-hand side of (4.28). As we see from the above, this is possible for $\gamma \in (1,3)$.

Below we present an example where $\rho(G_n)$ indeed converges to a random variable. However, due to slow convergence, a substantially larger computational capacity is needed in order to approximate the limiting distribution.

Take $((X_i,Y_i))_{i=1}^n$ to be an i.i.d. sample of integer random variables as in (2.10), where $\alpha_1 = \alpha_2 = \beta_1 = b$, $\beta_2 = ab$ for some $b > 0$ and $a > 1$. Then, for $i = 1,\dots,n$, we create a complete bipartite graph of X_i and Y_i vertices, respectively. These n complete bipartite graphs are not connected to one another. We denote such a collection of n bipartite graphs by G_n. The graph G_n has $|V_n| = \sum_{i=1}^n (X_i + Y_i)$ vertices and $|E_n'| = 2\sum_{i=1}^n X_i Y_i$ directed edges. Further, if D_v denotes the random degree of vertex v, then we obtain

$$\sum_{v\in V_n} D_v^p = \sum_{i=1}^n (X_i^p Y_i + Y_i^p X_i), \qquad \sum_{(u,v)\in E_n'} D_u D_v = 2\sum_{i=1}^n (X_i Y_i)^2.$$

Assume that the ξ_j's in (2.10) satisfy (2.11) with $\gamma \in (2,4)$, so that $\mathbb{E}[\xi^2] < \infty$, but $\mathbb{E}[\xi^4] = \infty$. As a result, $|E_n'|/n \xrightarrow{\mathbb{P}} 2\mathbb{E}[XY] < \infty$ and $\frac{1}{n}\sum_{v\in V} D_v^2 \xrightarrow{\mathbb{P}} \mathbb{E}[XY(X+Y)] < \infty$ when $\gamma \in (3,4)$, while, for $\gamma \in (2,3)$,

$$n^{-3/\gamma} \sum_{v\in V} D_v^2 = n^{-3/\gamma} \sum_{i=1}^n (X_i^2 Y_i + Y_i^2 X_i) \xrightarrow{d} Z, \tag{4.30}$$

for some random variable Z. [For $\gamma = 3$, this sum grows as a slowly varying function in n, but this case is very similar and will thus be omitted.] Further,

$$n^{-4/\gamma} b^{-4} \sum_{i=1}^n (X_i^3 Y_i + Y_i^3 X_i) \xrightarrow{d} (a^3+a)Z_1 + 2Z_2, \qquad n^{-4/\gamma} b^{-4} \sum_{i=1}^N (X_i Y_i)^2 \xrightarrow{d} a^2 Z_1 + Z_2,$$

where Z_1 and Z_2 and two independent stable distributions with parameter $\gamma/4$. Therefore, using (3.4) and the fact that $4/\gamma > (6/\gamma - 1) \wedge 1$, we arrive at

$$\rho(G_n) \xrightarrow{d} \frac{2a^2 Z_1 + 2Z_2}{(a+a^3)Z_1 + 2Z_2}, \text{ as } n \to \infty.$$

which is a proper random variable taking values in $(2a/(1+a^2), 1)$.

For convergence of the rank correlation, we note that

$$\mathbb{P}(X_n = k, Y_n = l) \to \mathbb{P}(X = k, Y = l) = \frac{kl}{\mathbb{E}[X_1 Y_1]} \mathbb{P}(X_1 = k, Y_1 = l),$$

where we recall that (X_1, Y_1) is as in (2.10), while (X, Y) are the degrees at either side of a uniformly chosen edge. Thus, convergence of the rank correlation follows from Theorem 3.2(a).

5 Numerical results

In this section, we present numerical examples that illustrate our results.

5.1 Numerical results for configuration models and preferential attachment model

We have generated random graphs of different sizes using the configuration model in Section 4.1, the configuration model with intermediate vertices in Section 4.2, and the Preferential Attachment model (PAM) in Section 4.3. For the undirected preferential attachment model, we use the basic version with $m = 1$ and $\delta = 0$, which implies $\gamma = 2$. In both configuration models (without and with intermediate vertices) we generate the degree sequences by rounding up i.i.d. values of a continuous

random variable η with Pareto distribution: $\mathbb{P}(\eta > x) = 4x^{-2}$, $x > 2$. The exponent $\gamma = 2$ is chosen for a fair comparison to PAM, and all degrees are at least three for the strongest disassortativity in the model with intermediate in the model with intermediate vertices, see (4.1). In case of the configuration graph in Section 4.1, we consider two versions: the original model with self-loops and double edges present, and the model where self-loops and double-edges are removed. The rank correlation coefficient $\rho^{\text{rank}}(G)$ is computed as in (3.2). The results are presented in Table 3.

Model	Characteristic	n			
		10^2	10^3	10^4	10^5
Configuration model with self-loops and double edges	$\mathbb{E}_N(\rho(G_n))$	-0.0070	-0.0018	-0.0011	0.0006
	$\sigma_N(\rho(G_n))$	0.0735	0.0221	0.0077	0.0017
	$\mathbb{E}_N(\rho^{\text{rank}}(G_n))$	0.0056	-0.0098	-0.0036	0.0005
	$\sigma_N(\rho^{\text{rank}}(G_n))$	0.0504	0.0150	0.0046	0.0019
Configuration model without self-loops and double edges	$\mathbb{E}_N(\rho(G_n))$	-0.0713	-0.0226	-0.0150	-0.0032
	$\sigma_N(\rho(G_n))$	0.0546	0.0188	0.0092	0.0029
	$\mathbb{E}_N(\rho^{\text{rank}}(G_n))$	-0.0409	-0.0094	-0.0032	-0.0006
	$\sigma_N(\rho^{\text{rank}}(G_n))$	0.0700	0.0201	0.0083	0.0021
Configuration model with intermediate vertices	$\mathbb{E}_N(\rho(\bar{G}_n))$	-0.2804	-0.1346	-0.0572	-0.0291
	$\sigma_N(\rho(\bar{G}_n))$	0.0742	0.0517	0.0279	0.0147
	$\mathbb{E}_N(\rho^{\text{rank}}(\bar{G}_n))$	-0.7523	-0.7498	-0.7498	-0.7500
	$\sigma_N(\rho^{\text{rank}}(\bar{G}_n))$	0.0081	0.0025	0.0008	0.0003
Preferential attachment	$\mathbb{E}_N(\rho(G_n))$	-0.2682	-0.1282	-0.0608	-0.0272
	$\sigma_N(\rho(G_n))$	0.0575	0.0271	0.0132	0.0064
	$\mathbb{E}_N(\rho^{\text{rank}}(G_n))$	-0.4347	-0.4263	-0.4288	-0.4289
	$\sigma_N(\rho^{\text{rank}}(G_n))$	0.0627	0.0272	0.0065	0.0020

Table 3: Estimated mean and standard deviation of $\rho(G_n)$ and $\rho^{\text{rank}}(G_n)$ obtained from 20 realizations of G_n for random graph models in Sections 4.1–4.3.

The results for the configuration model with intermediate vertices confirm our findings in Section 4.2: Pearson's coefficient converges to zero, while Spearman's rho quickly converges to -0.75 revealing the strong negative dependence. For the PAM, Pearson's coefficient converges to zero, as indicated in Theorem 3.1, while Spearman's rank correlation clearly indicates a negative dependence. This can be understood by noting that the majority of edges of vertices with high degrees, which are old vertices, come from vertices which are added late in the graph growth process and thus have small degree. On the other hand, by the growth mechanism of the PAM, vertices with low degree are more likely to be connected to vertices having high degree, which indeed suggests negative degree-degree dependencies.

We emphasize that under given model assumptions, the graphs of different sizes have been constructed by the same algorithm. Thus, their mixing patterns are exactly the same. As we predicted, the Pearson correlation coefficient fails to reflect the intrinsic properties of the model because its absolute value decreases with the graph size, and converges to zero for all models. On the contrary, Spearman's rho consistently shows neutral mixing for the classical configuration model, moderately disassortative mixing for the Preferential Attachment graph, and strongly disassortative mixing for the configuration model with intermediate vertices.

5.2 Numerical results for collections of bipartite graphs

We next compute the degree-degree dependencies in the collection of bipartite graphs discussed in Section 4.4. In Table 4 we present numerical results for $\rho(G_n)$ and $\rho^{\text{rank}}(G_n)$. Here we choose $b = 1/2$, $a = 2$, ξ has a generalized Pareto distribution $\mathbb{P}(\xi > x) = (1 + (x-1)/2.8)^{-2.8}$, $x > 1$, and the degrees X and Y are obtained by rounding up the values in (2.10).

Note that in this model there is a genuine dependence between the correlation measure and the graph size. Indeed, if $n = 1$ then the assortativity coefficient equals -1 because nodes with

n	10^2	10^3	10^4	10^5
$\mathbb{E}_N(\rho(G_n))$	0.6554	0.7247	0.8042	0.8265
$\sigma_N(\rho(G_n))$	0.1145	0.1406	0.0689	0.0654
$\mathbb{E}_N(\rho^{\rm rank}(G_n))$	0.7575	0.7950	0.8526	0.8615
$\sigma_N(\rho^{\rm rank}(G_n))$	0.0735	0.1377	0.0218	0.0074

Table 4: Estimated mean and standard deviation of $\rho(G_n)$ and $\rho^{\rm rank}(G_n)$ for the collection of n complete bipartite graphs. The number of realizations for each graph size is 20.

larger degrees are connected to nodes with smaller degrees. However, when the graph size grows, the positive correlations start dominating because of the positive linear dependence between X and Y. We see that again the rank correlation captures the relation faster and gives consistent results with decreasing dispersion of values. Finally, Figure 2 shows the changes in the empirical distribution of $\rho(G_n)$ as n grows. It is clear that a part of the probability mass is spread over the interval $(0.8, 1)$.

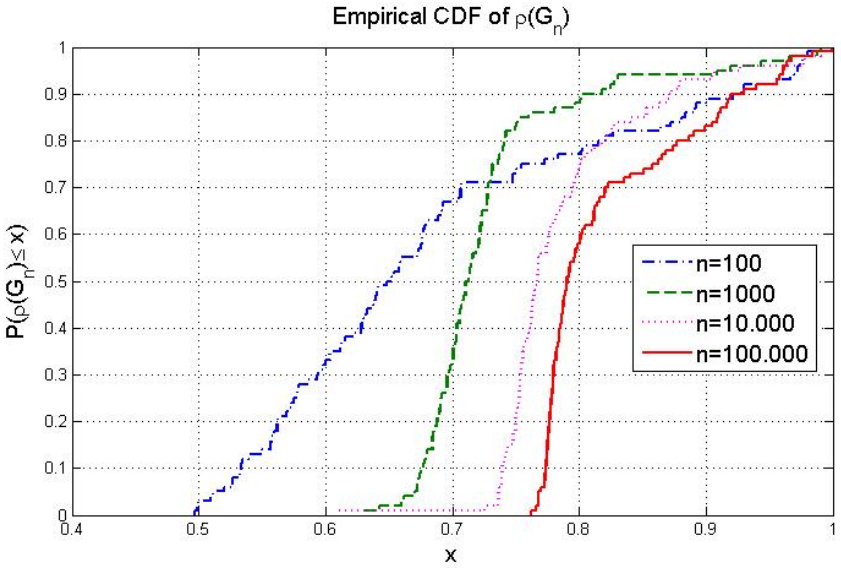


Figure 2: The empirical distribution function $\mathbb{P}(\rho(G_n) \le x)$ for 100 observed values of $\rho(G_n)$, where G_n is a collection of n complete bipartite graphs.

In the limit, $\rho(G_n)$ has a non-zero density on this interval. The difference between the crossproducts and the expectation squared in $\rho(G_n)$ is only of the order $n^{1-2/\gamma}$, which is about $n^{0.29}$ in our example, thus, the convergence is too slow to be observed at $n = 100.000$.

5.3 Web samples and social networks

For completeness, we present the numerical results for web samples and social networks from [24], see in Table 5. We used the compressed graph data from the Laboratory of Web Algorithms (LAW) at the Università degli studi di Milano [6, 5] with bvgraph MATLAB package [20]. The *stanford-cs* database [13] is a 2001 crawl that includes all pages in the cs.stanford.edu domain. In datasets (iv), (vii), (viii) we evaluate $\rho(G_n)$, $\rho^{\rm rank}(G_n)$ and $\rho^{-}(G_n)$ (see (3.5)) over 1000 random edges, and present the average over 10 such evaluations (in 10 samples of 1000 edges, the observed dispersion of the results was small).

We note that $\rho^{\rm rank}(G_n)$ here is an approximation of (3.2) computed as described in [24]: we define the random variables X and Y as the degrees on two ends of a random *undirected* edge in a graph (that is, here (u, v) and (v, u) represent the same edge); for each edge, when the observed degrees are a and b, we assign $[X = a, Y = b]$ or $[X = b, Y = a]$ with probability $1/2$; the ties are resolved randomly as in (3.2). The experiments on random graphs show that the values obtained by

nr	Dataset	Description	# nodes	# edges	max degree	$\rho(G_n)$	$\rho^{\text{rank}}(G_n)$	$\rho^{-}(G_n)$
(i)	stanford-cs	web domain	9,914	54,854	340	-.1656	-.1627	-.4648
(ii)	eu-2005	.eu web domain	862,664	5,477,938	68,963	-.0562	-.2525	-.0670
(iii)	uk@100,000	.uk web crawl	100,000	5,559,150	55,252	-.6536	-.5676	-1.117
(iv)	uk@1,000,000	.uk web crawl	1,000,000	77,123,940	403,441	-.0831	-.5620	-.0854
(v)	enron	e-mail exchange	69,244	506,898	1,634	-.1599	-.6827	-.1932
(vi)	dblp-2010	co-authorship	326,186	1,615,400	238	.3018	.2604	-.7736
(vii)	dblp-2011	co-authorship	986,324	6,707,236	979	.0842	.1351	-.2963
(viii)	hollywood-2009	co-starring	1,139,905	113,891,327	11,468	.3446	.4689	-0.6737

Table 5: (i)–(iv) Web crawls: nodes are web pages, and an (undirected) edge means that there is a hyperlink from one of the two pages to another; (iii),(iv) are breadth-first crawls around one page. (v) e-mail exchange by Enron employees (mostly part of the senior management): node are employees, and an edge means that an e-mail message was sent from one of the two employees to another. (vi), (vii) scientific collaboration networks extracted from the DBLP bibliography service: each vertex represents a scientist and an edge means a co-authorship of at least one article. (viii) vertices are actors, and two actors are connected by an edge if they appeared in the same movie.

this algorithm are very close to those computed by (3.2).

The most remarkable result here is obtained on the two .uk crawls (iii) and (iv). Here $\rho(G_n)$ is significantly smaller in magnitude on a larger crawl. Intuitively, mixing patterns should not depend on the crawl size. This is indeed confirmed by the value of Spearman's rho, which consistently shows strong negative correlations in both crawls. We could not observe a similar phenomenon so sharply in (vi) and (vii), probably because a larger co-authorship network incorporates articles from different areas of science, and the culture of scientific collaborations can vary greatly from one research field to another.

We also notice that, as predicted by our results, the small in magnitude values of $\rho^{-}(G_n)$ result in profound difference in magnitude between $\rho(G_n)$ and $\rho^{\text{rank}}(G_n)$. This is clearly seen in the data sets (ii), (iv) and (v). Again, (ii) and (iv) are the largest among the analyzed web crawls.

The observed behaviour of Pearson's coefficient is explained by the results proved in this paper in that $\rho(G_n)$ is strongly influenced by the large dispersion in the degree values, and particularly by the presence of hubs. The latter increases with graph size because of the scale-free phenomenon. As a result, $\rho(G_n)$ becomes smaller in magnitude when n increases, which makes it impossible to compare graphs of different sizes. In contrast, the *ranks* of the degrees are drawn from a uniform distribution on $[0, 1]$, scaled by the factor $|E'|$. Clearly, when a correlation coefficient is computed, the scaling factor cancels, and therefore Spearman's rho provides consistent results in the graphs of different sizes.

6 Discussion

In this paper, we have investigated dependency measures for power-law random variables. We have argued that Pearson's correlation coefficient, despite its appealing feature that it is always in $[-1, 1]$, is inappropriate to describe dependencies between heavy-tailed random variables. Indeed, the two main problems with the sample correlation coefficient are that (a) it can converge to a proper random variable when the sample size tends to infinity, indicating that it fluctuates tremendously as the sample size increases, and (b) that it is always asymptotically non-negative when dealing with non-negative random variables (even when these are obviously negatively dependent). In the context of random graphs, the first deficiency means that Pearson's coefficient can have a non-vanishing variance even when the size of the graph is huge, the second mistakenly suggests that there do not exist asymptotically disassortative scale-free graphs. We give proofs for the facts stated above, and illustrate the results using simulations.

Rank correlations are a special case of the broader concept of copulas that are widely used in

multivariate analysis, in particular in applications in mathematical finance and risk management. There is a heated discussion in this area about the adequacy and informativeness of such measures, see e.g. [32] and consequent reactions. There are several points of criticism. In particular, Spearman's rho uses rank transformation, which changes the observed values of the degrees. Then, first of all, what exactly does Spearman's rho tell us about the dependence between the original values? Second of all, no substantial justification exists for the rank transformation, besides its mathematical convenience. We thus do not claim that Spearman's rho is *the* solution to the problem. Nevertheless, compared to the Pearson's coefficient, Spearman's rho has a significant advantage that it is free from the undesirable size-dependency, and converges to meaningful value in the infinite volume limit.

We note that Spearman's rho has computational complexity $O(n \log(n))$ because the values of the random variables must be ranked first. Pearson's correlation coefficient is easier to evaluate because it uses the values of the degrees directly, and has computational complexity $O(n)$. Efficient methods for computing Spearman's rho in large graphs is an interesting topic for future research.

Raising the discussion to a higher level, random variables X and Y are positively dependent when a large realization of X typically implies a large realization of Y. A strong form of this notion is when $\mathbb{P}(X > x, Y > y) \geq \mathbb{P}(X > x)\mathbb{P}(Y > y)$ for every $x, y \in \mathbb{R}$, but for many purposes this notion is too restrictive. The covariance for non-negative random variables is obtained by integrating the above inequality over $x, y \geq 0$, so that it is true for 'typical' values of x, y. In many cases, however, we are particularly interested in certain values of x, y. Another class of methods for measuring rank correlations is based on the *angular measure*, a notion originating in the theory of multivariate extremes, for which the above inequality is investigated for *large* x and y, so that it describes the *tail dependence* for a random vector (X, Y), that is, the dependence between extremely large values of X and Y, see e.g. [43]. Such tail dependence is characterized by an probability-like measure, or, the angular measure, on $[0, 1]$. Informally, a concentration of the angular measure around the points 0 and 1 indicates independence of large values, while concentration around some other number $a \in (0, 1)$ suggests that a certain fraction of large values of Y comes together with large values of X. In [45, 46] a first attempt was made to compute the angular measure between in-degree of a node and its importance measured by the Google PageRank algorithm. Strikingly, completely different dependence structures were discovered in Wikipedia (independence), Preferential Attachment networks (complete dependence) and the Web (intermediate case).

Acknowledgments. We thank Yana Volkovich for the code generating a preferential attachment graph and Marie Albenque for a counter example that shows that negative dependence of (X, Y) in does not follow from negative dependence of (X_1, Y_1). We further thank Juli Komjáthy for her careful reading of the paper, which has tremendously improved the presentation and has corrected several typos and errors. This article is also the result of joint research in the 3TU Centre of Competence NIRICT (Netherlands Institute for Research on ICT) within the Federation of Three Universities of Technology in The Netherlands. The work of RvdH was supported in part by the Netherlands Organisation for Scientific Research (NWO). The work of NL is partially supported by the EU-FET Open grant NADINE (288956).

References

[1] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

[2] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967.

[3] E.A. Bender and E.R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978.

[4] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*, volume **27** of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1989.

[5] P. Boldi, M. Rosa, M. Santini, and S. Vigna. Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks. In *Proceedings of the 20th International World Wide Web Conference (WWW 2011)*. ACM Press, 2011.

[6] P. Boldi and S. Vigna. The WebGraph framework I: Compression techniques. In *Proceedings of the 13th International World Wide Web Conference (WWW 2004)*, pages 595–601, Manhattan, USA, 2004. ACM Press.

[7] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European J. Combin.*, 1(4):311–316, 1980.

[8] B. Bollobás. *Random Graphs*, volume 73. Cambridge Univ Pr, 2001.

[9] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. *Random Structures and Algorithms*, 18:279–290, 2001.

[10] C.B. Borkowf. Computing the nonnull asymptotic variance and the asymptotic relative efficiency of spearman's rank correlation. *Computational statistics & data analysis*, 39(3):271–286, 2002.

[11] D. Braha and Y. Bar-Yam. The statistical mechanics of complex product development: Empirical and analytical results. *Management Science*, 53(7):1127–1145, 2007.

[12] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Statac, A. Tomkins, and J. Wiener. Graph structure in the Web. *Computer Networks*, 33:309–320, 2000.

[13] P.G. Constantine and D.F. Gleich. Using polynomial chaos to compute the influence of multiple random surfers in the PageRank model. In Anthony Bonato and Fan Chung Graham, editors, *Proceedings of the 5th Workshop on Algorithms and Models for the Web Graph (WAW2007)*, volume 4863 of *Lecture Notes in Computer Science*, pages 82–95. Springer, 2007.

[14] S.N. Dorogovtsev, A.L. Ferreira, A.V. Goltsev, and J.F.F. Mendes. Zero Pearson coefficient for strongly correlated growing trees. *Physical Review E*, 81(3):031135, 2010.

[15] J.C. Doyle, D.L. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger. The robust yet fragile nature of the Internet. *PNAS*, 102(41):14497–14502, 2005.

[16] R. Durrett. *Random graph dynamics*. Cambridge University Press, 2007.

[17] V.M. Eguiluz and K. Klemm. Epidemic threshold in structured scale-free networks. *Physical Review Letters*, 89(10):108701, 2002.

[18] S. Eubank, H. Guclu, V.S. Anil Kumar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

[19] S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer. On local estimations of pagerank: A mean field approach. *Internet Mathematics*, 4(2-3):245–266, 2007.

[20] D.F. Gleich, A.P. Gray, C. Greif, and T. Lau. An inner-outer iteration for computing pagerank. *SIAM Journal on Scientific Computing*, 32(1):349, 2010.

[21] B. V. Gnedenko and A. N. Kolmogorov. *Limit distributions for sums of independent random variables*. Translated from the Russian, annotated, and revised by K. L. Chung. With appendices by J. L. Doob and P. L. Hsu. Revised edition. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills., Ont., (1968).

[22] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[23] R. van der Hofstad. Random graphs and complex networks, 2013. Available at `http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf`.

[24] R. van der Hofstad and N. Litvak. Uncovering disassortativity in large scale-free networks. *Physical Review E*, **87**(2):022801, 2013.

[25] S. Janson. The probability that a random multigraph is simple. *Combinatorics, Probability and Computing*, 18(1-2):205–225, 2009.

[26] M. Kendall. *Rank Correlation Methods*. Charles Griffin & Company, 1975.

[27] M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[28] R. Kumar and S. Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th international conference on World wide web*, pages 571–580. ACM, 2010.

[29] L. Li, D.L. Alderson, J.C. Doyle, and W. Willinger. Towards a theory of scale-free graphs: definition, properties, and implications. *Internet Mathematics*, 2(4):431–523, 2005.

[30] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat. Systematic topology analysis and generation using degree correlations. *ACM SIGCOMM Computer Communication Review*, 36(4):135–146, 2006.

[31] M. Mesfioui and A. Tajar. On the properties of some nonparametric concordance measures in the discrete case. *Nonparametric Statistics*, 17(5):541–554, 2005.

[32] T. Mikosch. Copulas: Tales and facts. *Extremes*, 9(1):3–20, 2006.

[33] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[34] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995.

[35] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. *Combinatorics Probability and Computing*, 7(3):295–305, 1998.

[36] J. Nevslehová. On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis*, 98(3):544–567, 2007.

[37] M. Newman. *Networks: an introduction*. Oxford University Press, Inc., 2010.

[38] M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.

[39] M.E.J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[40] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[41] M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.

[42] M. Raschke, M. Schläpfer, and R. Nibali. Measuring degree-degree association in networks. *Physical Review E*, 82(3):037102, 2010.

[43] S.I. Resnick. *Heavy-tail phenomena*. Springer, 2007.

[44] C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[45] Y. Volkovich, N. Litvak, and B. Zwart. Measuring extremal dependencies in Web graphs. In *WWW' 08: Proceedings of the 17th international conference on World Wide Web*, pages 1113–1114. ACM Press New York, NY, 2008.

[46] Y. Volkovich, N. Litvak, and B. Zwart. Extremal dependencies and rank correlations in power law networks. In J. Zhou, O. Akan, and P. Bellavista et al., editors, *Complex Sciences*, volume 5 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 1642–1653. Springer Berlin Heidelberg, 2009.

Quick Detection of Nodes with Large Degrees

Konstantin Avrachenkov, Nelly Litvak, Marina Sokol, Don Towsley

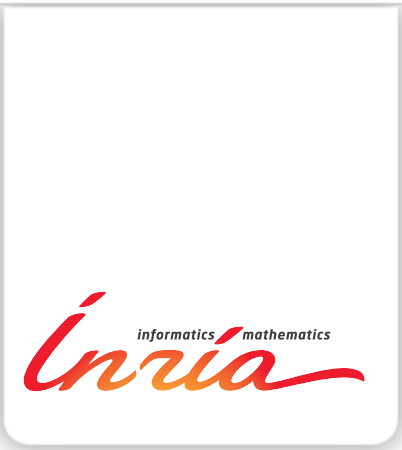
RESEARCH REPORT

N° 7881

February 2012

Project-Team Maestro

ISSN 0249-6399 ISRN INRIA/RR--7881--FR+ENG



Quick Detection of Nodes with Large Degrees

Konstantin Avrachenkov*, Nelly Litvak†, Marina Sokol‡,
Don Towsley§

Project-Team Maestro

Research Report n° 7881 — February 2012 — 13 pages

Abstract: Our goal is to quickly find top k lists of nodes with the largest degrees in large complex networks. If the adjacency list of the network is known (not often the case in complex networks), a deterministic algorithm to find a node with the largest degree requires an average complexity of $O(n)$, where n is the number of nodes in the network. Even this modest complexity can be very high for large complex networks. We propose to use the random walk based method. We show theoretically and by numerical experiments that for large networks the random walk method finds good quality top lists of nodes with high probability and with computational savings of orders of magnitude. We also propose stopping criteria for the random walk method which requires very little knowledge about the structure of the network.

Key-words: Complex networks, detection of nodes with the largest degrees, top k list, random walk, stopping criteria

* INRIA Sophia Antipolis, France, K.Avrachenkov@sophia.inria.fr
† University of Twente, the Netherlands, N.Litvak@utwente.nl
‡ INRIA Sophia Antipolis, France, Marina.Sokol@sophia.inria.fr
§ University of Massachusetts Amherst, USA, towsley@cs.umass.edu

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Détection Rapide de Noeuds à Degrés Élevés

Résumé : Notre objectif est de trouver rapidement dans les grands réseaux complexes top k listes de noeuds avec les plus grands degrés. Si la liste d'adjacence du réseau est connu (pas souvent le cas dans les réseaux complexes), un algorithme déterministe pour trouver un noeud avec le plus grand degré nécessite une complexité moyenne de $\mathrm{O}(n)$, où n est le nombre de noeuds dans le réseau. Même cette complexité modeste peut être très élevé pour les grands réseaux complexes. Nous proposons d'utiliser une méthode basé sur le marche aléatoire. Nous montrons théoriquement et par expérimentations numériques que pour les grands réseaux la méthode de marche aléatoire trouve top k listes de bonne qualité avec une forte probabilité de réussite et avec des économies de calcul de plusieurs ordres de grandeur. Nous proposons également des critères d'arrêt pour la méthode de marche aléatoire qui ne nécessite pas de connaissance de la structure du réseau.

Mots-clés : réseaux complexes, détection de noeuds avec les plus grands degrés, top k liste, marche aléatoire, critères d'arrêt

1 Introduction

We are interested in quickly detecting nodes with large degrees in very large networks. Firstly, node degree is one of centrality measures used for the analysis of complex networks. Secondly, large degree nodes can serve as proxies for central nodes corresponding to the other centrality measures as betweenness centrality or closeness centrality [8, 9]. In the present work we restrict ourself to undirected networks or symmetrized versions of directed networks. In particular, this assumption is well justified in social networks. Typically, friendship or acquaintance is a symmetric relation. If the adjacency list of the network is known (not often the case in complex networks), the straightforward method that comes to mind is to use one of the standard sorting algorithms like Quicksort or Heapsort. However, even their modest average complexity, $O(n \log(n))$, can be very high for very large complex networks. In the present work we suggest using random walk based methods for detecting a small number of nodes with the largest degree. The main idea is that the random walk very quickly comes across large degree nodes. In our numerical experiments random walks outperform the standard sorting procedures by orders of magnitude in terms of computational complexity. For instance, in our experiments with the web graph of the UK domain (about 18 500 000 nodes) the random walk method spends on average only about 5 400 steps to detect the largest degree node. Potential memory savings are also significant since the method does not require knowledge of the entire network. In many practical applications we do not need a complete ordering of the nodes and even can tolerate some errors in the top list of nodes. We observe that the random walk method obtains many nodes in the top list correctly and even those nodes that are erroneously placed in the top list have large degrees. Therefore, as typically happens in randomized algorithms [12, 13], we trade off exact results for very good approximate results or for exact results with high probability and gain significantly in computational efficiency.

The paper is organized as follows: in the next section we introduce our basic random walk with uniform jumps and demonstrate that it is able to quickly find large degree nodes. Then, in Section 3 using configuration model we provide an estimate for the necessary number of steps for the random walk. In Section 4 we propose stopping criteria that use very little information about the network. In Section 5 we show the benefits of allowing few erroneous elements in the top k list. Finally, we conclude the paper in Section 6.

2 Random walk with uniform jumps

Let us consider a random walk with uniform jumps which serves as a basic algorithm for quick detection of large degree nodes. The random walk with uniform jumps is described by the following transition probabilities [1]

$$p_{ij} = \begin{cases} \frac{\alpha/n+1}{d_i+\alpha}, & \text{if } i \text{ has a link to } j, \\ \frac{\alpha/n}{d_i+\alpha}, & \text{if } i \text{ does not have a link to } j, \end{cases} \tag{1}$$

where d_i is the degree of node i. The random walk with uniform jumps can be regarded as a random walk on a modified graph where all the nodes in the graph are connected by artificial edges with a weight α/n. The parameter α controls the rate of jumps. Introduction of jumps helps in a number of ways. As was shown in [1], it reduces the mixing time to stationarity. It also solves a problem encountered by a random walk on a graph consisting of two or more components, namely the inability to visit all nodes. The random walk with jumps also reduces the variance of the network function estimator [1]. This random walk resembles the PageRank random walk. However, unlike the PageRank random walk, the introduced random walk is reversible. One

important consequence of the reversibility of the random walk is that its stationary distribution is given by a simple formula

$$\pi_i(\alpha) = \frac{d_i + \alpha}{2|E| + n\alpha} \qquad \forall i \in V, \tag{2}$$

from which the stationary distribution of the original random walk can easily be retrieved. We observe that the modification preserves the order of the nodes' degrees, which is particularly important for our application.

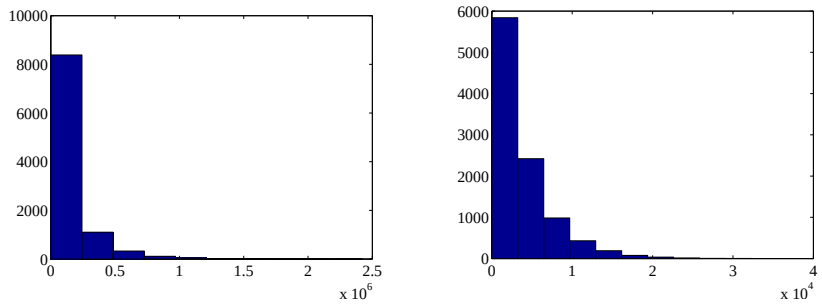
We illustrate on several network examples how the random walk helps us quickly detect large degree nodes. We consider as examples one synthetic network generated by the preferential attachment rule and two natural large networks. The Preferential Attachment (PA) network combines 100 000 nodes. It has been generated according to the generalized preferential attachment mechanism [6]. The average degree of the PA network is two and the power law exponent is 2.5. The first natural example is the symmetrized web graph of the whole UK domain crawled in 2002 [4]. The UK network has 18 520 486 nodes and its average degree is 28.6. The second natural example is the network of co-authorships of DBLP [5]. Each node represents an author and each link represents a co-authorship of at least one article. The DBLP network has 986 324 nodes and its average degree is 6.8.

We carry out the following experiment: we initialize the random walk (1) at a node chosen according to the uniform distribution and continue the random walk until we hit the largest degree node. The largest degrees for the PA, UK and DBLP networks are 138, 194 955, and 979, respectively. For the PA network we have made 10 000 experiments and for the UK and DBLP networks we performed 1 000 experiments (these networks were too large to perform more experiments).

In Figue 1 we plot the histograms of hitting times for the PA network. The first remarkable observation is that when $\alpha = 0$ (no restart) the average hitting time, which is equal to 123 000, is nearly three orders of magnitude larger than 3 720, the hitting time when $\alpha = 2$. The second remarkable observation is that 3 720 is not too far from the value

$$1/\pi_{max}(\alpha) = (2|E| + n\alpha)/(d_{max} + \alpha) = 2\,857,$$

which corresponds to the average return time to the largest degree node in the random walk with jumps.



(a) $\alpha = 0$ (b) $\alpha = 2$

Figure 1: Histograms of hitting times in the PA network.

We were not able to collect a representative number of experiments for the UK and DBLP networks when $\alpha = 0$. The reason for this is that the random walk gets stuck either in disconected

or weakly connected components of the networks. For the UK network we were able to make 1 000 experiments with $\alpha = 0.001$ and obtain the average hitting time 30 750. Whereas if we take $\alpha = 28.6$ for the UK network, we obtain the average hitting time 5 800. Note that the expected return time to the largest degree node in the UK network is given by

$$1/\pi_{max}(\alpha) = (2|E| + n\alpha)/(d_{max} + \alpha) = 5\,432.$$

For the DBLP graph we conducted 1 000 experiments with $\alpha = 0.00001$ and obtained an average hitting time of 41 131. Whereas if we take $\alpha = 6.8$, we obtain an average hitting time of 14 200. The expected return time to the largest degree node in the DBLP network is given by

$$1/\pi_{max}(\alpha) = (2|E| + n\alpha)/(d_{max} + \alpha) = 13\,607.$$

The two natural network examples confirm our guess that the average hitting time for the largest degree node is fairly close to the average return time to the largest degree node. Let us also confirm our guess with asymptotic analysis.

Theorem 1 *Without loss of generality, index the nodes such that node 1 has the largest degree, $(1, i) \in E, i = 2, ..., s, s = d_1 + 1$, and let ν denote the initial distribution of the random walk with jumps. Then, the expected hitting time to node 1 starting from any initial distribution ν is given by*

$$E_\nu[T_1] = \frac{\sum_{i=2}^n d_i + (n-1)\alpha}{d_1 + 2\alpha(1 - 1/n)} + o\left(\min_{i=2,...,s}\{(d_i + \alpha), n\}\right), \tag{3}$$

Proof: The expected hitting time from distribution ν to node 1 is given by the formula

$$E_\nu[T_1] = \nu[I - P_{-1}]^{-1}\underline{1}, \tag{4}$$

where P_{-1} is a taboo probability matrix (i.e., matrix P with the 1-st row and 1-st column removed). The matrix P_{-1} is substochastic but is very close to stochastic. Let us represent it as a stochastic matrix minus some perturbation term:

$$P_{-1} = \tilde{P} - \varepsilon Q = \tilde{P} - \begin{bmatrix} \frac{1+2\alpha/n}{d_2+\alpha} & 0 & & & & 0 \\ 0 & \ddots & & & & \\ & & \frac{1+2\alpha/n}{d_s+\alpha} & & & \\ & & & \frac{2\alpha/n}{d_{s+1}+\alpha} & & \\ & & & & \ddots & 0 \\ 0 & & & & 0 & \frac{2\alpha/n}{d_n+\alpha} \end{bmatrix}$$

We add missing probability mass to the diagonal of $\tilde{P}$, which corresponds to an increase in the weights for self-loops. The matrix $\tilde{P}$ represents a reversible Markov chain with the stationary distribution

$$\tilde{\pi}_j = \frac{d_j + \alpha}{\sum_{i=2}^n d_i + (n-1)\alpha}.$$

Now we can use the following result from the perturbation theory (see Lemma 1 in [2]):

$$[I - \tilde{P} + \varepsilon Q]^{-1} = \frac{\underline{1}\tilde{\pi}}{\tilde{\pi}(\varepsilon Q)\underline{1}} + X_0 + \varepsilon X_1 + ... \,, \tag{5}$$

where $\tilde{\pi}$ is the stationary distribution of the stochastic matrix $\tilde{P}$. In our case, the quantity $\max_{i=2,\dots,s}\{1/(d_i+\alpha), 1/n\}$ will play the role of ε. We apply the series (5) to approximate the expected hitting time. Towards this goal, we calculate

$$\tilde{\pi}(\varepsilon Q)\underline{1} = \sum_{j=2}^{n} \tilde{\pi}_j \varepsilon q_{jj}$$

$$= \sum_{j=2}^{s} \frac{d_j+\alpha}{\sum_{i=2}^{n} d_i + (n-1)\alpha} \frac{1+2\alpha/n}{d_j+\alpha} + \sum_{j=s+1}^{n} \frac{d_j+\alpha}{\sum_{i=2}^{n} d_i + (n-1)\alpha} \frac{2\alpha/n}{d_j+\alpha}$$

$$= \frac{d_1(1+2\alpha/n) + (n-d_1-1)(2\alpha/n)}{\sum_{i=2}^{n} d_i + (n-1)\alpha} = \frac{d_1 + 2\alpha(1-1/n)}{\sum_{i=2}^{n} d_i + (n-1)\alpha}.$$

Observing that $\nu\underline{1}\tilde{\pi}\underline{1} = 1$, we obtain (3).

□

Indeed, the asymptotic expression (3) is very close to $(2|E| + n\alpha)/(d_1 + \alpha)$, which is the expected return time to node 1.

Based on the notion of the hitting time we propose an efficient method for quick detection of the top k list of largest degree nodes. The algorithm maintains a top k candidate list. Note that once one of the k nodes with the largest degrees appears in this candidate list, it remains there subsequently. Thus, we are interested in hitting events. We propose the following algorithm for detecting the top k list of largest degree nodes.

Algorithm 1 Random walk with jumps and candidate list

1. *Set k, α and m.*
2. *Execute a random walk step according to* (1).
3. *Check if the current node has a larger degree than one of the nodes in the current top k candidate list. If it is the case, insert the new node in the top-k candidate list and remove the worst node out of the list.*
4. *If the number of random walk steps is less than m, return to Step 2 of the algorithm. Stop, otherwise.*

The value of parameter α is not crucial. In our experiments, we have observed that as long as the value of α is neither too small nor not too big, the algorithm performs well. A good option for the choice of α is a value slightly smaller than the average node degree. Let us explain this choice by calculating a probability of jump in the steady state

$$\sum_{j=1}^{n} \pi_j(\alpha) \frac{\alpha}{d_j+\alpha} = \sum_{j=1}^{n} \frac{d_j+\alpha}{2|E|+n\alpha} \frac{\alpha}{d_j+\alpha} = \frac{n\alpha}{2|E|+n\alpha} = \frac{\alpha}{2|E|/n+\alpha}.$$

If α is equal to $2|E|/n$, the average degree, the random walk will jump in the steady state on average every two steps. Thus, if we set α to the average degree or to a slightly smaller value, on one hand the random walk will quickly converge to the steady state and on the other hand we will not sample too much from the uniform distribution.

The number of random walk steps, m, is a crucial parameter. Our experiments indicate that we obtain a top k list with many correct elements with high probability if we take the number of random walk steps to be twice or thrice as large as the expected hitting time of the nodes in the top k list. From Theorem 1 we know that the hitting time of the large degree node is related to the value of the node's degree. Thus, the problem of choosing m reduces to the problem of estimating the values of the largest degrees. We address this problem in the following section.

3 Estimating the largest degrees in the configuration network model

The estimations for the values of the largest degrees can be derived in the configuration network model [7] with a power law degree distribution. In some applications the knowledge of the power law parameters might be available to us. For instance, it is known that web graphs have power law degree distribution and we know typical ranges for the power law parameters.

We assume that the node degrees $D_1, \ldots, D_n$ are i.i.d. random variables with a power law distribution F and finite expectation $E[D]$. Let us determine the number of links contained in the top k nodes. Denote

$$F(x) = P[D \le x], \quad \bar{F}(x) = 1 - F(x), \quad x \ge 0.$$

Further let $D_{(1)} \ge \ldots \ge D_{(n)}$ be the order statistics of $D_1, \ldots, D_n$. Under the assumption that D_j's obey a power law, we use the results from the extreme value theory as presented in [11], to state that there exist sequences of constants (a_n) and (b_n) and a constant δ such that

$$\lim_{n\to\infty} n\bar{F}(a_n x + b_n) = (1 + \delta x)^{-1/\delta}. \tag{6}$$

This implies the following approximation for high quantiles of F, with exceedance probability close to zero [11]:

$$x_p \approx a_n \frac{(pn)^{-\delta} - 1}{\delta} + b_n.$$

For the jth largest degree, where $j = 2, \ldots, k$, the estimated exceedance probability equals $(j-1)/n$, and thus we can use the quantile $x_{(j-1)/n}$ to approximate the degree $D_{(j)}$ of this node:

$$D_{(j)} \approx a_n \frac{(j-1)^{-\delta} - 1}{\delta} + b_n. \tag{7}$$

The sequences (a_n) and (b_n) are easy to find for a given shape of the tail of F. Below we derive the corresponding results for the commonly accepted Pareto tail distribution of D, that is,

$$\bar{F}(t) = Cx^{-\gamma} \quad \text{for } x > x', \tag{8}$$

where $\gamma > 1$ and x' is a fixed sufficiently large number so that the power law degree distribution is observed for nodes with degree larger than x'. In that case we have

$$\lim_{n\to\infty} n\bar{F}(a_n x + b_n) = \lim_{n\to\infty} nC(a_n x + b_n)^{-\gamma} = \lim_{n\to\infty} (C^{-1/\gamma} n^{-1/\gamma} a_n x + C^{-1/\gamma} n^{-1/\gamma} b_n)^{-\gamma},$$

which directly gives (6) with

$$\delta = 1/\gamma, \quad a_n = \delta C^{\delta} n^{\delta}, \quad b_n = C^{\delta} n^{\delta}. \tag{9}$$

Substituting (9) into (7) we obtain the following prediction for $D_{(j)}$, $j = 2, \ldots, k$, in the case of the Pareto tail of the degree distribution:

$$D_{(j)} \approx n^{1/\gamma}[C^{1/\gamma}(j-1)^{-1/\gamma} - C^{1/\gamma} + 1]. \tag{10}$$

It remains to find an approximation for $D_{(1)}$, the maximal degree in the graph. From the extreme value theory it is well known that if $D_1, \ldots, D_n$ obey a power law then

$$\lim_{n\to\infty} P\left(\frac{D_{(1)} - b_n}{a_n} \le x\right) = H_\delta(x) = \exp(-(1 + \delta x)^{-1/\delta}),$$

where, for Pareto tail, a_n, b_n and δ are defined in (9). Thus, as an approximation for the maximal node degree we can choose $a_n x + b_n$ where x can be chosen as either an expectation, a median or a mode of $H_\delta(x)$. If we choose the mode, $((1+\delta)^{-\delta} - 1)/\delta$, then we obtain an approximation, which is smaller than the one for the 2nd largest degree. Further, the expectation $(\Gamma(1-\delta)-1)/\delta$ is very sensitive to the value of $\delta = 1/\gamma$, especially when γ is close to one, which is often the case in complex networks. Besides, the parameter γ is hard to estimate with high precision. Thus, we choose the median $(\log(2))^{-\delta} - 1)/\delta$, which yields

$$D_{(1)} \approx a_n \frac{(\log(2))^{-\delta} - 1}{\delta} + b_n = n^{1/\gamma}[C^{1/\gamma}(\log(2))^{-1/\gamma} - C^{1/\gamma} + 1]. \tag{11}$$

For instance, in the PA network $\gamma = 2.5$ and $C = 3.7$, which gives according to (11) $D_{(1)} \approx 127$. (This is a good prediction even though the PA network is not generated according to the configuration model. We also note that even though the extremum distribution in the preferential attachment model is different from that of the configuration model their ranges seem to be very close [10].) This in turn suggests that for the PA network m should be chosen in the range 6 000-18 000 if $\alpha = 2$. As we can see from Figure 2 this is indeed a good range for the number of random walk steps. In the UK network $\gamma = 1.7$ and $C = 90$, which gives $D_{(1)} \approx 82\,805$ and suggests a range of 20 000-30 000 for m if $\alpha = 28.6$. Figure 3 confirms that this is a good choice. The degree distribution of the DBLP network does not follow a power law so we cannot apply the above reasoning to it.

4 Stopping criteria

Suppose now that we do not have any information about the range for the largest k degrees. In this section we design stopping criteria that do not require knowledge about the structure of the network. As we shall see, knowledge of the order of magnitude of the average degree might help, but this knowledge is not imperative for a practical implementation of the algorithm.

Let us now assume that node j can be sampled independently with probability $\pi_j(\alpha)$ as in (2). There are at least two ways to achieve this practically. The first approach is to run the random walk for a significant number of steps until it reaches the stationary distribution. If one chooses α reasonably large, say the same order of magnitude as the average degree, then the mixing time becomes quite small [1] and we can be sure to reach the stationary distribution in a small number of steps. Then, the last step of a run of the random walk will produce an i.i.d. sample from a distribution very close to (2). The second approach is to run the random walk uninterruptedly, also with a significant value of α, and then perform Bernoulli sampling with probability q after a small initial transient phase. If q is not too large, we shall have nearly independent samples following the stationary distribution (2). In our experiment, $q \in [0.2, 0.5]$ gives good results when α has the same order of magnitude as the average degree.

We now estimate the probability of detecting correctly the top k list of nodes after m i.i.d. samples from (2). Denote by X_i the number of hits at node i after m i.i.d. samples. We note that if we use the second approach to generate i.i.d. samples, we spend approximately m/q steps of the random walk. We correctly detect the top k list with the probability given by the multinomial distribution

$$P[X_1 \geq 1, ..., X_k \geq 1] =$$

$$\sum_{i_1 \geq 1, ..., i_1 \geq 1} \frac{m!}{i_1! \cdots i_k! (m - i_1 - ... - i_k)!} \pi_1^{i_1} \cdots \pi_k^{i_k} (1 - \sum_{i=1}^{k} \pi_i)^{m - i_1 - ... - i_k}$$

but it is not feasible for any realistic computations. Therefore, we propose to use the Poisson approximation. Let Y_j, $j = 1, ..., n$ be independent Poisson random variables with means $\pi_j m$. That is, the random variable Y_j has the following probability mass function $P[Y_j = r] = e^{-m\pi_j}(m\pi_j)^r/r!$. It is convenient to work with the complementary event of not detecting correctly the top k list. Then, we have

$$P[\{X_1 = 0\} \cup ... \cup \{X_k = 0\}] \leq 2P[\{Y_1 = 0\} \cup ... \cup \{Y_k = 0\}]$$

$$= 2(1 - P[\{Y_1 \geq 1\} \cap ... \cap \{Y_k \geq 1\}]) = 2(1 - \prod_{j=1}^{k} P[\{Y_j \geq 1\}])$$

$$= 2(1 - \prod_{j=1}^{k}(1 - P[\{Y_j = 0\}])) = 2(1 - \prod_{j=1}^{k}(1 - e^{-m\pi_j})) =: a, \tag{12}$$

where the first inequality follows from [12, Thm 5.10]. In fact, in our numerical experiments we observed that the factor 2 in the first inequality is very conservative. For large values of m, the Poisson bound works very well as proper approximation.

For example, if we would like to obtain the top 10 list with at most 10% probability of error, we need to have on average 4.5 hits per each top element. This can be used to design the stopping criteria for our random walk algorithm. Let $\bar{a} \in (0, 1)$ be the admissible probability of an error in the top k list. Now the idea is to stop the algorithm after m steps when the estimated value of a for the first time is lower than the critical number $\bar{a}$. Clearly,

$$\hat{a}_m = 2(1 - \prod_{j=1}^{k}(1 - e^{-X_j}))$$

is the maximum likelihood estimator for a, so we would like to choose m such that $\hat{a}_m \leq \bar{a}$. The problem, however, is that we do not know which X_j's are the realisations of the number of visits to the top k nodes. Then let $X_{j_1}, ..., X_{j_k}$ be the number of hits to the current elements in the top k candidate list and consider the estimator

$$\hat{a}_{m,0} = 2(1 - \prod_{i=1}^{k}(1 - e^{-X_{j_i}})),$$

which is the maximum likelihood estimator of the quantity

$$2(1 - \prod_{i=1}^{k}(1 - e^{-m\pi_{j_i}})) \geq a.$$

(Here π_{j_i} is a stationary probability of the node with the score X_{j_i}, $i = 1, \ldots, k$). The estimator $\hat{a}_{m,0}$ is computed without knowledge of the top k nodes or their degrees, and it is an estimator of an upper bound of the estimated probability that there are errors in the top k list. This leads to the following stopping rule.

Stopping rule 0. *Stop at $m = m_0$, where*

$$m_0 = \arg\min\{m : \hat{a}_{m,0} \leq \bar{a}\}.$$

The above stopping criterion can be simplified even further to avoid computation of $\hat{a}_{m,0}$. Since

$$\hat{a}_{m,1} := 2(1 - (1 - e^{-X_{j_k}})^k) \geq \hat{a}_{m,0} \geq \hat{a},$$

where X_{j_k} is the number of hits of the worst element in the candidate list. The inequality $\hat{a}_m \leq \bar{a}$ is guaranteed if $\hat{a}_{m,1} \leq \bar{a}$. This leads to the following stopping rule for the random walk algorithm.

Stopping rule 1. *Compute* $x_0 = \arg\min\{x \in \mathbb{N} : (1 - e^{-x})^k \geq 1 - \bar{\alpha}/2.\}$ *Stop at*

$$m_1 = \arg\min\{m : X_{j_k} = x_0\}.$$

We have observed in our numerical experiments that we obtain the best trade off between the number of steps of the random walk and the accuracy if we take α around the average degree and the sampling probability q around 0.5. Specifically, if we take $\bar{a}/2 = 0.15$ ($x_0 = 4$) in Stopping rule 1 for top 10 list, we obtain 87% accuracy for an average of 47 000 random walk steps for the PA network; 92% accuracy for an average of 174 468 random walk steps for the DBLP network; and 94% accuracy for an average of 247 166 random walk steps for the UK network. We have averaged over 1000 experiments to obtain tight confidence intervals.

5 Relaxation of top k lists

In the stopping criteria of the previous section we have strived to detect all nodes in the top k list. This costs us a lot of steps of the random walk. We can significantly gain in performance by relaxing this strict requirement. For instance, we could just ask for list of k nodes that contains 80% of top k nodes [3]. This way we can take an advantage of a generic 80/20 rule that 80% of result can be achieved with 20% of effort.

Let us calculate the expected number of top k elements observed in the candidate list up to trial m. Define by X_j the number of times we have observed node j after m trials and

$$H_j = \begin{cases} 1, & \text{node } j \text{ has been observed at least once,} \\ 0, & \text{node } j \text{ has not been observed.} \end{cases}$$

Assuming we sample in i.i.d. fashion from the distribution (2), we can write

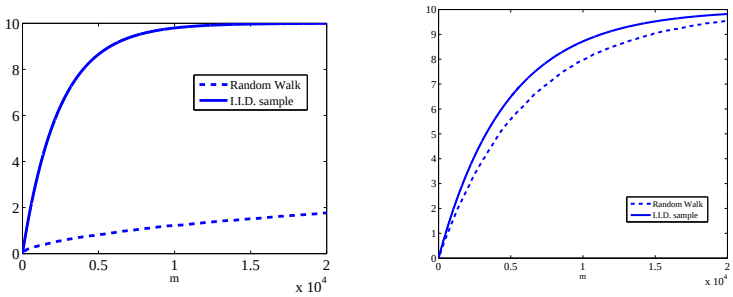
$$E[\sum_{j=1}^{k} H_j] = \sum_{j=1}^{k} E[H_j] = \sum_{j=1}^{k} P[X_j \geq 1] = \sum_{j=1}^{k} (1 - P[X_j = 0]) = \sum_{j=1}^{k} (1 - (1 - \pi_j)^m). \quad (13)$$

In Figure 2 we plot $E[\sum_{j=1}^{k} H_j]$ (the curve "I.I.D. sample") as a function of m for $k = 10$ for the PA network with $\alpha = 0$ and $\alpha = 2$. In Figure 3 we plot $E[\sum_{j=1}^{k} H_j]$ as a function of m for $k = 10$ for the UK network with $\alpha = 0.001$ and $\alpha = 28.6$. The results for the UK and DBLP networks are similar in spirit.

Here again we can use the Poisson approximation

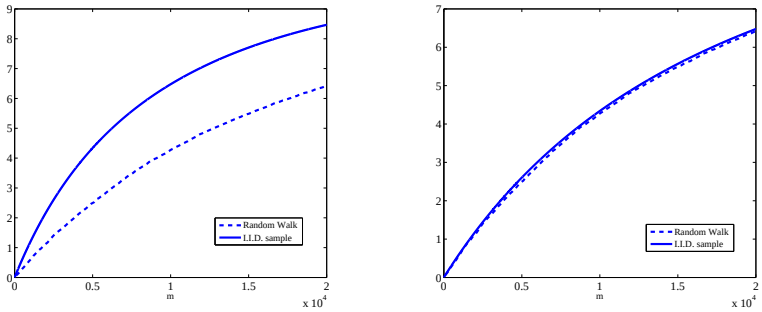
$$E[\sum_{j=1}^{k} H_j] \approx \sum_{j=1}^{k} (1 - e^{-m\pi_j}).$$

In fact, the Poisson approximation is so good that if we plot it on Figures 2 and 3, it nearly covers exactly the curves labeled "I.I.D. sample", which correspond to the exact formula (13). Similarly



(a) $\alpha = 0$ (b) $\alpha = 2$

Figure 2: Average number of correctly detected elements in top-10 for PA.



(a) $\alpha = 0.001$ (b) $\alpha = 28.6$

Figure 3: Average number of correctly detected elements in top-10 for UK.

to the previous section, we can propose stopping criteria based on the Poisson approximation. Denote

$$b_m = \sum_{i=1}^{k} (1 - e^{-X_{j_i}}).$$

Stopping rule 2. *Stop at $m = m_2$, where*

$$m_2 = \arg\min\{m : b_m \geq \bar{b}\}.$$

Now if we take $\bar{b} = 7$ in Stopping rule 3 for top-10 list, we obtain on average 8.89 correct elements for an average of 16 725 random walk steps for the PA network; we obtain on average 9.28 correct elements for an average of 66 860 random walk steps for the DBLP network; and we obtain on average 9.22 correct elements for an average of 65 802 random walk steps for the UK network. (We have averaged over 1000 experiments for each network.) This makes for the UK network the gain of more than two orders of magnitude in computational complexity with respect to the deterministic algorithm.

6 Conclusions and future research

We have proposed the random walk method with the candidate list for quick detection of largest degree nodes. We have also supplied stopping criteria which do not require knowledge of the graph structure. In the case of large networks, our algorithm finds top k list of largest degree nodes with few mistakes with the running time orders of magnitude faster than the deterministic sorting algorithm. In future research we plan to obtain estimates for the required number of steps for various types of complex networks.

References

[1] K. Avrachenkov, B. Ribeiro and D. Towsley, "Improving random walk estimation accuracy with uniform restarts", in Proceedings of WAW 2010, also Springer LNCS v.6516, pp.98-109, 2010.

[2] K. Avrachenkov, V. Borkar and D. Nemirovsky, "Quasi-stationary distributions as centrality measures for the giant strongly connected component of a reducible graph", *Journal of Comp. and Appl. Mathematics*, v.234, pp.3075-3090, 2010.

[3] K. Avrachenkov, N. Litvak, D. Nemirovsky, E. Smirnova and M. Sokol, "Quick detection of top-k personalized pagerank lists", in Proceedings of WAW 2011.

[4] P. Boldi and S. Vigna, "The WebGraph framework I: Compression techniques", in Proceedings of WWW 2004.

[5] P. Boldi, M. Rosa, M. Santini and S. Vigna, "Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks", in Proceedings of WWW 2011.

[6] S.N. Dorogovtsev, J.F.F. Mendes and A.N. Samukhin, "Structure of growing networks: Exact solution of the Barabasi-Albert model", *Phys. Rev. Lett.*, v.85, pp.4633-4636, 2000.

[7] R. van der Hofstad, *Random graphs and complex networks*, Lechture Notes, Available at *http://www.win.tue.nl/rhofstad/NotesRGCN.pdf*, 2009.

[8] Y. Lim, D.S. Menasche, B. Ribeiro, D. Towsley and P. Basu, "Online estimating the k central nodes of a network", in Proceedings of IEEE NSW 2011.

[9] A.S. Maiya and T.Y. Berger-Wolf, "Online sampling of high centrality individuals in social networks", in Proceedings of PAKDD 2010.

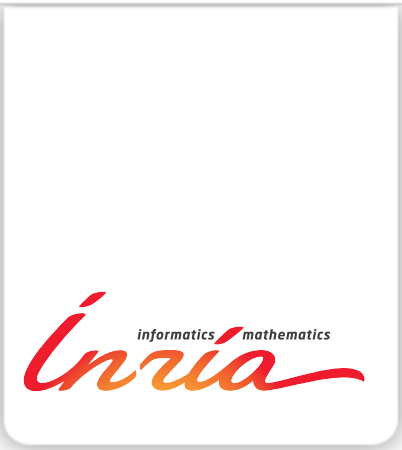
[10] A.A. Moreira, J.S. Andrade Jr. and L.A.N. Amaral, "Extremum statistics in scale-free network models", *Phys. Rev. Lett.*, v.89, 268703 4 pages, 2002.

[11] G. Matthys and J. Beirlant, "Estimating the extreme value index and high quantiles with exponential regression models", *Statistica Sinica*, v.13, no.3, pp.853-880, 2003.

[12] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*, Cambridge University Press, 2005.

[13] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, 1995.

Contents



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399

This figure "pagei.png" is available in "png" format from:

http://arxiv.org/ps/1202.3261v1

Alpha current flow betweenness centrality*

Konstantin Avrachenkov[1], Nelly Litvak[2], Vasily Medyanikov[3], Marina Sokol[1]

[1] Inria Sophia Antipolis, 2004 Route des Lucioles, Sophia-Antipolis, France
[2] University of Twente, P.O.Box 217, 7500AE, Enschede, The Netherlands
[3] St. Petersburg State University, 7-9, Universitetskaya nab., St. Petersburg, Russia

Abstract. A class of centrality measures called betweenness centralities reflects degree of participation of edges or nodes in communication between different parts of the network. The original shortest-path betweenness centrality is based on counting shortest paths which go through a node or an edge. One of shortcomings of the shortest-path betweenness centrality is that it ignores the paths that might be one or two steps longer than the shortest paths, while the edges on such paths can be important for communication processes in the network. To rectify this shortcoming a current flow betweenness centrality has been proposed. Similarly to the shortest path betwe has prohibitive complexity for large size networks. In the present work we propose two regularizations of the current flow betweenness centrality, α-current flow betweenness and truncated α-current flow betweenness, which can be computed fast and correlate well with the original current flow betweenness.

1 Introduction

A class of centrality measures called betweenness centralities reflects degree of participation of edges or nodes in communication between different parts of the network. The first notion of betweenness centrality was introduced by Freeman [8]. Let $s, t \in V$ be a pair of nodes in an undirected network $G = (V, E)$. We denote $|V| = n$, $|E| = m$, and let d_v be the degree of node v. Let $\sigma_{s,t}$ be the number of shortest paths connecting nodes s and t and denote $\sigma_{s,t}(e)$ the number of shortest path connecting nodes s and t passing through edge e. Then betweenness centrality of edge e is calculated as follows:

$$C_{\mathrm{B}}(e) = \frac{1}{n(n-1)} \sum_{s,t \in V} \frac{\sigma_{s,t}(e)}{\sigma_{s,t}} \tag{1}$$

Computational complexity of the best known algorithm for computing the betweenness in (1)is $\mathcal{O}(mn)$ [4]. This limits its applicability for large graphs.

One of shortcomings of the betweenness centrality in (1)is that it takes into accounts only the shortest paths, ignoring the paths that might be one or two

* This research is partially funded by Inria Alcatel-Lucent Joint Lab, by the European Commission within the framework of the CONGAS project FP7-ICT-2011-8-317672, see www.congas-project.eu, and by the EU-FET Open grant NADINE (288956).

steps longer, while the edges on such paths can be important for communication processes in the network. In order to take such paths into account, Newman [11] and Brandes and Fleischer [5] introduced the current flow betweenness centrality (CF-betweenness). In [11,5] the graph is regarded as an electrical network with edges being unit resistances. The CF-betweenness of an edge is the amount of current that flows through it, averaged over all source-destination pairs, when one unit of current is induced at the source, and the destination (sink) is connected to the ground. This exploits the well known relation between electrical networks and reversible Markov chains, see e.g. [1,7].

The computational difficulty of Betweenness and the CF-betweenness is that the computations must be done over the set of all source-destination pairs. The best previously known computational complexity for the CF-betweenness is $\mathcal{O}(I(n-1)+mn\log n)$ where $I(n-1)$ is the complexity of the inversion of matrix of dimension $n-1$.

In the present work we introduce new betweenness centrality measures: α-current flow betweenness (α-CF betweenness) and its truncated version. The main purpose of these new measures is to bring down the high cost of the CF-flow betweenness computation. Our proposed measures are very close in performance to the CF-betweenness, but they are comparable to the PageRank algorithm [6] in their modest computational complexity. Our goal is to provide and analyze efficient algorithms for α-CF betweenness and truncated α-CF betweenness, to compare the α-CF betweenness to other centrality measures.

2 Alpha current flow betweenness

We view the graph G as an electrical network where each edge has resistance $1/\alpha$, and each node is connected to ground node $n+1$ by an edge with resistance $1/(1-\alpha)$. This is in the spirit of the PageRank, indeed, the current (probability flow) is inversely proportional to the resistance, and thus the fraction α of the current from a node flows to the network, while the fraction $(1-\alpha)$ of the current is directed to the sink. Since the graph is undirected, we use a convention that (v,w) and (w,v) represent the same arc in E, but depending on the chosen direction the current along this arc is considered to be positive or negative.

Assume that a unit of current is supplied to a source node $s \in V$, and there is a destination node $t \in V$ connected to the ground. Let $\varphi_v^{(s,t)}$ denote the absolute potential of node $v \in V$, if s is a source s, and t is the destination. Assume without loss of generality that $s=1$ and $t=n$ ($\varphi_n^{(1,n)} = \varphi_{n+1}^{(1,n)} = 0$). The vector of absolute potentials of the other nodes $\varphi^{(1,n)} = [\varphi_1^{(1,n)}, ..., \varphi_{n-1}^{(1,n)}]^T$ is a solution of the following system of equations (Kirchhoff's current law):

$$[\tilde{D} - \alpha\tilde{A}]\varphi^{(1,n)} = \tilde{b}, \tag{2}$$

where $\tilde{D}$ and $\tilde{A}$ are the degree and adjacency matrices of the graph without node n and $\tilde{b} = [1, 0, ..., 0]^T$, see [5].

Obviously, we would not like to solve a separate linear system for each source-destination pair with different left hand side coefficient matrix $[\tilde{D} - \alpha\tilde{A}]$. In the

following theorem we demonstrate that we need to only invert the coefficient matrix $[D - \alpha A]$.

Theorem 1 *The voltage drop along the edge (v, w) is given by*

$$\varphi_v^{(s,t)} - \varphi_w^{(s,t)} = (c_{s,v} - c_{s,w}) + \frac{c_{s,t}}{c_{t,t}}(c_{t,w} - c_{t,v}), \tag{3}$$

where $(c_{v,w})_{v,w \in V}$, are the elements of the matrix $C = [D - \alpha A]^{-1}$.

Proof: Assume again without loss of generality that $s = 1$ and $t = n$. The matrix $[D - \alpha A]$ can be written in the following block structure

$$D - \alpha A = \begin{bmatrix} \tilde{D} - \alpha \tilde{A} & -\alpha \tilde{a} \\ -\alpha \tilde{a}^T & d_n \end{bmatrix}, \quad \text{with} \quad \tilde{a} = \begin{bmatrix} a_{1,n} \\ a_{2,n} \\ \vdots \\ a_{n-1,n} \end{bmatrix}.$$

Then, divide accordingly the elements of the inverse matrix

$$C = [D - \alpha A]^{-1} = \begin{bmatrix} \tilde{C} & \tilde{c} \\ \tilde{c}^T & c_{n,n} \end{bmatrix}.$$

Writing the relation $[D - \alpha A]C = I$ in the block form yields

$$[\tilde{D} - \alpha \tilde{A}]\tilde{C} - \alpha \tilde{a}\tilde{c}^T = I, \tag{4}$$

$$[\tilde{D} - \alpha \tilde{A}]\tilde{c} - \alpha \tilde{a} c_{n,n} = 0. \tag{5}$$

Premultiplying equation (4) by $[\tilde{D} - \alpha \tilde{A}]^{-1}$, we obtain

$$[\tilde{D} - \alpha \tilde{A}]^{-1} = \tilde{C} - \alpha [\tilde{D} - \alpha \tilde{A}]^{-1} \tilde{a}\tilde{c}^T. \tag{6}$$

And premultiplying (5) by $[\tilde{D} - \alpha \tilde{A}]^{-1}$, we obtain

$$\alpha [\tilde{D} - \alpha \tilde{A}]^{-1} \tilde{a} = \frac{1}{c_{n,n}} \tilde{c}. \tag{7}$$

Combining both equations (6) and (7) gives

$$[\tilde{D} - \alpha \tilde{A}]^{-1} = \tilde{C} - \frac{1}{c_{n,n}} \tilde{c}\tilde{c}^T,$$

and hence $\varphi^{(1,n)} = [\tilde{D} - \alpha \tilde{A}]^{-1} \tilde{b} = \tilde{C}_{\cdot,1} - \frac{c_{1,n}}{c_{n,n}} \tilde{c}$. Thus, we can write

$$\varphi_v^{(1,n)} - \varphi_w^{(1,n)} = (c_{v,1} - c_{w,1}) + \frac{c_{1,n}}{c_{n,n}}(c_{w,n} - c_{v,n})$$

The above expression is symmetric and can be rewritten for any source-target pair (s, t). That is,

$$\varphi_v^{(s,t)} - \varphi_w^{(s,t)} = (c_{v,s} - c_{w,s}) + \frac{c_{s,t}}{c_{t,t}}(c_{w,t} - c_{v,t}).$$

Furthermore, since matrix C is symmetric for symmetric graphs, we can rewrite the above equation as

$$\varphi_v^{(s,t)} - \varphi_w^{(s,t)} = (c_{s,v} - c_{s,w}) + \frac{c_{s,t}}{c_{t,t}}(c_{t,w} - c_{t,v}),$$

which completes the proof. □

The current $I_e^{(s,t)}$ through edge $e = (v, w)$ is equal to $\alpha(\varphi_v^{(s,t)} - \varphi_w^{(s,t)})$. Let

$$x_e^{(s,t)} = |\varphi_v^{(s,t)} - \varphi_w^{(s,t)}|, \quad (v, w) \in E$$

be the difference of potentials, that determines the absolute value of the current on the edge. The α-CF betweenness of edge e is defined by

$$x_e^{\alpha} = \frac{1}{n(n-1)} \sum_{s,t \in V, s \neq t} x_e^{(s,t)}, \quad e \in E. \tag{8}$$

Further, for each node $v \in V$ its α-CF betweenness is defined as the sum of the α-CF betweenness scores of its adjacent edges:

$$\alpha\text{-CF betweenness}(v) = \sum_{(v,w) \in E} x_{(v,w)}^{\alpha}, \quad v \in V. \tag{9}$$

With this definition, the node is central if a relatively large amount of current flows from this node to the network. This is in accordance to the original CF-betweenness of [11,5], except we introduced the additional sink ground node $n + 1$. This mitigates the computational complexity because the original CF-betweenness require the inversion of the ill-conditioned matrix $[\tilde{D} - \tilde{A}]$, while for computing α-CF betweenness we need to invert the matrix $[D - \alpha A]$, which is a well posed problem, and has many possible efficient solutions, for example, power iteration and Monte Carlo methods. In fact, as we shall show below, we need to obtain just a few rows of the inverse matrix $[D - \alpha A]^{-1}$. In the rest of the paper we will discuss the computations and the properties of the α-CF betweenness.

3 Computation of α-CF betweenness

Due to the presence of the auxiliary node $n + 1$, the value of $x_e^{(s,t)}$ on the right-hand side of (8) can be computed efficiently with high precision for any source-destination pair. However, the summation over all $n(n - 1)$ pairs is a problem of prohibitive computational complexity even for graphs of a modest size. The solution is to perform the computations for sufficiently many source-destination pairs. This presents two problems: how to sample the source-destination pairs and how many such pairs we need to achieve a good precision.

Ideally, we would like to choose the most representative source-destination pairs. In particular, we can expect large values of $x_e^{(s,t)}$ if the sum of all potentials

$\sum_{v\in V}\varphi_v^{(s,t)}$ is maximal. Let us take again $s=1$, $t=n$. Then we obtain

$$\sum_{v\in V}\varphi_v^{(1,n)} = \mathbf{1}^T[\tilde{D}-\alpha\tilde{A}]^{-1}\tilde{b} = \mathbf{1}^T[I-\alpha\tilde{P}]^{-1}\tilde{D}^{-1}\tilde{b}, \tag{10}$$

where $\mathbf{1}$ is a column vector of ones, and $\tilde{P}$ is the transition probability matrix for a simple random walk on G with absorption in n. Compare this to the well-known expression for PageRank vector $\pi=(\pi_1,\ldots,\pi_n)$ with uniform teleportation and damping factor α:

$$\pi = \frac{1-\alpha}{n}\mathbf{1}^T[I-\alpha P]^{-1}.$$

Note that the vector $\mathbf{1}^T[I-\alpha\tilde{P}]^{-1}$ in (10) is very similar to PageRank, except it nullifies the contribution of node n. We denote this vector by $\tilde{\pi}$ and recall that $\tilde{b}=(1,0,\ldots,0)^T$ to obtain

$$\sum_{v\in V}\varphi_v^{(1,n)} = \tilde{\pi}_1 d_1^{-1}.$$

It is well-known and is also confirmed by our experiments that the PageRank of a node in an undirected graph is strongly correlated to the degree of the node. Thus, with any choice of the source, the sum of the potentials is of similar magnitude, except for the cases when the contribution of the destination node is defining for the PageRank mass of the source. However, the destination node will mainly affect the PageRank of its close neighbours. Thus, we propose to choose the source-destination pair uniformly at random, so that there is no preference on the source, and the probability of choosing neighbour nodes is small. This results in the next algorithm for computing the α-CF betweenness.

Algorithm 1.

1. Select a set of pairs of nodes $(s_i,t_i), i=1,...,N$, uniformly at random;
2. For each s_i or t_i, $i=1,...,N$ compute the rows $c_{s_i,\cdot}$, $c_{t_i,\cdot}$. (this can be done either by power iteration or by Monte Carlo algorithm);
3. For each edge $e=(v,w)$ and each pair (s_i,t_i), use (3) to compute

$$x_e^{(s_i,t_i)} = |\varphi_v-\varphi_w|.$$

4. Average over source-destination pairs

$$\bar{x}_e^{\alpha} = \frac{1}{N}\sum_{i=1}^{N} x_e^{(s_i,t_i)}.$$

Since we chose the pairs (s_i,t_i) uniformly at random then for every edge e, $\bar{x}_e^{\alpha}$ is just a sample average where all values are between zero and one. Then using the standard approach for the analysis of the series of independent random variables we have the following result.

Theorem 2 *Algorithm 1 approximates the alpha current flow betweenness in* $O(m\log(n)\varepsilon^{-2}\log(\varepsilon)/\log(\alpha))$ *time and* $O(m)$ *space to within an absolute error of* ε *with arbitrarily high fixed probability.*

Proof: In addition to the proof of Theorem 3 in [5] we just need to note that we can compute Personalized PageRank with precision ε in $\mathrm{O}(\log(\varepsilon)/\log(\alpha))$ power iterations. □

4 Truncated α-CF betweenness

In the experiments we noticed that the values $x_e^{(s,t)}$ have a high variance, which results in poor precision when evaluating x_e^{α}. A closer analysis revealed that the edges adjacent to the source s receive large values of $x_e^{(s,t)}$. This is especially apparent when $e=(v,s)$, where v has degree 1, so (v,s) is its only edge, and s has a large degree. This can be explained using the random walk interpretation. Consider a PageRank-type random walk on G. At each node, with probability α, the random walk traverses a randomly chosen edge of this node, and with probability $1-\alpha$ it jumps to the sink $n+1$. Denote by T_B the number of steps of the random walk needed to hit set B. Then it follows from Proposition 10 of [1, Chapter 3] that $\varphi_v^{(s,t)}/\varphi_s^{(s,t)} = P_v(T_{\{s\}} < T_{\{t,n+1\}})$, where $P_v(\cdot)$ is a conditional probability given that the random walk starts at v. Hence, if s is the only neighbor of v then $\varphi_v^{(s,t)}/\varphi_s^{(s,t)} = \alpha$, the probability of no absorption before reaching s. Thus, $|\varphi_s^{(s,t)} - \varphi_v^{(s,t)}| = (1-\alpha)\varphi_s^{(s,t)}$, which can be large if e.g. $\alpha = 0.8$ because $\varphi_s^{(s,t)}$ is the largest potential in the network. Furthermore, the original CF-betweenness corresponds to $\alpha = 1$, implying that the current in (v,s) is zero.

This motivates for the truncated version of α-CF betweenness where for each edge (v,w) we only take into account the scores $x_{(v,w)}^{(s,t)}$ if $v,w \neq s$. In Figure 1 we present log-linear plots of the empirical complementary distribution function of $x_{(v,w)}^{(s,t)}$ over all pairs (s,t) (solid line), and its truncated version (dashed line). The plots are given for two edges in the Dolphin social network described in Section 5 below. Nodes 1 and 36 are central in the network, so the high α-CF betweenness of (1,36) is expected. Node 60 has degree 1, so edge (32,60) gains an unwanted high betweenness in the non-truncated version.

Since the truncated α-CF betweenness gives lower scores to the edges connected to nodes of degree 1, one can expect that it has a higher correlation with CF-betweenness, especially for not very large α. This is confirmed below in Figure 2. Moreover, the truncated version removes outliers, and does not have large spread in values, thus standard statistical procedures, based on the Central Limit Theorem can be applied. Also, because of the smaller variance, Algorithm 1 achieves a desired precision with a smaller sample of source-destination pairs.

5 Datasets

We consider the four graphs described below.

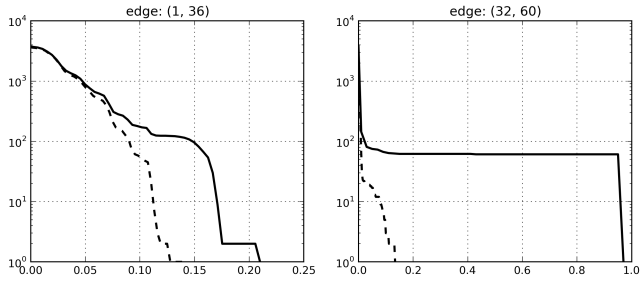


Fig. 1. The number of pairs s, t with $x^{(s,t)}_{(v,w)} > x$ over all pairs (s, t) (solid line) and only pairs with $v, w \neq s$. (dashed line)

Dolphin social network. This small graph represents a social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand [10].

Graph of VKontakte social network. We have collected data from a popular Russian social network VKontakte. We were considering subgraph representing one of the connected components of people who stated that they were studying at Applied Mathematics - Control Processes Faculty at the St. Petersburg State University in different years. We ran the breadth-first search (BFS) algorithm starting at one specific node on the network and then anonymized the obtained users' data leaving only information about connections between people. Collected network consists of 2092 individuals out of total 8859 denoted the specified faculty in the Education field.

Watts-Strogatz model. As an artificial example, we used a random graph generated by the Watts-Strogatz model. We have chosen this model as it combines high clustering and short average path length, thus different centrality measures give very different results on this graph. For other random models considered (Erdos-Renyi and Barabasi-Albert) all measures are highly correlated and behave very similar to each other.

Enron graph. Enron email communication network is a well known test dataset. It covers all the email communication within a dataset of around half million emails between Enron's employees. The node are e-mail addresses, and the edges appears if an e-mail message was sent from one e-mail to another. Although this graph is small compared to, say, web or Twitter samples, it is already prohibitively large for computing the CF-betweenness in its original form.

6 Numerical results for α-CF betweenness

To begin with, we compare the two versions of α-CF betweenness (truncated and without truncation) to the CF-betweenness scores defined as in [11,5]. Figure 2 presents the results for the three smaller graphs, in which the latter measure

	$\|V\|$	$\|E\|$	$\langle deg(v) \rangle$	$diam(G)$	$C_{\text{clustering}}$	$\langle d(u,v) \rangle$
Dolphin social network	62	159	5.13	8	0.259	3.357
VKontakte AMCP social graph	2092	14816	14.16	14	0.338	4.598
Watts-Strogatz ($n = 1000, k = 12, p = 0.150$)	1000	6000	12.00	6	0.422	3.713
Enron	36692	183831	10.02	11	0.4970	≈ 4.8

Table 1. Datasets characteristics

could be computed. As a correlation measure we use the Kendall tau rank correlation. We observe that the truncated version is better correlated with the

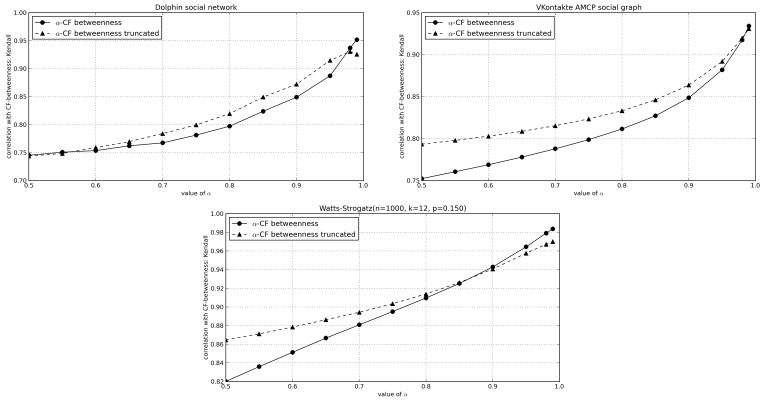


Fig. 2. Correlations between α-CF betweenness and truncated α-CF betweenness with CF-betweenness as a function of α.

CF-betweenness when α is not very close to one. As explained above, this is because the high probability of absorption results in a relatively high current in the edges connected to the source, which is not necessarily the case if absorption is only possible in the destination node.

Next, we demonstrate that that we can compute α-CF betweenness in the Enron graph, where the computation of CF-centrality is infeasible. We have evaluated α-CF betweenness, non-truncated and truncated, with $\alpha = 0.98$. We have run Algorithm 1 using with $N = 20 \cdot 10^6$ source-destination pairs. In the plot below we show the complementary distribution function in log-linear scale, of the score $x_e^{0.98}$ across the edges.

Note that distribution over edges (the left plot in Figure 3) does not have a large spread of values, except one outlier edge that connects two most important hubs. Since the weights of the edges are comparable, it is to be expected that in this graph the nodes of large degrees are also the ones with highest betweenness. Indeed, the Kendall's tau correlation between α-CF betweenness and degree of

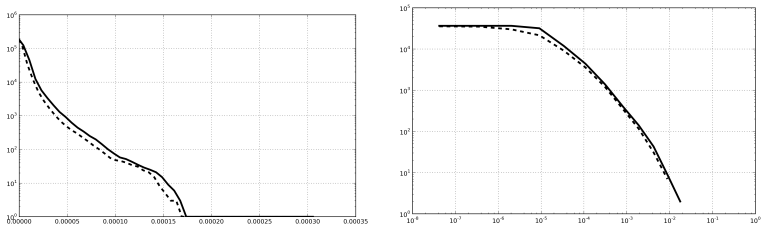


Fig. 3. Distribution of α-CF betweenness scores in the Enron graph, truncated (dashed line) and not truncated (solid line). Left: $x_e^{0.98}$ for edges $e \in E$. Right: α-CF betweenness (v) for $v \in V$. On the x-axis are the values of α-CF betweenness, on the y-axis the number of edges/nodes with the score larger than x.

the nodes turns out to be 0.808, which is higher than in small examples below. The reason can be either the graph size or its structure. In future research we will investigate how the CF-betweenness score, e.g. its maximum value across the edges, scales with the graph size in graphs with power law degrees.

We further present correlations between our proposed measures and other measure of betweenness. These are computed on smaller graphs where we could obtain exact values of all presented measures, see Tables 2–4. For completeness, we also include one distance-base centrality measure - the Closeness Centrality:

$$C_{\mathrm{C}}(v) = \frac{n-1}{\sum_{w \in V, w \neq v} d(v,w)},$$

where $d(v,w)$ is the graph distance between v and w. Betweenness (Between.) is computed as in (1), and PageRank(PR) is computed with $\alpha = 0.85$.

	Degree	PR	Closeness	Between.	CF	αCF(0.8)	αCF-tr(0.8)	αCF(0.98)
Degree	1.000	0.930	0.548	0.665	0.737	0.864	0.855	0.769
PageRank	0.930	1.000	0.458	0.658	0.733	0.872	0.827	0.757
Closeness	0.548	0.458	1.000	0.578	0.575	0.515	0.573	0.591
Betweenness	0.665	0.658	0.578	1.000	0.829	0.749	0.759	0.828
CF	0.737	0.733	0.575	0.829	1.000	0.798	0.820	0.939
αCF(0.8)	0.864	0.872	0.515	0.749	0.798	1.000	0.925	0.838
αCF-tr(0.8)	0.855	0.827	0.573	0.759	0.820	0.925	1.000	0.876
αCF(0.98)	0.769	0.757	0.591	0.828	0.939	0.838	0.876	1.000

Table 2. Kendall tau for centrality measures in Dolphin social network.

Note that α-CF betweenness is strongly correlated with CF-betweenness. The Closeness Centrality does not agree well with the CF-betweenness, even the PageRank and the degrees have a higher correlations with the CF-betweenness in real graphs. Recent paper [2] suggests more measures based on distance, and

	Degree	PR	Closeness	Between.	CF	αCF(0.8)	αCF-tr(0.8)	αCF(0.98)
Degree	1.000	0.655	0.679	0.521	0.545	0.659	0.668	0.599
PageRank	0.655	1.000	0.375	0.662	0.717	0.833	0.811	0.766
Closeness	0.679	0.375	1.000	0.382	0.356	0.424	0.445	0.395
Betweenness	0.521	0.662	0.382	1.000	0.761	0.760	0.749	0.778
Current Flow	0.545	0.717	0.356	0.761	1.000	0.812	0.833	0.917
αCF(0.8)	0.659	0.833	0.424	0.760	0.812	1.000	0.938	0.878
αCF-tr(0.8)	0.668	0.811	0.445	0.749	0.833	0.938	1.000	0.903
αCF(0.98)	0.599	0.766	0.395	0.778	0.917	0.878	0.903	1.000

Table 3. Kendall tau for centrality measures in the social graph VKontakte AMCP.

	Degree	PR	Closeness	Between.	CF	αCF(0.8)	αCF-tr(0.8)	αCF(0.98)
Degree	1.000	0.891	0.462	0.526	0.610	0.643	0.581	0.612
PageRank	0.891	1.000	0.415	0.485	0.565	0.610	0.546	0.567
Closeness	0.462	0.415	1.000	0.655	0.613	0.647	0.666	0.628
Betweenness	0.526	0.485	0.655	1.000	0.853	0.819	0.852	0.857
Current Flow	0.610	0.565	0.613	0.853	1.000	0.910	0.914	0.979
αCF(0.8)	0.643	0.610	0.647	0.819	0.910	1.000	0.935	0.923
αCF-tr(0.8)	0.581	0.546	0.666	0.852	0.914	0.935	1.000	0.930
αCF(0.98)	0.612	0.567	0.628	0.857	0.979	0.923	0.930	1.000

Table 4. Kendall tau for centrality measures in the Watts-Strogatz graph (n=1000, k=12, p=0.150).

efficient computation method for such measures is presented in [3]. In future it will be interesting to compare these new measures to α-CF betweenness.

7 Centrality measures and network vulnerability

We now consider how well the CF-betweenness and α-CF betweenness can indicate the nodes responsible for maintaining the network connectivity. We follow the methodology in [9]. As measures of connectivity we choose the average inverse distance

$$< d^{-1} >= \frac{1}{n(n-1)} \sum_{u,v \in V, u \neq v} \frac{1}{d(u,v)}$$

and the size of the largest connected component. In the experiment, we remove the top nodes one by one, according to different betweenness measures, and observe how the connectivity of the network changes. In Figure 4 the results are presented for the inversed average distance.

The results for the social graph VKontakte are especially interesting, because this network turns out to be less vulnerable to the removal of nodes with large degree than nodes with large betweenness and its modifications (CF-betweenness, α-CF betweenness, and truncated α-CF betweenness). On the small Dolphin

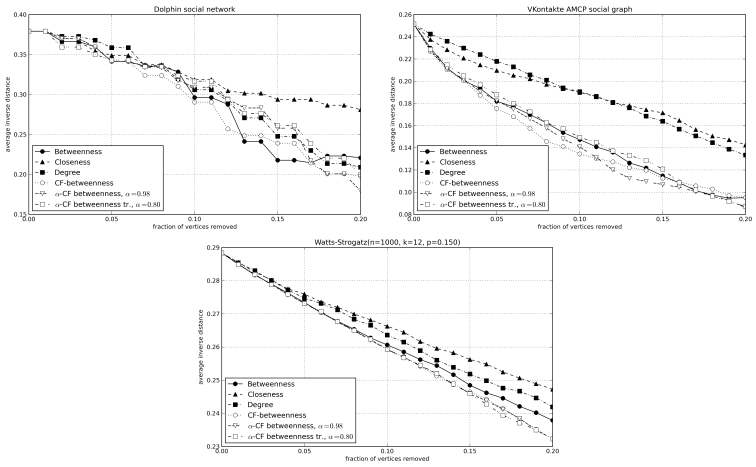


Fig. 4. Inverse average distance as a function of the fraction of removed top-nodes according to different betweenness centrality measures.

social network there is no much difference in vulnerability with respect to different centrality measures. Finally, on the artificial Watts-Strogatz graph the CF-betweenness and our proposed two versions of α-CF betweenness find the nodes that are most essential for the network connectivity.

Another connectivity measure of the network is the size of its larges connected component. In Figure 5 we plot the size of the largest connected components against the fraction of removed top-nodes. We do not present the plot for the Watts-Strogatz graph because it remains entirely connected, so the size of its largest connected component equals to the number of remaining nodes irrespectively of which nodes are removed first. For the two real graphs, the CF-betweenness is most efficient in reducing the size of the giant component. On the Dolphin graph, α-CF betweenness performs closely to CF-betweenness, except the interval when 13-18% of nodes are removed. On the graph VKontakte, α-CF betweenness and its truncated version perfom comparably to the CF-betweenness. Again, on this graph, degree and Closeness centrality fail to reveal the nodes responsible for the network connectivity. The α-CF betweenness with $\alpha = 0.98$ appears to be a better measure for betweenness of a node than the truncated α-CF betweenness with $\alpha = 0.8$. The latter however also gives gives good results, and can be computed easier on large graphs due to the faster convergence of the power iteration algorithm.

We conclude that both α-CF betweenness and truncated α-CF betweenness provide an adequate measure for the role of a node in network's connectivity. Furthermore, their computational costs are lower than for known measures of betweenness, and the computations can be done in parallel easily. Thus, α-CF betweenness can be applied in large graphs, for which computations of other measures of betweenness are merely infeasible.

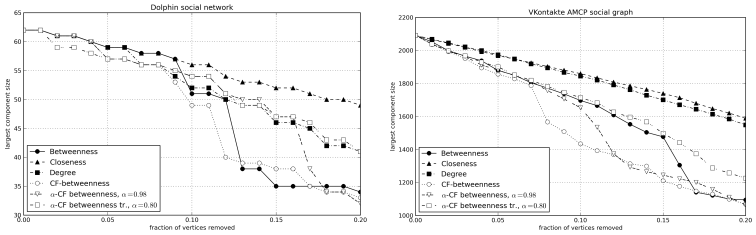


Fig. 5. The size of the largest connected component as a function of the fraction of removed top-nodes according to different betweenness centrality measures.

References

1. D. Aldous and J. Fill. Reversible Markov chains and random walks on graphs. 1999.
2. P. Boldi and S. Vigna. Axioms for centrality. *arXiv:1308.2140*.
3. P. Boldi and S. Vigna. In-core computation of geometric centralities with hyperball: A hundred billion nodes and beyond. *arXiv:1308.2144*.
4. U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(1994):163–177, 2001.
5. U. Brandes and D. Fleischer. Centrality measures based on current flow. In *Proceedings of the 22nd annual conference on Theoretical Aspects of Computer Science*, pages 533–544, 2005.
6. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and {ISDN} Systems*, 30(17):107–117, 1998.
7. P.G. Doyle and J.L. Snell. *Random walks and electric networks*. Mathematical Association of America, 1984.
8. L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 1977.
9. P. Holme, B.J. Kim, C.N. Yoon, and S.K. Han. Attack vulnerability of complex networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 65(5 Pt 2):056109, May 2002.
10. D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, and S.M. Dawson. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, September 2003.
11. M.E.J. Newman. A measure of betweenness centrality based on random walks. *Social networks*, pages 1–15, 2005.

Quick detection of popular entities in large directed networks

ABSTRACT

In this paper, we address a problem of quick detection of popular entities in large online social networks. Practical importance of the problem is attested by a large number of companies that continuously collect and update statistics about popular entities. We suggest an efficient two-stage algorithm for solving this problem. For instance, our algorithm needs only one thousand API requests in order to find the top-50 most popular users in Twitter, a network with more than a billion of registered users. Our algorithm is easy to implement, it outperforms existing methods, and serves many different purposes, such as finding most popular users or most popular interest groups in social networks. An important contribution of this work is the analysis of the proposed algorithm using the Extreme Value Theory – a branch of probability that studies extreme events and properties of largest order statistics in random samples. Using this theory, we derive accurate predictions for the algorithm's performance and show that the number of API requests for finding top-k most popular entities is sublinear in the number of entities. Moreover, we formally show that the high variability among the entities, expressed through heavy-tailed distributions, is the reason for the algorithm's efficiency. We quantify this phenomenon in a rigorous mathematical way.

1. INTRODUCTION

In this paper, we propose a randomized algorithm for quick detection of popular entities in large online social networks. The entities can be, for example, users or interest groups, user categories, geographical locations, etc. For instance, one can be interested in finding out a list of Twitter users with many followers or Facebook interest groups with many members. Practical importance of the problem is attested by a large number of companies that continuously collect and update statistics about popular entities (*twittercounter.com*, *followerwonk.com*, *twitaholic.com*, *www.insidefacebook.com*, *yavkontakte.ru* just to name a few).

The problem at hand may seem trivial, if one assumes that the network structure and the relation between entities are known. However, even then finding, for example, top-k in-degree nodes in a directed graph G of size N takes the time $\mathrm{O}(N)$. For large networks, such linear complexity is already too high. In fact, for any practical purpose, it is much more valuable to find an approximate result in a sublinear time than an exact result in a linear time. Furthermore, the data of current social graphs is typically available only to the owners of the social network, and can be obtained by other interested parties only through API requests. The rate of allowed API requests is usually quite small. For instance, Twitter has the limit of one access per minute for one standard API account. Then, in order to crawl the entire network with more than 500 million users one need more than 950 years. Clearly, we would like to find most popular entities using only a small number of API requests.

Formally, the problem addressed in this paper is as follows. Let V be a set of entities, usually users, that can be accessed using API requests. Let W also be another set of entities (possibly equal to V). We represent V and W as vertices of a bipartite graph (V, W, E), where a directed edge $(v, w) \in E$, with $v \in V$, $w \in W$, represents a relation between v and w. For instance, if $V = W$ is a set of Twitter users, then $(v, w) \in E$ may mean that v follows w, or that v retweeted a tweet from w. Note that any directed graph $G = (V, E)$ can be represented equivalently by the bi-partite graph (V, V, E). One can also suppose that V is a set of users and W is a set of interest groups, while the edge (v, w) represents that user v belongs to group w. Our goal is to quickly find top-k in-degree entities in W. In this setting, throughout the paper, we use the terms 'nodes' and 'entities' interchangeably.

The algorithm proposed in this paper can detect popular entities with high precision using very small number of API requests. Most of our experiments are performed on the Twitter graph, because it is a good example of a huge network (billion of registered users) and very limited rate of requests to API. We use only 1000 API request to find top-50 Twitter users with very high precision. We also demonstrate the efficacy of our approach on the example of online social network (to be specified in the camera-ready version) which had, at the time of article preparation, more than 200 million registered users. We use our algorithm to quickly detect most popular interest groups in this social network. Experiments on random graph models show that our algorithm outperforms the baselines algorithms from [4] and [14]. Moreover, our algorithm can be used in a very general settings for finding most popular entities, while the baseline algorithms can only be use for finding nodes of largest de-

grees in directed ([14]) or undirected ([4]) graphs.

An important contribution of this work is the novel analysis of proposed algorithm using classical results of the Extreme Value Theory (EVT) – a branch of probability that studies extreme events and properties of largest order statistics in random samples. We refer to [8] for a comprehensive introduction to EVT. Specifically, we treat the largest in-degrees in W as high order statistics of a heavy-tailed distribution, and use EVT to obtain the limiting properties of these order statistics. This way we obtain statistical estimation of the magnitude of the largest in-degrees in W. Using these mathematical tools, we can, for instance, accurately predict the average fraction of correctly identified top-100 most followed users in Twitter using only the knowledge of top-20 degrees, which can be detected by our algorithm very quickly with practically 100% accuracy.

We derive the complexity of our algorithm in terms of the number of entities in W show that the complexity is sublinear if the in-degree distribution in W is heavy tailed. Intuitively, this should be the case because the high variability of the in-degrees implies that the largest entities have extremely large number of in-links and thus are easy to find. We formalize this argument using the EVT results.

The algorithm consists of two stages. The parameters of the algorithm, n_1 and n_2, are the number of API requests used on the first and the second stage, respectively. The performance of the algorithm is very robust with respect of the parameters' values. We find the optimal scaling for n_1 and n_2 with respect to three measures of the algorithm performance: the average fraction of correctly identified top-k entities, the first-error index (the number of the highest statistics within top-k that was not included in the identified top-k list), and the the sum of incoming degrees of identified n_2 entities. Notice that for fixed n, the latter performance measure does not monotonically grows with n_2 because with small n_1 the number of links received from n_1 random users is a poor indication of the node's actual degree. This can be clearly seen in Figure 2 for the Twitter graph.

The rest of the paper is organized as follows. In Section 2, we give a short overview of the related work. In Section 3, we formally describe our algorithm. We empirically show the efficiency of our algorithm and compare it to baseline strategies in Section 4. We present a detailed analysis of the algorithm in Section 5 and evaluate its optimal parameters with respect to the above mentioned performance characteristics. Section 6 concludes the paper.

2. RELATED WORK

Over the last years data sets have become increasingly massive. For such large data any complexity higher than linear (in dataset size) is unacceptable, and even linear complexity may be too high. It is also well understood that an algorithm, which runs in sublinear time, cannot return an exact answer. In fact, such algorithms often use randomization, and then errors occur with positive probability. Nevertheless, in practice, a rough but quick answer is often more valuable than exact but computationally demanding solution. Therefore, sublinear time algorithms become increasingly important, and many studies of such algorithms have appeared in recent years (see, e.g., [10, 13, 15, 16]).

An essential assumption of this work is that the network structure is not available, and has to be discovered using the API requests. This setting is similar to on-line computations, when information is obtained and immediately processed while crawling the network graph (for instance the World Wide Web). There is a large body of literature where such on-line algorithms are developed and analyzed. Many of these algorithms are developed for computing and updating the PageRank vector [1, 6]. In particular, the algorithm recently proposed in [6] computes the PageRank vector in sublinear time. Furthermore, the probabilistic Monte Carlo methods [2, 11] allow to continuously update the PageRank as the structure of the Web changes.

Randomized algorithms are also used for discovering the structure of social networks. Often random walks are designed in such a way that the desired nodes are easily found. For example, in [12] an unbiased random walk, where each node is visited with equal probability, is constructed in order to find the degree distribution on Facebook. A different random walk is designed in [4] for finding nodes with largest degrees in undirected graphs. This random walk has jumps, so that it does not get stuck around just one hub, but unlike PageRank, its a stationary distribution completely defined by the nodes' degrees.

The problem of finding the most popular entities in large networks has been analyzed in several papers. In Section 4.3 we show that our algorithm outperforms two baselines: the random walk algorithm in [4], and the crawling algorithm in [14]. The latter algorithm [14] is designed to efficiently discover the correct set of pages with largest incoming degrees in a fixed network, and to track these pages over time when the network is changing. Their setting is different from ours in several aspects. For example, in our case we can use API to get indegree of any given item, while in the World Wide Web this information is not available. On the other hand, the algorithm in [14] is designed to discover the graph structure, and cannot be easily adopted for other tasks, such as finding most popular use categories or interest groups.

To the best our knowledge, this is the first work that presents and analyzes an efficient algorithm for retrieving most popular entities under realistic API constraints.

3. ALGORITHM DESCRIPTION

Recall that we consider a bipartite graph (V, W, E), where V and W are sets of entities, and $(v, w) \in E$ represents a relation between the entities.

Let $n = n_1 + n_2$. Our algorithm has two stages, described below. See Algorithm 1 for the pseudocode.

First stage. We start by sampling uniformly at random a set A of n_1 entities (users, or nodes) $v_1, \ldots, v_{n_1} \in V$. The nodes are sampled independently, so the same node may appear in A more than once, in which case we regard each copy of this node as a different node. Since multiplicities occur with very small probability this does not affect the efficiency of the algorithm but simplifies the implementation. For each node in A we record its out-neighbors in W. In practice, we bound the number of recorded out-links by the maximal number of id's that can be retrieved within one API request, thus the first stage uses exactly n_1 API requests.

Second stage. Let S_w, $w \in W$, be the number of nodes in A that have a (recorded) edge to w, and let w_i be the node in W with i-th largest values of S_w, so that $S_{w_1} \geqslant S_{w_2} \geqslant \cdots \geqslant S_{w_N}$. Then we use another n_2 API requests to retrieve the actual in-degrees of the n_2 top-nodes $w_1, \ldots, w_{n_2}$.

The set $\{w_1, w_2, \ldots, w_{n_2}\}$ is supposed to contain nodes from W with large in-degrees. For example, if we are inter-

ested in top-k in-degree nodes in a directed graph, we hope to identify these nodes with high precision if k is significantly smaller than n_2.

Algorithm 1: Find entities with large incoming degrees

input : Set of entities V of size M, set of entities W of size N, number of random nodes n_1, number of candidate nodes n_2
output: Nodes $w_1, \dots w_{n_2} \in W$, their degrees $d_1, \dots, d_{n_2}$

```
for i ← 1 to N do
    S[i] ← 0;
for i ← 1 to n_1 do
    v ← random(M);
    F ← OutNeighbors(v);
    foreach j in F do
        S[j] ← S[j] + 1;
w_1, ..., w_{n_2} ← Top_n2(S) // S[w_1], ..., S[w_{n_2}] are top n_2 values in S;
for i ← 1 to n_2 do
    d_i ← InDegree(w_i);
```

4. EXPERIMENTS

4.1 Twitter graph

First, we show that our algorithm quickly finds the most popular users in Twitter graph. Formally, V is a set of Twitter users, $W = V$, and $(v, w) \in E$ iff v is a follower of w. Twitter is an example of a huge network with limited access to its structure. Information on the Twitter graph can be obtained via Twitter API. The standard rate of requests to API is one per minute. Every vertex has an id, which is an integer number starting from 12. The largest id of a user is $\sim$ 1460M (at the time when we performed the experiments). Due to such id assignment, a random user in Twitter can be easily chosen. The only problem is that some users in this range have been deleted, some are suspended, and therefore errors occur when addressing the id's of these pages. In our implementation we usually skip errors and assume that we do not spend resources on such nodes. The fraction of errors is $\approx 20\%$.

Given an id of a user, a request to API can return one of the following: i) the number of followers (indegree), ii) the number of followees (outdegree), or iii) at most 5000 id's of followers or followees. If a user has more than 5000 followees, then all their id's can be retrieved only by using several API requests. Instead, as described above, we record only the first 5000 of the followees and ignore the rest. This does not affect the performance of the algorithm because we record followees of randomly sampled users, and the fraction of Twitter users with more than 5000 followees, is small.

In order to obtain the ground truth, we first took $n_1 = n_2 = 500\,000$ and found top-1000 users with a very high precision. We used the obtained list for evaluating the performance of our algorithm.

Figure 1 shows the average fraction of correctly identified users from top-k for different k over 100 experiments, as a function of n_2, when $n = 1000$. Remarkably, we can find top-50 users with very high precision.

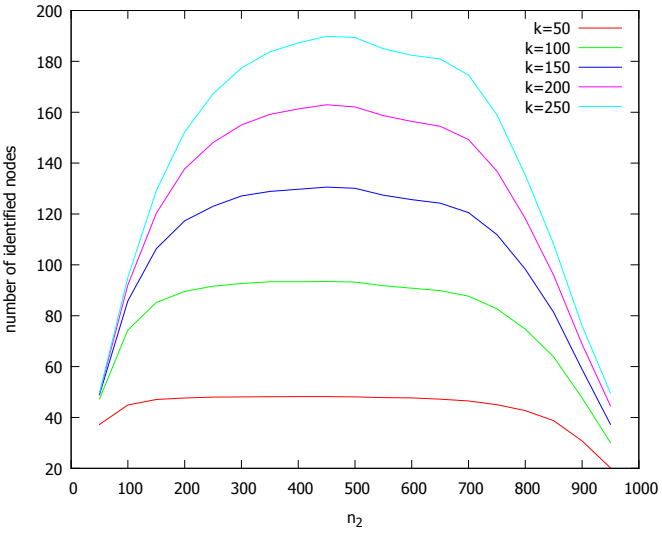


Figure 1: The number of correctly identified top-k most followed Twitter users as a function of n_2, with $n = 1000$.

We have also looked at the first-error index – the position of the first mistake in the top-k list. Formally, if we correctly identified top-$(i-1)$ users, but did not find the ith user, then the first-error index is i. Again, we have averaged the results over 100 experiments. Results are shown in Figure 4 below (red line). Note that with only 1000 API requests we can correctly identify more than 50 users without any omission.

The sums of the degrees of the identified top-n_2 entities, with $n = 1000$, are depicted in Figure 2. Observe that here the optimal value of n_2 is larger than in two previously discussed metrics. Thus, in order to to discover as many true in-links as possible, we may want to check more incoming degrees in the second stage of the algorithm, so that we have a large output list, but with less precision. We will discuss this in more detail in Section 5.3.

4.2 Finding largest interest groups

Let V be a set of users, W be a set of interest groups, and $(v, w) \in E$ iff v is a member of w.

We will demonstrate that our algorithm can find the most popular groups in a large social network with more than 200M registered users (to be specified in the camera-ready version). As for Twitter, information on the network under consideration can be obtained via API. Again, all users have ids: integer numbers starting from 1. Due to this id assignment, a random user in this network can be easily chosen. In addition, all interest groups also have their own id's.

We are interested in the following requests to API: i) given id of a user, return his or her interest groups, ii) given id of a group return its number of members. If a user decide to hide the list of groups, then an error occurs. The portion of such errors is $\approx 30\%$.

As before, first we used our algorithm with $n_1 = n_2 = 50000$ in order to find the most popular groups with high precision. Table 1 presents some statistics on the most popular groups. Then, we took $n_1 = 700$, $n_2 = 300$ and com-

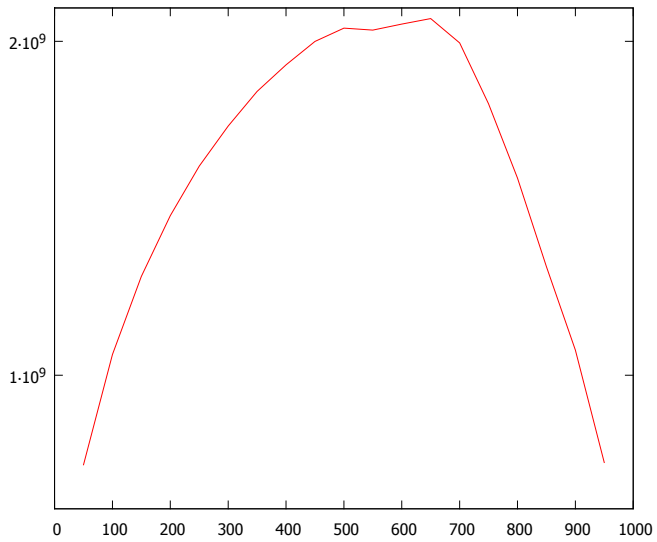


Figure 2: The sum of incoming degrees of identified users as a function of n_2, $n = 1000$.

Table 1: The most popular groups

Rank	Number of participants	Topic
1	4,35M	humor
2	4,1M	humor
3	3,76M	movies
4	3,69M	humor
5	3,59M	humor
6	3,58M	facts
7	3,36M	cookery
8	3,31M	humor
9	3,14M	humor
10	3,14M	movies
100	1,65M	success

puted the fraction of correctly identified groups from top-100. Using only 1000 API requests, our algorithm identifies on average 73.2 from the top-100 interest groups (averaged over 25 experiments). The standard deviation is 4.6.

4.3 Comparison with baseline algorithms

In this section we compare our algorithm with the algorithms suggested in [4] and [14]. We start with the description of these algorithms.

Random walk based algorithm [4]. The algorithm in [4] is a randomized algorithm for undirected graphs that finds a top-k list of nodes with largest degrees in sublinear time. It is based on the random walk with uniform jumps, described by the following transition probabilities [5]:

$$p_{ij} = \begin{cases} \frac{\alpha/N+1}{d_i+\alpha}, & \text{if } i \text{ has a link to } j, \\ \frac{\alpha/N}{d_i+\alpha}, & \text{if } i \text{ does not have a link to } j, \end{cases} \quad (1)$$

where N is the number of nodes in the graph and d_i is the degree of node i. The parameter α controls how often the random walk makes an artificial jump. In [4] it is suggested to take the parameter α equal to the average degree

Algorithm 2: Random walk based algorithm

input : Graph G with N nodes, E edges, number of steps n, size of output list k, parameter α
output: Nodes $v_1, \dots v_k$, their degrees $d_1, \dots, d_k$

```
v ← random(N);
F ← Neighbors(v);
D[v] ← size(F);
for i ← 2 to n do
    r ←(sample)— U[0, 1];
    if r < D[v]/(D[v]+α) then
        v ← random from F;
    else
        v ← random(N);
    F ← Neighbors(v);
    D[v] ← size(F);
v_1, ..., v_k ← Top_k(D) // D[v_1], , D[v_k] are top k values in D;
```

in order to maximize the number of independent samples. Interestingly, this implies that the random walk, in stationarity, makes on average just one step between the jumps. With such choice of α the random walk method of [4] mimics most closely the suggested algorithm with independent sampling and exactly one step from entity in V to entity in W. We should note that the random walk method could be very valuable when the independent uniform sampling is expensive, for example, when the id space is very sparse.

The random walk keeps a candidate list of k nodes. Once a new node is discovered according to the transition probability (1), we check its degree and compare it with degrees of the nodes in the candidate list. If this newly discovered node has a degree larger than degrees of some nodes in the candidate list, the newly discovered node is inserted in the candidate list and a node with the smallest degree in the candidate list is pushed out. See Algorithm 2 for more detailed description. The algorithm can be run for a predefined number of steps or can be terminated according to one of the stopping criteria provided in [4].

Crawl-Al and Crawl-GAI [14]. At each step we consider one node and ask for its outgoing edges. At step n any node j has its *apparent indegree* S_j, $j = 1, \dots, N$: the number of discovered edges pointing to this node. In Crawl-Al the next node to consider is a random node, with probability proportional to the apparent indegree. In Crawl-GAI, the next node is the node with the highest apparent indegree. After n steps we get a list of nodes with largest apparent indegrees. See Algorithm 3 for the pseudocode of the algorithm Crawl-GAI.

In the experiments of the present paper we take the same budget for all tested algorithms to compare their performance.

Note that we cannot compare the algorithms on the Twitter graph for several reasons. First, Algorithm 2 works only on undirected graphs. Second, in order to choose a random edge of a node, we need at least two request to API, to ask for followees and followers. Also, the random walk often hits nodes of high degree, and then many additional requests are needed to retrieve their followers and followees, because the

Algorithm 3: Crawl-GAI

input : Graph G with N nodes, number of steps n, size of output list k
output: Nodes $v_1, \dots v_k$

for $i \leftarrow 1$ **to** N **do**
  $S[i] \leftarrow 0$;
for $i \leftarrow 1$ **to** N **do**
  $v \leftarrow \text{argmax}(S[i])$;
  $F \leftarrow OutNeighbors(v)$;
  foreach j *in* F **do**
    $S[j] \leftarrow S[j] + 1$;
$v_1, \dots, v_k \leftarrow Top\_k(S)$ // $S[w_1], \dots, S[w_k]$ are top k values in S;

Table 2: Number of correctly identified nodes from top-100 averaged over 100 experiments, $n = 1000$.

Algorithm	mean	standard deviation
Our (directed)	91.9	4.88
Crawl GAI (directed)	81.9	2.42
Crawl AI (directed)	82.9	2.38
Our (undirected)	97.9	1.71
Random walk (undirected)	60.7	4.76

number of id's that can be obtained in one request is limited (5000 in Twitter). For example, we need 6K request to get the followers of a user with 30M followers. Algorithm 3 crawls only out-degrees, that are usually much smaller, but it can potentially suffer from the API constraints, for example, when in-degrees and out-degrees are dependent.

Therefore, in order to compare Algorithms 1–3, we have generated a random directed graph according to the configuration model (see [7]). Our artificial graph has 1M nodes, 6M edges, and the parameter of the power law degree distribution is 2. This directed graph is used to compare our algorithm to Crawl-Al and Crawl-GAI. In order to compare our method to the random walk based algorithm, we treat the generated graph as undirected. As prescribed by [4], we took α slightly smaller than the average degree in the graph (in our case $\alpha = 10$) and we considered a random walk with 1000 steps.

For the algorithm suggested in this paper we took $n_1 = 700$, $n_2 = 300$. The results of comparison can be seen in Table 2.

We expect our algorithm with $n_1 = 1000$ to be close to Crawl-GAI. Indeed, in the directed case our algorithm with $n_1 = 1000$ identifies 81.4 nodes from top-100 on average (this number is not presented in the table). Further improvement of our algorithm over the baselines is obtained because of the right balance between n_1 and n_2.

5. ANALYSIS OF THE ALGORITHM

In this section, we present the theoretical analysis of Algorithm 1. The goal of this analysis is: 1) to mathematically justify our suggested two-steps procedure; 2) to prove that the total number of API requests, n, scales sublinearly with the network size, N; 3) to find the optimal scaling of n_1 and n_2 (the number of API requests in the first and the second stage of the algorithm) with respect to n.

We number the nodes in W by $1, 2, \dots, N$ according to the number of incoming links, from most popular to least popular. As prescribed by Algorithm 1, we pick n_1 nodes in V uniformly at random. The first important observation is that S_j follows a binomial distribution. Indeed, let F_j be the unknown random in-degree of node $j \in W$, so that $F_1 \geqslant F_2 \geqslant \dots \geqslant F_N$. Then, if we label all nodes from V that have a edge to j (we call such nodes followers of j), then S_j is exactly the number of labeled nodes in a random sample of n_1 nodes, so its distribution is $Binomial\Big(n_1, \frac{F_j}{N}\Big)$. Hence, we have

$$\mathbb{E}(S_j|F_j) = n_1 \frac{F_j}{N}, \quad \text{Var}(S_j) = n_1 \frac{F_j}{N}\Big(1 - \frac{F_j}{N}\Big). \qquad (2)$$

For the top nodes with large F_j this distribution can be approximated with the Poisson distribution $Poisson\Big(\frac{n_1 F_j}{N}\Big)$.

5.1 Candidate list

The quality of the top-k lists produced by Algorithm 1 is defined by the events whether or not the value of S_j, $j = 1, \dots, k$, is among the top-n_2 values of $S_1, S_2, \dots, S_N$, obtained in the first stage of the algorithm. This is justified by the intuition that if $F_j > F_l$, then we are likely to see $S_j > S_l$. Note, however, that the case when S_j is as small as 1, the event $1 = S_j > S_l = 0$ is not informative.

EXAMPLE 1. *Let us take $n_1 = n_2 = 500$ in the case of the Twitter graph. Then the average number of nodes i among the top-10000 with $S_i = 1$ is already*

$$\sum_{i=1}^{10^4} P(S_i = 1) \approx \sum_{i=1}^{10^4} \frac{500 F_i}{5 \cdot 10^8} e^{-500 F_i / 5 \cdot 10^8} = 2539.1,$$

hence, many more than n_2 nodes will have $S_i = 1$ and can make it to the top n_2 values of $S_1, S_2, \dots, S_N$ only with a small probability.

Motivated by the above considerations, we formulate our approach in terms of a statistical test as follows. Let our data be $S_1, S_2, \dots, S_N$. We assume that the observations are realizations of independent Poisson random variables with parameters $n_1 F_1/N, n_1 F_2/N, \dots, n_1 F_N/N$. For the two numbers $j, l \in 1, \dots, N$, we test the null-hypothesis $H_0 : F_j \leqslant F_l$ against the alternative $H_1 : F_j > F_l$. Let $S_{i_1} \geqslant S_{i_2} \geqslant \dots \geqslant S_{i_{n_2}}$ be the top-n_2 order statistics of $S_1, \dots, S_N$ obtained by Algorithm 1. Then the first stage of the algorithm is equivalent to rejecting $H_0 : F_{i_j} \leqslant F_{i_{n_2}}$ for $j = 1, \dots, n_2 - 1$ such that

$$S_{i_j} > \max\{S_{i_{n_2}}, 1\}. \qquad (3)$$

Here the strict inequality is necessary to guarantee that i_j is on the top-n_2 list after the first stage of the algorithm. If H_0 is rejected, then the actual degree of entity i_j will be retrieved in the second stage of the algorithm.

Note that in contrast to the classical hypothesis testing, here we do not draw the conclusions solely from the observed random data S_1, S_2, …, S_N but we obtain the true values of the parameters in the second stage of the algorithm. Hence, if we use $S_{i_{n_2}}$ as a proxy for S_{n_2}, then, given F_1, F_2, …, F_N, the quality of the top-k list is expressed as the power of

the test as follows:

$$\begin{aligned} &P(\text{node } j \text{ is found}|F_j, F_{n_2}) \\ &= P(S_j > \max\{S_{i_{n_2}}, 1\}|F_j, F_{i_{n_2}}) \quad (4) \\ &\approx P(S_j > \max\{S_{n_2}, 1\}|F_j, F_{n_2}) \\ &\approx \sum_{s=0}^{\infty} e^{-\frac{n_1 F_{n_2}}{N}} \frac{(n_1 F_{n_2})^s}{N^s s!} \sum_{r>\max\{s,1\}} e^{-\frac{n_1 F_j}{N}} \frac{(n_1 F_j)^r}{N^r r!} \\ &=: P_j(n_1), \quad j = 1, \ldots, k. \quad (5) \end{aligned}$$

5.2 Performance criteria

The main constraint of Algorithm 1 is the number of API requests that we can use. In order to measure the performance of the algorithm, we propose three objectives, described formally in this section.

The first objective is the average number of correctly identified top-k nodes. This is defined in the same way as in [3]:

$$\begin{aligned} &E[\text{fraction of correctly indentified top-}k \text{ entities}] \\ &= \frac{1}{k}\sum_{j=1}^{k} P(\text{node } j \text{ is found}|F_j, F_{n_2}) \approx \frac{1}{k}\sum_{j=1}^{k} P_j(n_1). \quad (6) \end{aligned}$$

The second objective is the first-error index, which is equal to i if the top $(i-1)$ entities are identified correctly, but the top-i entity is not identified. If all top-n_2 entities are identified correctly, we set the first-error index equal to $n+1$. Using that for a discrete random variable X with values $1, 2, \ldots, k+1$ holds $E(X) = \sum_{l=0}^{k} P(X > l)$, we obtain the average first-error index as follows:

$$\begin{aligned} E[\text{1st-error index}] &= \sum_{l=0}^{n_2} P(\text{1st-error index} > l) \\ &= \sum_{j=1}^{n_2+1} \prod_{l=1}^{j-1} P(S_j > \max\{S_{i_{n_2}}, 1\}|F_j, \ldots, F_{i_{n_2}}) \\ &\approx \sum_{j=1}^{n_2} \prod_{l=1}^{j-1} P_l(n_1). \quad (7) \end{aligned}$$

Finally, our last objective is the sum of the identified top-n_2 degrees, that can be written in a very simple form:

$$U := [\text{sum of identified } n_2 \text{ degrees}] = \sum_{l=1}^{n_2} F_{i_l}. \quad (8)$$

5.3 EVT performance predictions

In order to compute the values in (6), (7), we need to make assumptions on the top-n_2 in-degrees of entities in W: F_1, F_2, …, F_{n_2}. To this end, we employ the quantile estimation techniques from the Extreme Value Theory (EVT).

In most social networks the degrees of the entities show a great variability. This is often modeled using power laws, although it has been often argued that classical Pareto distribution does not always fit the observed data. In our analysis we assume that the incoming degrees of the entities in W are independent random variables following a *regularly varying* distribution G:

$$1 - G(x) = L(x)x^{-1/\gamma}, \quad x > 0, \quad (9)$$

where $L(\cdot)$ is a slowly varying function, that is,

$$\lim_{x\to\infty} L(tx)/L(x) = 1, \quad t > 0$$

($L(\cdot)$ can be, for example, a constant or a logarithm). We note that (9) describes a broad class of heavy-tailed distributions, for which the EVT arguments presented below are valid, without imposing the rigid Pareto assumption.

Observe that $F_1, F_2, \ldots, F_N$ are the order statistics of G. Assume now that we know the correct values of the top-m nodes, $m < k$. This is plausible because, for instance, in Twitter, with $n = 1000$, the top-50 nodes are identified with a very high precision, see Figure 1. Then, in order to estimate the value of γ, we can use the classical Hill's estimator $\hat{\gamma}$, based on the top-m order statistics:

$$\hat{\gamma} = \frac{1}{m-1}\sum_{i=1}^{m-1} \log(F_i) - \log(F_m). \quad (10)$$

Next, we use the quantile estimator, given by formula (4.3) in [9], but we replace their two-moment estimator by the Hill's estimator in (10). This is possible because both estimators are consistent (under slightly different conditions). Under the assumption $\gamma > 0$, we have the following estimator f_j for the $(j-1)/N$-th quantile of G:

$$f_j = F_m \left(\frac{m}{j-1}\right)^{\hat{\gamma}}, \qquad j > 1, j << N. \quad (11)$$

We propose to use f_j as a prediction of F_j.

Note that our argument is inspired but not entirely justified by [9] because the consistency of the proposed quantile estimator (11) is only proved for $j < m$, while we want to use it for $j > m$. However, in the experiments we observe that expressions (6) and (7) are very robust with respect to the estimated values $F_1, \ldots, F_{n_2}$. Moreover, $\hat{\gamma}$ increases with m, and it is easy to see that with smaller $\hat{\gamma}$ the predictions of the algorithm performance are more conservative.

In Figure 3 we compare the true fraction of the correctly identified top-k followed Twitter users to the performance prediction (6) for $n = 1000$ and $k = 100$. The magenta line shows the prediction for the fraction of correctly identified nodes in (6), where we used the correct values of $F_1, F_2, \ldots, F_{n_2}$. The green line represents the results for the estimated values of $F_1, \ldots, F_k$ and F_{n_2}, based on the true values of the top-20 degrees. We see that it is very close to the magenta line, which is based on the true values of the degrees.

Similarly, we use formula (7) and the estimator (11) in order to provide the prediction of the first-error index. The results are given in Figure 4. We see again that the EVT predictions are more pessimistic than the experimental results, so we find the lower bound for the algorithm's actual performance. Note also that the shape of the plot and the optimal value of n_2 have been captures correctly by both predictors.

It is also clear that there is a principal difficulty in finding similar analytical predictions for the objective U in (8) because is it is based not on the *actual* degrees $F_1, F_2, \ldots$, but on the degrees $F_{i_1}, F_{i_2}, \ldots, F_{i_{n_2}}$, where $S_{i_1} \geqslant S_{i_2} \geqslant \cdots \geqslant S_{i_N}$ are the order statistics of the S_j's. The exact expressions for such order statistics are rather messy. However, we can get some insight in the behavior of U in Figure 2. Indeed, clearly, the sum of correct top-n_2 degrees, $\sum_{i=1}^{n_2} F_i$, is an increasing function of n_2. Moreover, if we use the estimator (11), then we observe that the largest values of F_j's

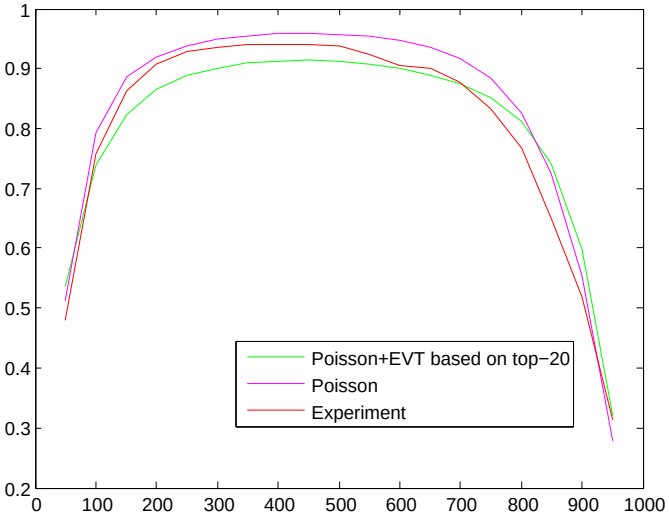


Figure 3: Fraction of correctly predicted nodes out of top-100 as a function of n_2, with $n = 1000$: experiments (red); prediction (5) based on the true values of the degrees (magenta); prediction (5) based on top-m degrees and estimator (11) with $m = 20$, $\hat{\gamma} = 2.2$ (green).

are of the same order of magnitude:

$$F_j/F_l \approx \left(\frac{l-1}{j-1}\right)^{\hat{\gamma}}.$$

Thus, as long as n_1 large enough so that a large entity j receives large S_j, we have that U is comparable to $\sum_{i=1}^{n_2} F_i$, and hence U increases in n_2. However, as n_1 becomes smaller, then small entities will constitute a large proportion of the set $\{i_1, i_2, \ldots, i_{n_2}$. For example, if $n_2 = 800$, $n_1 = 200$, then we obtain, for the true values of in-degrees in Twitter graph with $N \approx 500$M:

$$\sum_{i=1}^{800} P(S_i > 1) \approx 280.9,$$

thus on average about 520 out of the top-800 nodes will be undistinguishable from other, much smaller nodes (see Example 1). Moreover, in this case

$$\sum_{i=1}^{10^5} P(S_i > 1) \approx 485.18,$$

thus, on average, more than 300 nodes will be included into $\{i_1, \ldots, i_{800}\}$ essentially on a random basis. Since large majority of the nodes has very small degrees, this will drastically affect the magnitude of U. This is exactly what we observe in Figure 2.

5.4 Optimal scaling for n_1 and n_2

In this section our goal is to find the ratio n_2 to n_1 which maximizes the performance of the Algorithm 1. For simplicity, as a performance criterion we consider the fraction of correctly identified nodes from top-k in (6):

$$\frac{1}{k}\sum_{j=1}^{k} P_j(n_1) \to \max.$$

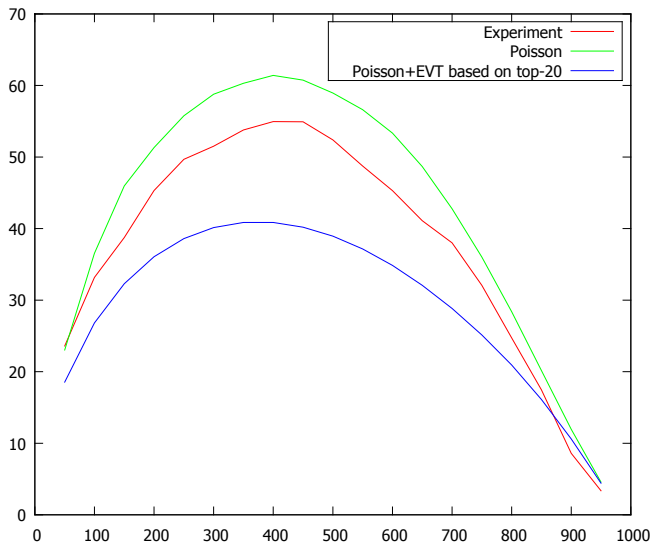


Figure 4: The position of first error as a function of n_2, with $n = 1000$.

We start with analyzing the optimal scaling for n_1. Intuitively, after the first stage of the algorithm, only $O(n_1)$ nodes j will have $S_j > 1$, and thus there is no need to check more than $n_2 = O(n_1)$ nodes in the second stage, which implies that n_1 should grow at least proportionally to n. This is formalized in the next proposition.

PROPOSITION 1. *It is optimal to choose $n = O(n_1)$.*

PROOF. Let J be a randomly chosen node, and J_l, $l = 1, \ldots, n_1$ be independent realizations of J in the first stage of Algorithm 1. Denote by M the maximal number of neighbors that a given API allows to retrieve. The first stage of the algorithm returns a list of candidate nodes, for which we require $S_j > 1$. Observe that the number of such nodes is bounded by

$$U := \frac{1}{2}\sum_{l=1}^{n_1} \max\{M, \text{out-degree}(J_l)\}.$$

Assuming that the out-degrees of each node are independent, we obtain that

$$E(U) = \frac{1}{2} n_1 E(\max\{M, \text{out-degree}(J)\}),$$

$$Var(U) = \frac{1}{4} n_1 \text{Var}(\max\{M, \text{out-degree}(J)\}).$$

Note that the API restriction simplifies the derivation because the variance of $\max\{M, \text{out-degree}(J)\}$ is finite. The formal argument for $M = \infty$ and infinite variance of out-degrees will be similar but requires some more work. Using, e.g. Chernoff bound or Chebyshev bound we obtain that $P(U > E(U)(1+\varepsilon)) \to 0$ as $n_1 \to \infty$. Thus, the number of nodes j with $S_j > 1$ is at most $O(n_1)$ with high probability, so we choose $n_2 = O(n_1)$ which results in $n = O(n_1)$. □

Note that if n is large enough, then the top nodes (first, second, etc.) can be found with very high probability. Figure 1 shows that if $n = 1000$, then for a wide range of n_2 the fraction of correctly identified nodes from top-50 is the

same. As k grows, the optimization becomes much more important. Motivated by this observation, we maximize the value $P_k(n_1)$. We prove the following theorem.

THEOREM 1. *Assume that $k = o(n)$ as $n \to \infty$. The maximizer n_2^* of probability $P_k(n - n_2)$ is close to the maximal root of the equation*

$$\frac{1}{3\gamma k^\gamma} x^{\gamma+1} + x - n = 0, \tag{12}$$

that is,

$$n_2^* = x(1 + o(1)), \quad as \quad k/n_2^* \to 0.$$

If in addition $n_2^* = o(n)$ as $n \to \infty$, then n_2^* can be given in a closed-form asymptotic expression

$$n_2 = (3\gamma k^\gamma n)^{\frac{1}{\gamma+1}} + o(n^{\frac{1}{\gamma+1}}).$$

PROOF. Consider first an extreme regime: $x = \mathrm{O}(k)$. Thus, we exclude the regime $n - x = \mathrm{o}(n)$. Consequently, $n_1 \to \infty$ as $n \to \infty$ and we can apply the following normal approximation

$$\begin{aligned} P_k(n_1) &\approx P\left(N\left(\frac{n_1(F_k - F_{n_2})}{N}, \frac{n_1(F_k + F_{n_2})}{N}\right) > 0\right) \\ &= P\left(N(0,1) > -\sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}}\right). \end{aligned} \tag{13}$$

(A completely formal justification can be given by the Berry-Esseen theorem.) Thus, in order to maximize the above probability, we need to maximize $\sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}}$. From EVT it follows that F_k decays as $k^{-\gamma}$. So, we can maximize

$$\frac{\sqrt{n_1}\left(k^{-\gamma} - n_2^{-\gamma}\right)}{\sqrt{k^{-\gamma} + n_2^{-\gamma}}}. \tag{14}$$

Now if $x = \mathrm{O}(k)$, $\sqrt{n - x} = \sqrt{n}(1 + o(1))$, and the maximization of (14) mainly depends on the remaining term in the product, which is an increasing function of n_2. This suggests that n_2 has to be chosen considerably greater than k. Hence, we proceed assuming the only interesting asymptotic regime where $k = o(n_2)$. In this asymptotic regime, we can simplify (14) as follows:

$$\frac{\sqrt{n - x}\left(k^{-\gamma} - x^{-\gamma}\right)}{\sqrt{k^{-\gamma} + x^{-\gamma}}} =$$

$$\frac{1}{k^{\gamma/2}}\sqrt{n - x}\left(1 - \frac{3}{2}\left(\frac{k}{x}\right)^\gamma\right) + \mathrm{o}\left(\left(\frac{k}{x}\right)^\gamma\right).$$

Next, we differentiate the function

$$f(x) := \sqrt{n - x}\left(1 - \frac{3}{2}\left(\frac{k}{x}\right)^\gamma\right)$$

and set the derivative to zero. This results in equation (12). If we assume further that $n_2^* = \mathrm{o}(n)$, then only the highest order term will remain in (12) and we immediately obtain the following approximation

$$n_2 = (3\gamma k^\gamma n)^{\frac{1}{\gamma+1}} + \mathrm{o}(n^{\frac{1}{\gamma+1}}).$$

□

For example, for $n = 1000$, $k = 100$, and $\gamma = 0.35$ we get $n_2 \approx 570$.

5.5 Sublinear complexity

The normal approximation (13) immediately implies the following proposition.

PROPOSITION 2. *For large enough n_1, the inequality*

$$\sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}} \geqslant x_{1-\varepsilon}$$

guarantees that on average we can find the fraction $1 - \varepsilon$ of top-k nodes in W.

For the inequality in (2) to hold, it is necessary that $\sqrt{n_1}(F_k - F_{n_2})$ is at least of the same order of magnitude as $N\sqrt{F_k + F_{n_2}}$. Moreover, it follows from Proposition 1 that $n = O(n_1)$, and thus the complexity n of the algorithm is defined by n_1. In the theorem below we use the results from Extreme Value Theory to show that n_1 scales sublinearly with N.

Theorem 1, and estimator (11), we can already provide a rough indication of the number of API request we need to use. Indeed, $k > m$, rough estimation with $n - n_2 \approx n$ and $F_k >> F_{n_2}$ gives

$$n \geqslant \frac{N x_{1-\varepsilon}^2 k^{\hat{\gamma}}}{F_m m^\gamma}. \tag{15}$$

For finding top-100 most followed users on Twitter with good precision, this will result in about 5000 of API requests (with $N = 500M$, $m = 20$, $k = 100$, $x_{1-\varepsilon} \approx 2$, $\hat{\gamma} = 2.2$).

For a better result, we may take into account the value of n_2, and substitute the value $n_2 = \left(3k^{\hat{\gamma}} n\hat{\gamma}\right)^{\frac{1}{\gamma+1}}$ obtained in Proposition 2:

$$\begin{aligned} &\frac{k^{-\hat{\gamma}/2}}{2\sqrt{n}}\left(2n - \left(3k^{\hat{\gamma}} n\hat{\gamma}\right)^{\frac{1}{\gamma+1}}\right)\left(1 - \frac{3}{2}\left(3k^{\hat{\gamma}} n\hat{\gamma}\right)^{\frac{-\hat{\gamma}}{\gamma+1}} k^{\hat{\gamma}}\right) \\ &\geqslant x_{1-\varepsilon}\sqrt{\frac{N}{F_m m^\gamma}}. \end{aligned}$$

From (15) we can also already anticipate that n is sublinear in N because $F_m m^\gamma$ grows with N. This argument is formalized in Theorem 2 below.

Notice that, interestingly, the obtained complexity is in terms of the cardinality of W, not V. In particular, this makes the problem of finding popular groups easier than the problem of finding popular users.

THEOREM 2. *If the in-degrees of the nodes are independent realizations of a regularly varying distribution G with exponent $1/\gamma$ as defined in (9), and $F_1 \geqslant F_2 \geqslant \cdots \geqslant F_N$ are their order statistics. Let $(a_N)_{N\geqslant 1}$, $(b_N)_{N\geqslant 1}$ be sequences such that*

$$\lim_{N\to\infty} N(1 - G(a_N x + b_N)) = (1 + \gamma x)^{-1/\gamma}.$$

Then Algorithm 1 finds $(1 - \varepsilon)$ of the top-k nodes with high probability in

$$n_1 = O(N/a_N),$$

of API requests. In particular, n scales sublinearly in N, and

$$\log(n_1) = (1 - \gamma)\log(N).$$

PROOF. For a regularly varying G, Theorem 2.1.1 in [8] can be applied, and thus for any finite m

$$\left(\frac{F_1 - b_N}{a_N}, \cdots \frac{F_m - b_N}{a_N}\right)$$

converges in distribution, as $N \to \infty$, to

$$\left(\frac{E_1^{-\gamma}-1}{\gamma}, \cdots, \frac{(E_1+\cdots+E_m)^{-\gamma}-1}{\gamma}\right),$$

where E_i's are independent exponential random variables with parameter 1. This implies, in particular, that $a_N/b_N = \mathrm{O}(1)$ and that for large enough N and any $\varepsilon > 0$, there exist l_i, u_i such that $P[l_i a_N \leqslant F_i \leqslant u_i a_N] > 1-\varepsilon$. It follows that for fixed k

$$\sqrt{\frac{n_1}{N}}\sqrt{F_k} = O(1)$$

with high probability when $n_1 = O(N/a_N)$, and the first statement of the theorem follows because $k = o(n_2)$ implying that $F_{n_2} = \mathrm{o}(F_k)$. In particular, if G is a Pareto distribution, $1 - G(x) = Cx^{-1/\gamma}$, $x \geqslant x_0$, then

$$a_N = \gamma C^\gamma N^\gamma, \quad b_N = C^\gamma n^\gamma.$$

For a general regularly varying distribution in (9) the slowly varying function will influence a_N but the logarithmic asymptotics of a_N will be still determined by the power law:

$$\log(a_N) = \gamma \log(N),$$

which gives the result. □

6. CONCLUSION

We proposed a randomized algorithm for quick detection of popular entities in large online social networks whose architecture has underlying directed graphs. Examples of social network entities are users and interest groups. We have analyzed the algorithm with respect to three criteria and compared with two baseline methods. Our analysis demonstrates that the algorithm has nonlinear complexity on networks with heavy-tailed in-degree distribution and that the performance of the algorithm is robust with respect to the values of its few parameters. The algorithm outperforms the two baseline methods and has much wider applicability. An important ingredient of our analysis is substantial use of the extreme value theory. The extreme value theory is not so well know in computer science and sociology but appears to be a very useful tool in the analysis of social networks. We feel that our work could be a good reference point for other researchers to start applying EVT in social network analysis. We have validated our theoretical results on two very large online social networks.

We see several extensions of the present work. A top list of popular entities is just one type of properties of social networks. We expect that our approach based on extreme value theory and using referral links can be extended to infer and to analyze other properties such as power law index and the tail, network functions and network motifs, degree-degree correlation. It will be very interesting and useful to develop quick and effective statistical tests to check for network assortativity and presence of heavy tails.

Since our approach requires very small numbers of API accesses, we believe that it will trace well network changes. Of course, a formal justification of the algorithm applicability for dynamic networks is needed.

7. REFERENCES

[1] S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. *Proceedings of the 12-th International World Wide Web Conference*, 2003.

[2] K. Avrachenkov, N. Litvak, D. Nemirovsky, and N. Osipova. Monte carlo methods in pagerank computation: When one iteration is sufficient. *SIAM J. Numer. Anal.*, 45(2):890–904, 2007.

[3] K. Avrachenkov, N. Litvak, D. Nemirovsky, E. Smirnova, and M. Sokol. Quick detection of top-k personalized pagerank lists. In *Proceeding on the 8th Workshop on Algorithms and Models for the Web Graph, WAW 2011*, pages 50–61. Springer, 2011.

[4] K. Avrachenkov, N. Litvak, M. Sokol, and D. Towsley. Quick detection of nodes with large degrees. In *Proceeding on the 9th Workshop on Algorithms and Models for the Web Graph*, pages 54–65. Springer, 2012.

[5] K. Avrachenkov, B. Ribeiro, and D. Towsley. Improving random walk estimation accuracy with uniform restarts. In *Proceeding on the 7th Workshop on Algorithms and Models for the Web Graph, WAW 2010*, pages 98–109. Springer, 2010.

[6] C. Borgs, M. Brautbar, J. Chayes, and S.-H. Teng. A sublinear time algorithm for pagerank computations. *Lecture Notes in Computer Science*, 7323:41–53, 2012.

[7] T. Britton, M. Deijfen, and A. Martin-Löf. Generating simple random graphs with prescribed degree distribution. *J. Stat. Phys.*, 124(6):1377–1397, 2006.

[8] L. De Haan and A. Ferreira. *Extreme value theory.* Springer, 2006.

[9] A. L. M. Dekkers, J. H. J. Einmahl, and L. de Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, 17(4):1833–1855, 1989.

[10] E. Fischer. The art of uninformed decisions: A primer to property testing. *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS*, 75:97–126, 2001.

[11] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlósa. Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. *Internet Mathematics*, 2(3):333–358, 2005.

[12] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. *Proceedings of IEEE INFOCOM'10*, 2010.

[13] O. Goldreich. Combinatorial property testingŮa survey. *Randomization Methods in Algorithm Design, DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 45–60, 1998.

[14] R. Kumar, K. Lang, C. Marlow, and A. Tomkins. Efficient discovery of authoritative resources. *IEEE 24th International Conference on Data Engineering*, pages 1495–1497, 2008.

[15] R. Rubinfeld and A. Shapira. Sublinear time algorithms. *SIAM J. Discrete Math.*, 25(4):1562–1588, 2011.

[16] M. Sudan. Invariance in property testing. *Property Testing: Current Research and Surveys, O. Goldreich, ed., Lecture Notes in Comput. Sci.*, pages 211–227, 2010.

Degree-degree correlations in directed networks with heavy-tailed degrees

Pim van der Hoorn, Nelly Litvak
University of Twente

October 25, 2013

Abstract

In network theory, Pearson's correlation coefficients are most commonly used to measure the degree assortativity of a network. We investigate the behavior of these coefficients in the setting of directed networks with heavy-tailed degree sequences. We prove that for graphs where the in- and out-degree sequences satisfy a power law, Pearson's correlation coefficients converge to a non-negative number in the infinite network size limit. We propose alternative measures for degree-degree correlations in directed networks based on Spearman's rho and Kendall's tau. Using examples and calculations on the Wikipedia graphs for nine different languages, we show why these rank correlation measures are more suited for measuring degree assortativity in directed graphs with heavy-tailed degrees.

Keywords degree assortativity, degree-degree correlations, scale free directed networks, power laws, rank correlations.

1 Introduction

In the analysis of the topology of complex networks a feature that is often studied is the degree-degree correlation, also called degree assortativity of the network. A network has positive degree-degree correlation, is called assortative, when nodes with high degree have a preference to be connected to nodes of similar large degree. When nodes with large degree have a connection preference for nodes with low degree the network is said to have negative degree-degree correlation, it is disassortative. A measure for degree assortativity was first given for undirected networks by Newman [15], which corresponds to Pearson's correlation coefficient of the degrees at the ends of a random edge in the network. A similar definition for directed networks was introduced in [16] and later adopted for analysis of directed complex networks in [18] and [8]. Analysis of the degree-degree correlation has been applied to networks in a variety of scientific fields such as neuroscience, molecular biology, information theory and social network sciences. In [10, 12] degree-degree correlations are used to investigate the structure of collaboration networks of a social news sharing website and Wikipedia discussion pages, respectively. Another

example is [9], where the influence of the phenotopic viability of a family of plants on the degree-degree correlations of their genetic network is investigated. Degree assortativity has also been found to influence several properties of networks. For instance, neural networks with high assortativity seem to behave more efficiently under the influence of noise [7]. Information content has been shown to depend on the absolute value of the degree assortativity [19] and networks with high degree assortativity have been shown to be less stable [4].

Recently it has been shown [13, 14] that for undirected networks of which the degree sequence satisfies a power law distribution with exponent $\gamma \in (1, 3)$, Pearson's correlation coefficient scales with the network size, converging to a non-negative number in the infinite network size limit. Because most real world networks have been reported to be scale free with exponent in $(1, 3)$, c.f. [1, 17, Table II], this could then explain why large networks are rarely classified as disassortative. In the same paper a new measure, corresponding to Spearman's rho [20], has been proposed as an alternative.

In this paper we will extend the analysis in [13] to the setting of directed networks. Here we have to consider four types of degree-degree correlations, depending on the choice for in- or out-degree on either side of an edge. Our message is, similar to that of [13], that Pearson's correlation coefficients are size biased and produce undesirable results, hence we should look for other means to measure degree-degree correlations. Although these results give some insights into the workings of these correlations we still do not fully understand the differences between the four correlation types or what they mean for the structural properties of the network.

We consider networks where the in- and out-degree sequences have a power law distribution. We will give conditions on the exponents of the in- and out-degree sequences for which the assortativity measures defined in [18] and [8] converge to a non-negative number in the infinite network size limit. This result is a strong argument against the use of Pearson's correlation coefficients for measuring degree-degree correlations in such directed networks. To strengthen this argument we also give examples which clearly show that the values given by Pearson's correlation coefficients do not represent the correlation between the degrees, which it is suppose to measure. As an alternative we propose correlation measures based on Spearman's rho [20] and Kendall's tau [11]. These measures are based on the ranking of the degrees rather then their value and hence do not exhibit the size bias observed in Pearson's correlation coefficients. We will give several examples where the difference between these three measures is shown. We also include an example for which one of the four Pearson's correlation coefficients converges to a random variable in the infinite network size limit and therefore will obviously produce uninformative results. Finally we calculate all four degree-degree correlations on the Wikipedia network for nine different languages using all the assortativity measures proposed in this paper.

This paper is structured as follows. In Section 2 we introduce notations. Pearson's correlation coefficients are introduced in Section 3 and a convergence theorem is given for these measures. We introduce the rank measures Spearman's rho and Kendall's tau for degree-degree correlations in Section 4. Example graphs that illustrate the differ-

ence between the three measures are presented in Section 5. Finally the degree-degree correlations for the Wikipedia graphs are presented in Section 6.

2 Definitions and notations

We start with the formal definition of the problem and introduce the notations that will be used throughout this paper.

2.1 Graphs, vertices and degrees

We will denote by $G = (V, E)$ a directed graph with vertex set V and edge set $E \subseteq V \times V$. For an edge $e \in E$, we denote its source by e_* and its target by e^*. With each directed graph we associate two functions $D^+, D^- : V \to \mathbb{N}$ where $D^+(v) := |\{e \in E | e_* = v\}|$ is the out-degree of the vertex v and $D^-(v) := |\{e \in E | e^* = v\}|$ the in-degree. When considering sequences of graphs, we denote by $G_n = (V_n, E_n)$ an element of the sequence $(G_n)_{n\in\mathbb{N}}$. We will further use subscripts to distinguish between the different graphs in the sequence. For instance, D_n^+ and D_n^- will denote the out- and in-degree functions of the graph G_n, respectively.

2.2 Four types of degree-degree correlations

In this paper we are interested in measuring the correlation between the degrees at both sides of an edge. That is, we measure the correlation between two vectors X and Y as function of the edges $e \in E$ corresponding to the degrees of e_* and e^*, respectively. In the undirected case this is called the degree-degree correlation. In the directed setting however, we can consider any combination of the two degree types resulting in four types of degree-degree correlations, illustrated in Figure 1.

From Figure 1 one can already observe some interesting features of these correlations. For instance, in the Out/In correlation the edge that we consider contributes to the degrees on both sides. We will later see that the Out/In correlation actually generalizes the degree-degree correlation in the undirected case. To be more precise, our result for this correlation type generalizes the result obtained in [14] when we transform from the undirected case by making every edge bi-directional.

For the other three correlation types we observe that there is always at least one side where the considered edge does not contribute towards the degree on that side. We will later see that for these correlation types the correlation of the in- and out-degree of a vertex will play a role.

3 Pearson's correlation coefficient

Among all correlation measures, the measure proposed by Newman [15, 16] has been widely used. This measure is the statistical estimator for the Pearson correlation coefficient of the degrees on both sides of a random edge. However, for undirected networks

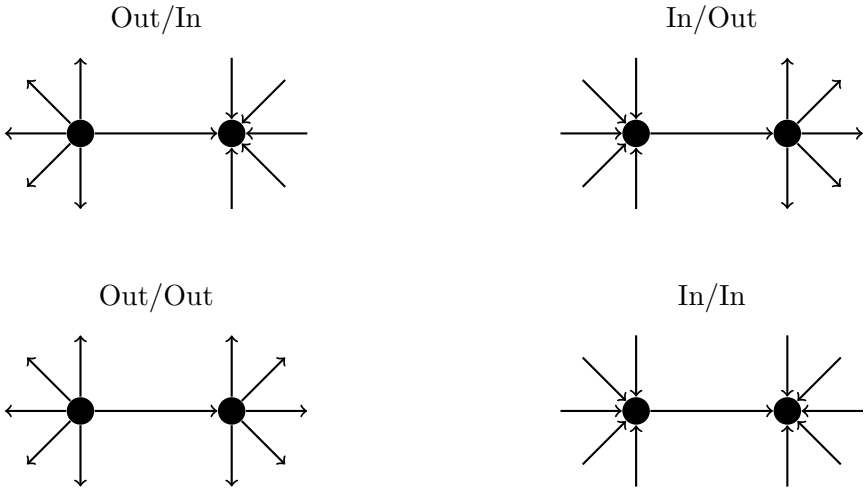


Figure 1: Four degree-degree correlation types

with heavy tailed degrees with exponent $\gamma \in (1,3)$ it was proved [14] that this measure converges, in the infinite size network limit, to a non-negative number. Therefore, in these cases, Pearson's correlation coefficient is not able to correctly measure negative degree-degree correlations. In this section we will extend this result to directed networks proving that also here Pearson's correlation coefficients are not the right tool to measure degree-degree correlations.

Let us consider Pearson's correlation coefficients as in [15, 16], adjusted to the setting of directed graphs as in [8, 18]. This will constitute four formula's which we combine into one. Take $\alpha, \beta \in \{+,-\}$, that is, we let α and β index the type of degree (out- or in-degree). Then we get the following expression for the four Pearson's correlation coefficients:

$$r_\alpha^\beta(G) = \frac{1}{\sigma_\alpha(G)\sigma^\beta(G)} \left(\frac{1}{|E|} \sum_{e \in E} D^\alpha(e_*) D^\beta(e^*) - \frac{1}{|E|^2} \sum_{e \in E} D^\alpha(e_*) \sum_{e \in E} D^\beta(e^*) \right), \quad (1)$$

where

$$\sigma_\alpha(G) = \sqrt{\frac{1}{|E|} \sum_{e \in E} D^\alpha(e_*)^2 - \frac{1}{|E|^2} \left(\sum_{e \in E} D^\alpha(e_*) \right)^2} \text{ and} \quad (2)$$

$$\sigma^\beta(G) = \sqrt{\frac{1}{|E|} \sum_{e \in E} D^\beta(e^*)^2 - \frac{1}{|E|^2} \left(\sum_{e \in E} D^\beta(e^*) \right)^2}. \quad (3)$$

Here we utilize the notations for the source and target of an edge by letting the superscript index denote the specific degree type of the target e^* and the subscript index the

degree type of the source e_*. For instance r_+^- denotes the Pearson correlation coefficient for the Out/In correlation.

It is convenient to rewrite the summations over edges to summations over vertices by observing that

$$\sum_{e\in E} D^\alpha(e_*)^k = \sum_{v\in V} D^+D^\alpha(v)^k$$

and similarly

$$\sum_{e\in E} D^\alpha(e^*)^k = \sum_{v\in V} D^-D^\alpha(v)^k$$

for all $k > 0$. Plugging this into (1)-(3) we arrive at the following definition.

Definition 3.1. *Let $G = (V, E)$ be a directed graph and let $\alpha, \beta \in \{+, -\}$. Then the Pearson's α-β correlation coefficient on G is defined by*

$$r_\alpha^\beta(G) = \frac{1}{\sigma_\alpha(G)\sigma^\beta(G)} \frac{1}{|E|} \sum_{e\in E} D^\alpha(e_*)D^\beta(e^*) - \hat{r}_\alpha^\beta(G), \tag{4}$$

where

$$\hat{r}_\alpha^\beta(G) = \frac{1}{\sigma_\alpha(G)\sigma^\beta(G)} \frac{1}{|E|^2} \sum_{v\in V} D^+(v)D^\alpha(v) \sum_{v\in V} D^-(v)D^\beta(v), \tag{5}$$

$$\sigma_\alpha(G) = \sqrt{\frac{1}{|E|}\sum_{v\in V} D^+(v)D^\alpha(v)^2 - \frac{1}{|E|^2}\left(\sum_{v\in V} D^+(v)D^\alpha(v)\right)^2}, \tag{6}$$

$$\sigma^\beta(G) = \sqrt{\frac{1}{|E|}\sum_{v\in V} D^-(v)D^\beta(v)^2 - \frac{1}{|E|^2}\left(\sum_{v\in V} D^-(v)D^\beta(v)\right)^2}. \tag{7}$$

Just as in the undirected case, c.f. [13, 14], the wiring of the network only contributes to the positive part of (4). All other terms are completely determined by the in- and out-degree sequences. This fact enables us to analyze the behavior of $r_\alpha^\beta(G)$, see Section 3.1. Observe also that in contrast to undirected graphs in the directed case the correlation between the in- and out-degrees of a vertex can play a role, take for instance $\alpha = -$ and $\beta = +$.

Note that in general $r_\alpha^\beta(G)$ might not be well defined, for either $\sigma_\alpha(G)$ or $\sigma^\beta(G)$ might be zero. For example, when G is a directed cyclic graph of arbitrary size. From equations (2) and (3) it follows that $\sigma_\alpha(G)$ and $\sigma^\beta(G)$ are the variance of X and Y, where $X = D^\alpha(e_*)$ and $Y = D^\beta(e^*)$, $e \in E$, with probability $1/|E|$. Thus, $\sigma_\alpha(G) \neq 0$ is only possible if $D^\alpha(v) \neq D^\alpha(w)$ for some $v, w \in V$. Moreover, v and w must have non-zero out-degree for at least one such pair v, w, so that $D^\alpha(v)$ and $D^\alpha(w)$ are counted when we traverse over edges. This argument is formalized in the next lemma, which provides necessary and sufficient conditions so that $\sigma_\alpha(G)$, $\sigma^\beta(G) \neq 0$.

Lemma 3.2. *Let $G = (V, E)$ be a graph and take $\alpha, \beta \in \{+, -\}$. Then the following holds:*

$$\frac{1}{|E|}\left(\sum_{v\in V} D^\alpha(v)D^\beta(v)\right)^2 \leq \sum_{v\in V} D^\alpha(v)D^\beta(v)^2 \tag{8}$$

and strict inequality holds if and only if there exits distinct $v, w \in V$ such that $D^\alpha(v)$, $D^\alpha(w) > 0$ and $D^\beta(v) \neq D^\beta(w)$.

Proof. Recall that $|E| = \sum_{v\in V} D^\alpha(v)$ for any $\alpha \in \{+, -\}$. Then we have:

$$\begin{aligned}
&|E| \sum_{v\in V} D^\alpha(v)D^\beta(v)^2 - \left(\sum_{v\in V} D^\alpha(v)D^\beta(v)\right)^2 \\
&= \sum_{w\in V}\sum_{v\in V\setminus w} D^\alpha(w)D^\alpha(v)D^\beta(v)^2 - D^\alpha(w)D^\beta(w)D^\alpha(v)D^\beta(v) \\
&= \frac{1}{2}\sum_{w\in V}\sum_{v\in V\setminus w} D^\alpha(w)D^\alpha(v)\left(D^\beta(w)^2 - 2D^\beta(w)D^\beta(v) + D^\beta(v)^2\right) \\
&= \frac{1}{2}\sum_{w\in V}\sum_{v\in V\setminus w} D^\alpha(w)D^\alpha(v)\left(D^\beta(w) - D^\beta(v)\right)^2 \geq 0,
\end{aligned}$$

which proves (8). From the last line one easily sees that strict inequality holds if and only if there exits distinct $v, w \in V$ such that $D^\alpha(v)$, $D^\alpha(w) > 0$ and $D^\beta(v) \neq D^\beta(w)$. □

3.1 Convergence of Pearson's correlation coefficients

In this section we will prove that under rather general conditions Pearson's correlation coefficients (4) converges to a non-negative value. We start by recalling the definition of big theta.

Definition 3.3. *Let $f, g : \mathbb{N} \to \mathbb{R}_{>0}$ be positive functions. Then $f = \Theta(g)$ if there exist $k_1, k_2 \in \mathbb{R}_{>0}$ and an $N \in \mathbb{N}$ such that for all $n \geq N$*

$$k_1 g(n) \leq f(n) \leq k_2 g(n).$$

When we have two sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ we write $a_n = \Theta(b_n)$ for $(a_n)_{n\in\mathbb{N}} = \Theta((b_n)_{n\in\mathbb{N}})$.

Next, we will provide the conditions that our sequence of graphs needs to satisfy and prove the result. Then we will motivate the chosen conditions. From here on we denote by $x \vee y$ and $x \wedge y$ the maximum and minimum of x and y, respectively.

Definition 3.4. *For $\gamma_-, \gamma_+ \in \mathbb{R}_{>0}$ we denote by $\mathcal{G}_{\gamma_-\gamma_+}$ the space of all sequences of graphs $(G_n)_{n\in\mathbb{N}}$ with the following properties:*

G1 $|V_n| = n$.

G2 *There exists and $N \in \mathbb{N}$ such that for all $n \geq N$ there exist $v, w \in V_n$ with $D_n^\alpha(v)$, $D_n^\alpha(w) > 0$ and $D_n^\alpha(v) \neq D_n^\alpha(w)$, for all $\alpha \in \{+,-\}$.*

G3 *For all $p, q \in \mathbb{R}_{>0}$,*

$$\sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q = \Theta(n^{p/\gamma_+ \vee q/\gamma_- \vee 1}).$$

G4 *For all $p, q \in \mathbb{R}_{>0}$, if $p < \gamma_+$ and $q < \gamma_-$ then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q := d(p,q) \in (0,\infty).$$

Where the limits are such that for all $a, b \in \mathbb{N}$, $k, m > 1$ with $1/k + 1/m = 1$, $a + p < \gamma_+$ and $b + q < \gamma_-$ we have,

$$d(a,b)^{\frac{1}{m}} d(p,q)^{\frac{1}{k}} > d(\frac{a}{m} + \frac{p}{k}, \frac{b}{m} + \frac{q}{k}).$$

Now we are ready to give the convergence theorem for Pearson's correlation coefficients, Definition 3.1.

Theorem 3.5. *Let $\alpha, \beta \in \{+,-\}$. Then there exists an area $A_\alpha^\beta \subseteq \mathbb{R}^2$ such that for $(\gamma_+, \gamma_-) \in A_\alpha^\beta$ and $(G_n)_{n\in\mathbb{N}} \in \mathcal{G}_{\gamma_-\gamma_+}$,*

$$\lim_{n\to\infty} \hat{r}_\alpha^\beta(G_n) = 0$$

and hence any limit point of $r_\alpha^\beta(G_n)$ is non-negative.

Proof. Let $(G_n)_{n\in\mathbb{N}}$ be an arbitrary sequence of graphs. It is clear that if $\hat{r}_\alpha^\beta(G_n) \to 0$ then any limit point of $r_\alpha^\beta(G_n)$ is non-negative. Therefore we need only to prove the first statement. To this end we define the following sequences,

$$a_n = \frac{1}{|E_n|}\left(\sum_{v\in V_n} D_n^+(v) D_n^\alpha(v)\right)^2, \qquad b_n = \frac{1}{|E_n|}\left(\sum_{v\in V_n} D_n^-(v) D_n^\beta(v)\right)^2,$$
$$c_n = \sum_{v\in V_n} D_n^+(v) D_n^\alpha(v)^2, \qquad d_n = \sum_{v\in V_n} D_n^-(v) D_n^\beta(v)^2,$$

and observe that $\hat{r}_\alpha^\beta(G_n)^2 = a_n b_n/(c_n - a_n)(d_n - b_n)$. Now if $(G_n)_{n\in\mathbb{N}} \in \mathcal{G}_{\gamma_-\gamma_+}$ then because of G2 and Lemma 3.2 there exists an $N \in \mathbb{N}$ such that for all $n \geq N$ we have

$c_n > a_n$ and $d_n > b_n$, so $\hat{r}_\alpha^\beta(G_n)$ is well-defined for all $n \geq N$. Next, using G3, we get that $a_n = \Theta(n^a)$, $b_n = \Theta(n^b)$, $c_n = \Theta(n^c)$ and $d_n = \Theta(n^d)$ for certain constants a, b, c and d, which depend on γ_-, γ_+ and the degree-degree correlation type chosen. Because $\hat{r}_\alpha^\beta(G_n) \to 0$ if and only if $\hat{r}_\alpha^\beta(G_n)^2 \to 0$, we need to find sufficient conditions for which $a_n b_n/(c_n - a_n)(d_n - b_n) \to 0$. It is clear that either $a < c$ and $b_n/(d_n - b_n)$ is bounded or $b < d$ and $a_n/(c_n - a_n)$ is bounded are sufficient. It turns out that this is exactly the case when either $a < c$ and $b \leq d$ or $a \leq c$ and $b < d$. We will do the analysis for the In/Out degree-degree correlation. The analysis for the other three correlation types is similar. Figure 2 shows all four areas A_α^β.

When $\alpha = -$ and $\beta = +$ we get the following constants

$$\begin{aligned} a, b &= 2\left(\frac{1}{\gamma_+} \vee \frac{1}{\gamma_-} \vee 1\right) - 1 \\ c &= \left(\frac{1}{\gamma_+} \vee \frac{2}{\gamma_-} \vee 1\right) \\ d &= \left(\frac{2}{\gamma_+} \vee \frac{1}{\gamma_-} \vee 1\right) \end{aligned}$$

It is clear that when $1 < \gamma_-, \gamma_+ < 2$ then $a < c$ and $b < d$ and hence $\hat{r}_\alpha^\beta \to 0$. Now if $1 < \gamma_- < 2$ and $\gamma_+ \geq 2$ then $a = b = d = 1 < c$. Using G4 we get that $\lim_{n\to\infty} d_n/n = d(2,1)$ and

$$\begin{aligned} \lim_{n\to\infty} \frac{b_n}{n} &= \lim_{n\to\infty} \frac{\left(\sum_{v\in V_n} D_n^-(v) D_n^+(v)\right)^2}{n^2} \frac{n}{|E_n|} \\ &= \lim_{n\to\infty} \left(\frac{\sum_{v\in V_n} D_n^-(v) D_n^+(v)}{n}\right)^2 \left(\frac{\sum_{v\in V_n} D_n^-(v)}{n}\right)^{-1} \\ &= \frac{d(1,1)^2}{d(0,1)} < d(2,1) = \lim_{n\to\infty} \frac{d_n}{n}, \end{aligned}$$

where, for the last part, we again used G4. From this it follows that $b_n/(d_n - b_n)$ is bounded and so $\hat{r}_\alpha^\beta \to 0$. A similar argument applies to the case $\gamma_- \geq 2$ and $1 < \gamma_+ < 2$, where the only difference is that $a = b = c = 1 < d$, hence

$$A_-^+ = \{(x,y) \in \mathbb{R} | 1 < x < 2, \quad y > 1\} \cup \{(x,y) \in \mathbb{R} | 1 < y < 2, \quad x > 1\}.$$

Using similar arguments, we obtain:

$$\begin{aligned} A_+^- &= \{(x,y) \in \mathbb{R}^2 | 1 < x < 3, \quad y > 1\} \cup \{(x,y) \in \mathbb{R}^2 | 1 < y < 3, \quad x > 1\}, \\ A_+^+ &= \{(x,y) \in \mathbb{R}^2 | 1 < x < 3, \quad y > 1\} \text{ and} \\ A_-^- &= \{(x,y) \in \mathbb{R}^2 | 1 < y < 3, \quad x > 1\}. \end{aligned}$$

□

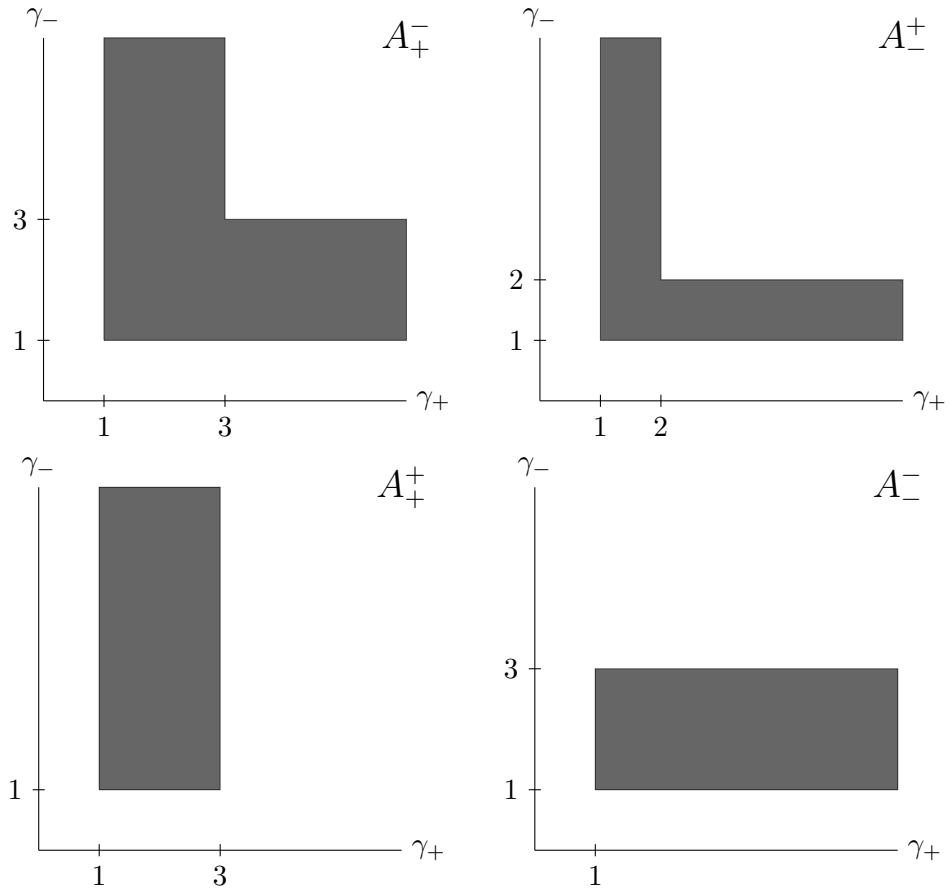


Figure 2

Let us now provide an intuitive explanation for the areas A_α^β, as depicted in Figure 2. The key observation is that due to G3 the terms with the highest power of either D_n^+ or D_n^- will dominate in $\hat{r}_\alpha^\beta(G_n)$. Therefore, if these moments do not exist, then the denominator will grow at a larger rate then the numerator, hence $\hat{r}_\alpha^\beta \to 0$.

Taking $\alpha = + = \beta$, we see that D^- only has terms of order one while D^+ has terms up to order three. This explains why $A_+^+ = \{(x, y) \in \mathbb{R} | 1 < x \leq 3, y > 1\}$. Area A_-^- is then easily explained by observing that the expression for $r_-^-(G)$ is obtained from $r_+^+(G)$ by interchanging D^+ and D^-.

For the Out/In correlation, i.e. $\alpha = +$ and $\beta = -$, we see from equations (5)-(7) that $\hat{r}_+^-(G)$ splits into a product of two terms, each completely determined by either in- or out-degrees,

$$\frac{\frac{1}{|E|}\sum_{v\in V} D^\alpha(v)^2}{\sqrt{\frac{1}{|E|}\sum_{v\in V} D^\alpha(v)^3 - \frac{1}{|E|^2}\left(\sum_{v\in V} D^\alpha(v)^2\right)^2}},$$

with $\alpha \in \{+, -\}$. These terms are of the exact same form as the expression in [13] for

the undirected degree-degree correlation. Because both D^+ and D^- have terms of order three, one sees that

$$A_+^- = \{(x,y) \in \mathbb{R}^2 | 1 < x < 3, \quad y > 1\} \cup \{(x,y) \in \mathbb{R}^2 | 1 < y < 3, \quad x > 1\}.$$

Now take a undirected network and make it directed by replacing each undirected edge with a bi-directional edge. Then $D^+(v) = D^-(v)$ for all $v \in V$ and hence $r_+^-(G)$ equals the expression of equation (3.4) in [13] when we replace D by either D^+ or D^-.

Theorem 3.5 has several consequences. First of all, no matter what mechanism is used for generating networks, if the conditions of the theorem are satisfied then for large enough networks the degree-degree correlations will always be non-negative. This could explain why most large networks are said not to have disassortative degree-degree correlations. In Section 5 we will give examples where this behavior can be observed. Second, if the underlying model that governs the topology of the network is in line with the conditions of the theorem, then one cannot compare networks of different sizes that arise from this model. For in this case, the degree-degree correlation coefficients r_α^β will decrease with the network size.

3.2 Motivation for $\mathcal{G}_{\gamma_-\gamma_+}$

In this section we will motivate Definition 3.4. G1 is easily motivated, for we want to consider infinite network size limits. G2 combined with Lemma 3.2 ensures that from a certain N, $r_\alpha^\beta(G_n)$ will always be well-defined. Conditions G3 and G4 are related to heavy-tailed degree sequences that are modeled using regularly varying random variables.

A random variable X is called regularly varying with exponent γ if $\mathbb{P}(X > t) = L(t)t^{-\gamma}$ for some slowly varying function L, that is $\lim_{t\to\infty} L(tx)/L(t) = 1$ for all x. We write $\mathcal{R}_{-\gamma}$ for the space of all regularly varying random variables with exponent γ. For a regularly varying random variable $X \in \mathcal{R}_{-\gamma}$ we have that $\mathbb{E}[X^p] < \infty$ for all $0 < p < \gamma$.

Through experiments it has been shown that many real world networks, both directed and undirected, have degree sequences whose distribution closely resembles a power law distribution, c.f. Table II of [1] and [17]. Suppose we take two random variables $\mathcal{D}^+ \in \mathcal{R}_{\gamma_+}$, $\mathcal{D}^- \in \mathcal{R}_{\gamma_-}$ and consider, for each n, the degree sequences $(D_n^\pm(v))_{v\in V_n}$ as i.i.d. copies of these random variables. Then for all $0 < p < \gamma_+$ and $0 < q < \gamma_-$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q = \mathbb{E}\left[(\mathcal{D}^+)^p (\mathcal{D}^-)^q\right].$$

Moreover, since $\mathcal{D}^\pm$ is non-degenerate, we have $\mathbb{E}\left[(\mathcal{D}^\pm)^k\right] > \mathbb{E}[\mathcal{D}^\pm]^k$, and thus by taking $d(p,q) = \mathbb{E}[(\mathcal{D}^+)^p(\mathcal{D}^-)^q]$, we get G4 where the second part follows from Hölder's inequality. Although i.i.d. sequences for the in- and out-degrees do not in general constitute a graphical sequence, it is often the case that one can modify this sequence into a graphical sequence preserving i.i.d. properties asymptotically. Consider for example [5], where a directed version of the configuration model is introduced and it is proven that the degree sequences are asymptotically independent.

The property G3 is associated with the scaling of the sums $\sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q$ and is related to the central limit theorem for regularly varying random variables. When we model the degrees as i.i.d. copies of independent regularly varying random variables $\mathcal{D}^+ \in \mathcal{R}_{-\gamma_+}$, $\mathcal{D}^- \in \mathcal{R}_{-\gamma_-}$ and take $p \geq \gamma_+$ or $q \geq \gamma_-$ then $\sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q$ is in the domain of attraction of a γ-stable random variable $S(\gamma)$, where $\gamma = (\gamma_+/p \wedge \gamma_-/q)$, c.f. [6]. This means that

$$\frac{1}{a_n} \sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q \xrightarrow{d} S(\gamma_+/p \wedge \gamma_-/q), \quad \text{as } n \to \infty \tag{9}$$

for some sequence $a_n = \Theta(n^{q/\gamma_- \vee p/\gamma_+})$, where $\xrightarrow{d}$ denotes convergence in distribution. Informally, one could say that $\sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q$ scales as $n^{q/\gamma_- \vee p/\gamma_+}$ when either the p or q moment does not exist and as n when both moments exist, hence, $\sum_{v\in V_n} D_n^+(v)^p D_n^-(v)^q$ scales as $n^{q/\gamma_- \vee p/\gamma_+ \vee 1}$, which is what G3 states. For completeness we include the next lemma, which shows that (9) implies that G3 holds with high probability. We remark that although this motivation is based on results where the regularly varying random variables are assumed to be independent the dependent case can be included. For this one then needs to adjust the scaling parameters in G3 for the specified dependence.

Lemma 3.6. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of positive random variables such that*

$$\frac{X_n}{a_n} \xrightarrow{d} X, \quad \text{as } n \to \infty,$$

for some sequence $(a_n)_{n\in\mathbb{N}}$ and positive random variable X. Then for each $0 < \varepsilon < 1$, there exists an $N_\varepsilon \in \mathbb{N}$ and $\kappa_\varepsilon \geq \ell_\varepsilon > 0$ such that for all $n \geq N_\varepsilon$

$$\mathbb{P}(\ell_\varepsilon a_n \leq X_n \leq \kappa_\varepsilon a_n) \geq 1 - \varepsilon.$$

Proof. Let $0 < \varepsilon < 1$ and take $\delta > 0$, $0 < \ell \leq \kappa$ such that

$$\mathbb{P}(\ell \leq X \leq \kappa) \geq 1 - \varepsilon + \delta.$$

Then, because $X_n/a_n \xrightarrow{d} X$ as $n \to \infty$, there exists an $N \in \mathbb{N}$ such that for all $n \geq N$,

$$|\mathbb{P}(\ell \leq X \leq \kappa) - \mathbb{P}(\ell a_n \leq X_n \leq \kappa a_n)| < \delta.$$

Now we get for all $n \geq N$,

$$1 - \varepsilon + \delta - \mathbb{P}(\ell a_n \leq X_n \leq \kappa a_n) \leq \mathbb{P}(\ell \leq X \leq \kappa) - \mathbb{P}(\ell a_n \leq X_n \leq \kappa a_n) \leq \delta,$$

hence $\mathbb{P}(\ell a_n \leq X_n \leq \kappa a_n) \geq 1 - \varepsilon$. □

4 Rank correlations

In this section we consider two other measures for degree-degree correlations, Spearman's rho [20] and Kendall's tau [11], which are based on the rankings of the degrees rather then their actual value. We will define these correlation measures and argue that they do not have unwanted behavior as we observed for Pearson's correlation coefficients. We will later use examples to enforce this argument and show that Spearman's rho and Kendall's tau are better candidates for measuring degree-degree correlations.

4.1 Spearman's rho

Spearman's rho [20] is defined as the Pearson correlation coefficient of the vector of ranks. Let $G = (V, E)$ be a directed graph and $\alpha, \beta \in \{+, -\}$. In order to adjust the definition of Spearman's rho to the setting of directed graphs we need to rank the vectors $(D^\alpha(e_*))_{e\in E}$ and $(D^\beta(e^*))_{e\in E}$. These will, however, in general have many tied values. For instance, suppose that $D^\alpha(v) = m$ for some $v \in V$, then edges $e \in E$ with $e_* = v$ satisfy $D^\alpha(e_*) = D^\alpha(v)$. Therefore, we will encounter the value $D^\alpha(v)$ at least m times in the vector $(D^\alpha(e_*))_{e\in E}$. We will consider two strategies for resolving ties: uniformly at random (Section 4.1.1), and using an average ranking scheme (Section 4.1.2).

4.1.1 Resolving ties uniformly at random

Given a sequence $\{x_i\}_{1\le i\le n}$ of distinct elements in $\mathbb{R}$ we denote by $R(x_j)$ the rank of x_j, i.e. $R(x_j) = |\{i|x_i \ge x_j\}|$, $1 \le j \le n$. The definition of Spearman's rho in the setting of directed graphs is then as follows.

Definition 4.1. *Let $G = (V, E)$ be a directed graph, $\alpha, \beta \in \{+, -\}$ and let $(U_e)_{e\in E}$, $(W_e)_{e\in E}$ be i.i.d. copies of independent uniform random variables U and W on $(0, 1)$, respectively. Then we define the α-β Spearman's rho of the graph G as*

$$\rho_\alpha^\beta(G) = \frac{12\sum_{e\in E} R^\alpha(e_*)R^\beta(e^*) - 3|E|(|E|+1)^2}{|E|^3 - |E|}, \tag{10}$$

where $R^\alpha(e_*) = R(D^\alpha(e_*) + U_e)$ and $R^\beta(e^*) = R(D^\beta(e^*) + W_e)$.

From (10) we see that the negative part of $\rho_\alpha^\beta(G)$ depends only on the number of edges

$$\frac{3(|E|+1)^2}{(|E|^2-1)} = 3 + \frac{6|E|+4}{|E|^2-1},$$

while for $r_\alpha^\beta(G)$ it depended on the values of the degrees, see Definition 3.1. When $(G_n)_{n\in\mathbb{N}} \in \mathcal{G}_{\gamma_+,\gamma_-}$, with $\gamma_+, \gamma_- > 1$ then it follows that $|E_n| = \theta(n)$ hence $3 + (6|E| + 4)/(|E|^2-1) \to 3$, as $n \to \infty$. Therefore we see that the negative contribution will always be at least 3 and so $\rho_\alpha^\beta(G_n)$ does not in general converge to a non-negative number while $r_\alpha^\beta(G_n)$ does.

When calculating $\rho_\alpha^\beta(G)$ on a graph G one has to be careful, for each instance will give different ranks of the tied values. This could potentionally give rise to very different results among several instances, see Section 5.1.2 for an example. Therefore, in experiments, we will take an average of $\rho_\alpha^\beta(G)$ over several instances of the uniform ranking.

4.1.2 Resolving ties with average ranking

A different approach for resolving ties is to assign the same average rank to all tied values. Consider, for example, the sequence $(1, 2, 1, 3, 3)$. Here the two values of 3 have ranks 1 and 2, but instead we assign the rank $3/2$ to both of them. With this scheme the sequence of ranks becomes $(9/2, 3, 9/2, 3/2, 3/2)$. This procedure can be formalized as follows.

Definition 4.2. *Let $(x_i)_{1\leq i\leq n}$ be a sequence in $\mathbb{R}$ then we define the average rank of an element x_i as*

$$\overline{R}(x_i) = |\{j|x_j > x_i\}| + \frac{|\{j|x_j = x_i\}| + 1}{2}.$$

Observe that in the above definition the total average rank is preserved: $\sum_{i=1}^n \overline{R}(x_i) = n(n+1)/2$. The difference with resolving ties uniformly at random is that we in general do not know $\sum_{i=1}^n \overline{R}(x_i)^2$, for this depends on how many ties we have for each value. We now define the average Spearman's rho of graphs as follows.

Definition 4.3. *let $G = (V, E)$ be a directed graph, $\alpha, \beta \in \{+, -\}$ and denote by $\overline{R}^\alpha(e_*)$ and $\overline{R}^\beta(e^*)$ the average ranks of $D^\alpha(e_*)$ among $(D^\alpha(e_*))_{e\in E}$ and $D^\beta(e^*)$ among $(D^\beta(e^*))_{e\in E}$, respectively. Then we define the average α-β Spearman's rho of the graph G by*

$$\overline{\rho}_\alpha^\beta(G) = \frac{4\sum_{e\in E} \overline{R}^\alpha(e_*)\overline{R}^\beta(e^*) - |E|(|E|+1)^2}{\overline{\sigma}_\alpha(G)\overline{\sigma}^\beta(G)}, \tag{11}$$

where

$$\overline{\sigma}_\alpha(G) = \sqrt{4\sum_{e\in E} \overline{R}^\alpha(e_*)^2 - |E|(|E|+1)^2}$$

and

$$\overline{\sigma}^\beta(G) = \sqrt{4\sum_{e\in E} \overline{R}^\beta(e^*)^2 - |E|(|E|+1)^2}.$$

4.2 Kendall's Tau

Another common rank correlation measure is Kendall's Tau [11], which measures the weighted difference between the number of concordant and disconcordant pairs of the

joint observations $(x_i, y_i)_{1\le i\le n}$. More precisely, a pair (x_i, y_i) and (x_j, y_j) of joint observations is concordant if $x_i < x_j$ and $y_i < y_j$ or if $x_i > x_j$ and $y_i > y_j$. They are called disconcordant if $x_i < x_j$ and $y_i > y_j$ or if $x_i > x_j$ and $y_i < y_j$.

Definition 4.4. *Let $G = (V, E)$ be a directed graph, $\alpha, \beta \in \{-, +\}$ and denote by $\mathcal{N}_c$ and $\mathcal{N}_d$, respectively, the number of concordant and disconcordant pairs among $\left(D^\alpha(e_*), D^\beta(e^*)\right)_{e\in E}$. Then we define the α-β Kendall's tau of G by*

$$\tau_\alpha^\beta(G) = \frac{2(\mathcal{N}_c - \mathcal{N}_d)}{|E|(|E|-1)}.$$

It might seem at first that τ does not suffer from ties. However, note that the numerator of τ includes only strictly (dis)concordant pairs, while the denominator is equal to the number of all possible pairs, irregardless of the presence of ties. Hence, when the number of ties is large, the denominator may become much larger than the numerator resulting in small, even vanishing in the graph size limit, values of τ_α^β. We will provide such example in Section 5. Since, as discussed above, the sequences $(D^\alpha(e_*))_{e\in E}$ and $\left(D^\beta(e^*)\right)_{e\in E}$ naturally have a large number of ties, we cannot expect $\tau_\alpha^\beta(G)$ to take very large (positive or negative) values.

5 Bridge graph example

In this section we will provide a sequences of graphs to illustrate the difference between the four correlation measures in directed networks. We start with a deterministic sequence and will later adapt this to a randomized sequence using regularly varying random variables.

5.1 A deterministic in-out bridge graph

Let $k, m \in \mathbb{N}_{>0}$, then we define the bridge graph $G(k, m) = (V(k, m), E(k, m))$, displayed in Figure 3a, as follows:

$$V(k,m) = v \cup w \cup \bigcup_{i=1}^{k} v_i \cup \bigcup_{j=1}^{m} w_j, \quad E(k,m) = g \cup \bigcup_{i=1}^{k} e_i \cup \bigcup_{j=1}^{m} f_j, \text{ where}$$

$$e_i = (v_i, v),\ f_j = (w, w_j) \text{ and } g = (v, w).$$

It follows that $|E(k, m)| = m + k + 1$. For the degrees of $G(k, m)$ we have:

$$\begin{aligned}
&D^+(v_i) = 1, && D^-(v_i) = 0, && \text{for all } 1 \le i \le k;\\
&D^+_{n,a}(w_j) = 0, && D^-_{n,a}(w_j) = 1, && \text{for all } 1 \le j \le m;\\
&D^+(v) = 1, && D^-(v) = k,\\
&D^+(w) = m, && D^-(w) = 1.
\end{aligned}$$

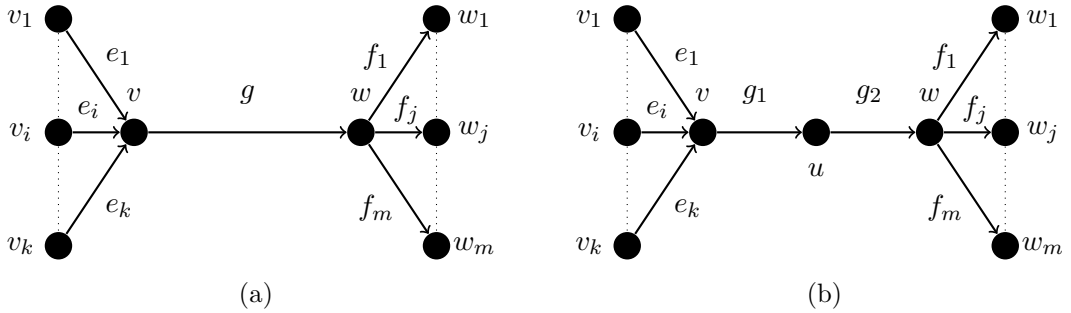


Figure 3: A graphical representation of the graphs $G(k,m)$ (a) and $\hat{G}(k,m)$ (b).

Looking at the scatter plot of $(D^-(e_*), D^+(e^*))_{e\in E(k,m)}$, Figure 4a, we see that the point (k,m) contributes towards a positive correlations while the points $(0,1)$ and $(1,0)$ contribute towards a negative correlation. Hence, depending on how much weight we put on each of these points we could argue equally well that this graph has positive or negative In/Out correlation. We can however extend the in-out bridge graph to a graph for which we do have a clear expectation for the In/Out correlation.

We define the disconnected in-out bridge graph $\hat{G}(k,m) = (\hat{V}(k,m), \hat{E}(k,m))$ from $G(k,m)$ by adding a vertex u and replacing the edge $g = (v,w)$ by the edges $g_1 = (v,u)$ and $g_2 = (u,w)$, see Figure 3b. In this graph the node with the largest in-degree, v, is connected to node u, of out-degree 1. Similarly u, which has in-degree 1, is connected to the node with the highest out-degree, w. Therefore we would expect a negative In/Out correlation. This intuition is supported by the scatter plot of $(D^+(e^*), D^-(e_*))_{e\in\hat{E}(k,m)}$, Figure 4b.

Now consider for a fixed $a \in \mathbb{N}$ the sequence of graphs $G_n^a := G(n, an)$ and $\hat{G}_n^a := \hat{G}(n, an)$. Then, following the above reasoning we would expect that any In/Out correlation measure of $\hat{G}_n^a$ would converge to -1.

In Sections 5.1.1 – 5.1.3 we will show that $\lim_{n\to\infty} r_-^+(\hat{G}_n^a) = 0$ while the other three measures indeed yield negative results. Furthermore, we show that $\lim_{n\to\infty} r_-^+(G_n^a) = 1$ while $\lim_{n\to\infty} \overline{\rho}_-^+(G_n^a) = -1$ reflecting the two possibilities for the In/Out correlation represented in the scatter plot, Figure 4a.

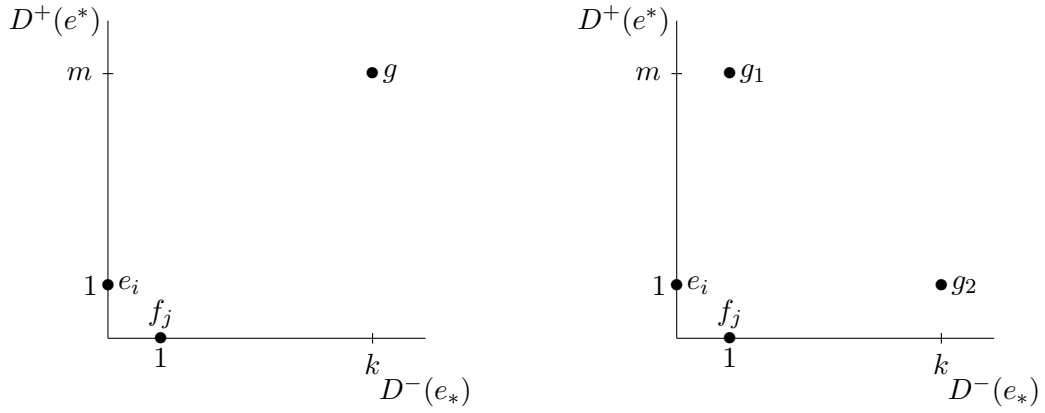


Figure 4: The scatter plots for the degrees of (a) $G(k,m)$ and (b) $\hat{G}(k,m)$.

5.1.1 Pearson In/Out correlation

We start with the graph G_n^a. Basic calculations yield that

$$\sum_{e\in E_n^a} D^-(e_*)D^+(e^*) = an^2, \tag{12}$$

$$\sum_{v\in V_n^a} D^-(v)D^+(v) = (1+a)n, \tag{13}$$

$$\sum_{v\in V_n^a} D^-(v)^2D^+(v) = n^2+an, \tag{14}$$

$$\sum_{v\in V_n^a} D^-(v)D^+(v)^2 = n+a^2n^2, \tag{15}$$

hence, using (6) and (7), we obtain:

$$\begin{aligned}|E_n^a|\sigma_-(G_n^a) &= \sqrt{((1+a)n+1)(n^2+an)-(1+a)^2n^2}\\ &= \sqrt{(1+a)n^3-(n-1)an}\end{aligned}$$

and

$$\begin{aligned}|E_n^a|\sigma^+(G_n^a) &= \sqrt{((1+a)n+1)(n+a^2n^2)-(1+a)^2n^2}\\ &= \sqrt{(1+a)n^3-(an-1)n}.\end{aligned}$$

When we plug this into (4) with $\alpha=-$ and $\beta=+$ we get

$$\begin{aligned}r_-^+(G_n^a) &= \frac{|E_n^a|an^2-(1+a)^2n^2}{|E_n^a|\sigma_\alpha(G_n^a)|E_n^a|\sigma^\beta(G_n^a)}\\ &= \frac{a(1+a)n^3-(a^2+a+1)n^2}{a\sqrt{(1+a)n^3-(n-1)an}\sqrt{(1+a)n^3-(an-1)n}}.\end{aligned} \tag{16}$$

From (16) it follows that if $a \in \mathbb{N}$ is fixed, then $\lim_{n\to\infty} r_-^+(G_n^a) = 1$, thus $r_-^+(G_n^a)$ in fact reflects the connection between v and w where the point (n, an) in the scatter plot received the most mass. However, when we turn to $\hat{G}_n^a$ we get a less expected result. Splitting the edge g in two adds one to equations (13)-(15), while equation (12) becomes $(a+1)n$ which is linear in n instead of quadratic. Because all other terms keep their scale with respect to n we easily deduce that for a fixed $a \in \mathbb{N}$, $\lim_{n\to\infty} r_-^+(\hat{G}_n^a) = 0$. This is undesirable for we would expect any correlation measure on $\hat{G}_n^a$ to converge to -1.

5.1.2 Spearman In/Out correlation

We start by calculation $\overline{\rho}_-^+(G_n^a)$. For this observe that by (11) and the definition of G_n^a we have that,

$$\begin{aligned}
&\overline{R}^+((e_i)^*) = 1 + \frac{n+1}{2}, && \overline{R}^-((e_i)_*) = an + 1 + \frac{n+1}{2};\\
&\overline{R}^+((f_j)^*) = n + 1 + \frac{an+1}{2}, && \overline{R}^-((f_j)_*) = 1 + \frac{an+1}{2};\\
&\overline{R}^+(g^*) = 1, && \overline{R}^-(g_*) = 1.
\end{aligned}$$

After some basic calculations we get

$$\overline{\rho}_-^+(G_n^a) = \frac{-(a^2+a)n^3 + (a+1)^2n^2 + (a+1)n}{(a^2+a)n^3 + (a+1)^2n^2 + (a+1)n} \to -1 \quad \text{as } n \to \infty.$$

This result is in striking contrast to the one for $r_-^+(G_n^a)$. Indeed, $\overline{\rho}_-^+$ places all the weight on the points $(0,1)$ and $(1,0)$. However, based on the scatter plot, see Figure 4a, both results could be plausible.

Let us now compute $\overline{\rho}_-^+(\hat{G}_n^a)$. For the rankings we have

$$\begin{aligned}
&\overline{R}^+((e_i)^*) = 2 + \frac{n}{2}, && \overline{R}^-((e_i)_*) = an + 2 + \frac{n+1}{2};\\
&\overline{R}^+((f_j)^*) = n + 2 + \frac{an+1}{2}, && \overline{R}^-((f_j)_*) = 2 + \frac{an}{2};\\
&\overline{R}^+((g_1)^*) = 2 + \frac{n}{2}, && \overline{R}^-((g_1)_*) = 1;\\
&\overline{R}^+((g_2)^*) = 1, && \overline{R}^-((g_2)_*) = 2 + \frac{an}{2}.
\end{aligned}$$

Filling this into equation (11) we get

$$\overline{\rho}_-^+(\hat{G}_n^a) = \frac{-(a^2+a)n^3 - (a^2+a)n^2 + (a+1)n - 2}{\bar{\sigma}_-(\hat{G}_n^a)\bar{\sigma}^+(\hat{G}_n^a)},$$

where

$$\begin{aligned}
&\bar{\sigma}_-(\hat{G}_n^a) = \sqrt{(a^2+a)n^3 + (a^2+4a+2)n^2 + (3a+4)n - 2} \text{ and}\\
&\bar{\sigma}^+(\hat{G}_n^a) = \sqrt{(a^2+a)n^3 + (2a^2+4a+1)n^2 + (4a+3)n + 2}.
\end{aligned}$$

Because

$$\lim_{n\to\infty}\frac{1}{n^3}\bar{\sigma}_-(\hat{G}_n^a)\bar{\sigma}^+(\hat{G}_n^a) = (a^2+a)$$

it follows that

$$\lim_{n\to\infty}\overline{\rho}_-^+(\hat{G}_n^a) = \lim_{n\to\infty}\frac{1/n^3}{1/n^3}\frac{-(a^2+a)n^3-(a^2+a)n^2+(a+1)n-2}{\bar{\sigma}_-(\hat{G}_n^a)\bar{\sigma}^+(\hat{G}_n^a)} = -1,$$

which equals $\lim_{n\to\infty}\rho(G_n^a)$. We have already argued that based on the graph and the scatter plot we would expect negative In/Out correlation for the sequence $(\hat{G}_n^a)_{n\in\mathbb{N}}$. This result is in agreement with what we would expect, while $r_-^+(\hat{G}_n^a)$ converges to 0 as $n\to\infty$.

Now we turn to $\rho_-^+(G_n^a)$. We will show that the choice of ranking of the tied values can have a great effect on the outcome of the In/Out correlation. In this example we will pick two rankings, one will yield a positive correlation while the other will give a negative correlation.

It is clear from the definition of G_n^a that the in- and out-degrees of all e_i are the same and similar for f_j. Let us now impose the following ranking of the vectors $(D^+(e^*))_{e\in E_n^a}$ and $(D^-(e_*))_{e\in E_n^a}$:

$$\begin{aligned}
&R^+((e_i)^*) = an+i, && R^-((e_i)_*) = i, && \text{for all } 1\le i\le n;\\
&R^+((f_j)^*) = j, && R^-((f_j)_*) = n+j, && \text{for all } 1\le j\le an;\\
&R^+(g^*) = 1+(a+1)n, && R^-(g_*) = 1+(a+1)n.
\end{aligned}$$

Here we ordered the ties by the order of their indices. We calculate that

$$\rho_-^+(G_n^a) = \frac{(a^3-3a^2-3a+1)n^3+3(a+1)^2n^2+2(a+1)n}{(a^3+3a^2+3a+1)n^3+3(a+1)^2n^2+2(a+1)n}. \tag{17}$$

Now let us now order $(D^+(e^*))_{e\in E_n^a}$ and $(D^-(e_*))_{e\in E_n^a}$ as follows:

$$\begin{aligned}
&R^+((e_i)^*) = (a+1)n+1-i, && R^-((e_i)_*) = i, && \text{for all } 1\le i\le n;\\
&R^+((f_j)^*) = an+1-j, && R^-((f_j)_*) = n+j, && \text{for all } 1\le j\le an;\\
&R^+(g^*) = 1+(a+1)n, && R^-(g_*) = 1+(a+1)n.
\end{aligned}$$

This order differs from the first one only on the vector $(D^+(e^*))_{e\in E_n^a}$, where we now ordered the ties based on the reversed order of their indices. Here we get, after some calculations,

$$\rho_-^+(G_n^a) = \frac{-(a+1)^3n^3+3(a+1)^2n^2+2(a+1)n}{(a+1)^3n^3+3(a+1)^2n^2+2(a+1)n} \tag{18}$$

When we compare (18) with (17) we see that for the former $\lim_{n\to\infty}\rho_-^+(G_n^a) = -1$ for all $a\in\mathbb{N}$ while for the latter we have $\lim_{n\to\infty}\rho_-^+(G_n^a) = (a^3-3a^2-3a+1)/(a+1)^3$. This means that increasing a will actually increase the limit of (17), which becomes positive when $a\ge 4$. This indicates what was already mentioned in Section 4.1.1, that changing the order of the ties can have a large impact on the value of $\rho_\alpha^\beta(G)$.

5.1.3 Kendall's Tau In/Out correlation

The last correlation measure we compute is Kendall's Tau. In order to do this we need to determine the number of concordant and disconcordant pairs. Starting with G_n^a, we observe that we have three kinds of joint observations, namely

$$\begin{aligned} I &: \left(D^-(e_{i*}), D^+(e_i^*)\right), \\ II &: \left(D^-(f_{j*}), D^+(f_j^*)\right) \text{ and} \\ III &: \left(D^-(g_*), D^+(g^*)\right). \end{aligned}$$

The combinations I and III, and II and III are concordant while I and II are disconcordant. From this it follows that $\mathcal{N}_c = (a+1)n$ while $\mathcal{N}_d = an^2$. Hence we get, see Definition 4.4.

$$\tau_-^+(G_n^a) = \frac{2(a+1)n - 2an^2}{(a+1)^2n^2 + (a+1)n},$$

which gives $\lim_{n\to\infty} \tau_-^+(G_n^a) = -\frac{2a}{(a+1)^2}$.

For the graph $\hat{G}_n^a$ we have four kinds of joint observations:

$$\begin{aligned} I &: \left(D^-(e_{i*}), D^+(e_i^*)\right), \\ II &: \left(D^-(f_{j*}), D^+(f_j^*)\right), \\ III &: \left(D^-(g_{1*}), D^+(g_1^*)\right) \text{ and} \\ IV &: \left(D^-(g_{2*}), D^+(g_2^*)\right). \end{aligned}$$

Again the combinations I and II are disconcordant, while now I and III, and II and IV are concordant. Therefore we get $\mathcal{N}_c = (a+1)n$ and $\mathcal{N}_d = an^2$, hence $\lim_{n\to\infty} \tau_-^+(G_n^a) = -\frac{2a}{(a+1)^2}$ which equals the limit for $\tau_-^+(G_n^a)$.

Note that $\lim_{n\to\infty} \tau_-^+(G_n^a)$ decreases when we increase a. This is because the number of tied values among the degrees increases with a. We already mentioned that τ_α^β gives smaller values when more ties are involved. Here this behavior is clearly present.

5.2 A collection of random In/Out bridge graphs

Let us now consider a collection of In/Out bridge graphs $G(W, Z)$ as defined in Section 5.1, where the values of W and Z are integer regularly varying random variables.

Let $X, Y \in \mathcal{R}_{-\gamma}$ be independent and integer valued and fix $a \in \mathbb{R}_{>0}$. For each $n \in \mathbb{N}$ take $(X_i)_{1\leq i\leq n}$ and $(Y_i)_{1\leq i\leq n}$ to be i.i.d. copies of X and Y, respectively, and define $W_i = X_i + Y_i$ and $Z_i = \lfloor X_i + aY_i \rfloor$. Then we define the graph $\mathcal{G}_n^a$ as the disconnected collection of the graphs $(G(W_i, Z_i))_{1\leq i\leq n}$. We will calculate $r_-^+(\mathcal{G}_n^a)$ and prove that it converges to a random variable, which can have support on $(\varepsilon, 1)$ for a specific choice of a.

Using the calculations in Section 5.1.1 we obtain:

$$\sum_{e\in E_n^a} D^-(e_*)D^+(e^*) = \sum_{i=1}^n \left(X_i^2 + aY_i^2 + (1+a)X_iY_i\right),$$

$$\sum_{v\in V_n^a} D^-(v)D^+(v) = \sum_{i=1}^n \left(2X_i + (1+a)Y_i\right),$$

$$\sum_{v\in V_n^a} D^-(v)^2D^+(v) = \sum_{i=1}^n \left(X_i^2 + Y_i^2 + 2X_iY_i + X_i + aY_i\right),$$

$$\sum_{v\in V_n^a} D^-(v)D^+(v)^2 = \sum_{i=1}^n \left(X_i^2 + a^2Y_i^2 + 2aX_iY_i + X_i + Y_i\right) \text{ and}$$

$$|E_n^a| = \sum_{i=1}^n \left(2X_i + (1+a)Y_i + 1\right).$$

By the stable limit law we have a sequence $(a_n)_{n\in\mathbb{N}}$ such that

$$\frac{1}{a_n}\sum_{i=1}^n X_i^2 \xrightarrow{d} S_X \quad \text{and} \quad \frac{1}{a_n}\sum_{i=1}^n Y_i^2 \xrightarrow{d} S_Y \quad \text{as } n\to\infty,$$

where S_X and S_Y are stable random variables. Further, due to Lemma 2.2 in [13] we have

$$\frac{1}{a_n}\sum_{i=1}^n X_iY_i \xrightarrow{d} 0, \quad \frac{1}{a_n}\sum_{i=1}^n X_i \xrightarrow{d} 0 \quad \text{and} \quad \frac{1}{a_n}\sum_{i=1}^n Y_i \xrightarrow{d} 0 \quad \text{as } n\to\infty.$$

Combining this we get

$$\frac{1}{\sqrt{a_n}}\sigma_-(\mathcal{G}_n^a) \xrightarrow{d} \sqrt{S_X+S_Y}, \quad \frac{1}{\sqrt{a_n}}\sigma^+(\mathcal{G}_n^a) \xrightarrow{d} \sqrt{S_X+a^2S_Y} \quad \text{as } n\to\infty,$$

and hence

$$r_-^+(\mathcal{G}_n^a) \xrightarrow{d} \frac{S_X+aS_Y}{\sqrt{S_X+S_Y}\sqrt{S_X+a^2S_Y}} \quad \text{as } n\to\infty,$$

which has support on $(0,1)$. Now, take $0<\varepsilon\le 1$ and consider the function $f(x) : (0,\infty)\to\mathbb{R}$ defined as

$$f(x) = \frac{1+ax}{\sqrt{1+x}\sqrt{1+a^2x}}.$$

This function attains its minimum in $1/a$ and by solving $f(1/a)=\varepsilon$ for a we get that for

$$a = \frac{2-\varepsilon^2 \pm \sqrt{1-\varepsilon}}{\varepsilon^2}$$

this minimum equals ε. If we now introduce the random variable $T = S_Y/S_X$ we see that for a defined as above $\frac{1+aT}{\sqrt{1+T}\sqrt{1+a^2T}}$ has support contained in $(\varepsilon,1)$.

This example shows that Pearson's correlation coefficients r_α^β can converge to a non-negative random variable in the infinite size network limit. This behavior is undesirable for if we consider two instances of the same model $\mathcal{G}_n^a$ then the values of r_-^+ will be random and hence could be very far apart. Therefore r_-^+ is not suitable for measuring the In/Out correlation if we would like to find one number (population value) that characterizes the In/Out correlation in this model.

6 Experiments

In this section we present experimental results for the degree-degree correlations introduced in Sections 3 and 4. For the calculations we used the WebGraph framework [2, 3] and the fastutil package from The Laboratory for Web Algorithmics (LAW) at the Universit degli studi di Milano, `http://law.di.unimi.it`. The calculations where done on the Wikepedia graphs, `http://wikipedia.org`, of nine different languages, obtained from the LAW dataset database. For each Wikipedia graph we calculated all four degree-degree correlations using the four measures introduced in this paper.

In an attempt to quantify the results we compared them to a randomized setting. For this we did 20 reconfigurations of the degree sequences of each graph, using the scheme decribed in Section 3 of [5]. More precisely, we used the *erased directed configuration model*. In this scheme we first assign to each vertex v, $D^+(v)$ outbound stubs and $D^-(v)$ inbound stubs. Then we randomly select an available outbound stub and combine it with a inbound stub, selected uniformly at random from all available inbound stubs, to make an edge. When this edge is a selfloop we remove it. When we end up with multiple edges between two vertices we combine them into one edge. Proposition 3.7 of [5] now tells us that the distribution of the degrees of the resulting simple graph will, with high probability, be the the same as the original distribution. For each of these reconfigurations, all correlations where calculated using all four measures and then for each correlation type and measure we took the average. The results are presented in Table 1.

The first observation is that for each Wikipedia graph and correlation type, the measures ρ, $\overline{\rho}$ and τ have the same sign while r in many cases has a different sign. Furthermore, there are many cases where the absolute value of the three rank correlations is at least an order of magnitute larger than that of Pearson's correlation coefficients. See for instance the Out/In correlations for DE, EN, FR and NL or the In/Out correlation for KO and RU.

These examples illustrate the fact that Pearson's correlation coefficients are scaled down by the high variance in the degree sequences which in turn gave rise to Theorem 3.5, while the rank correlations do not have this deficiency. Another interesting observation is that the values for ρ and $\overline{\rho}$ are almost in full agreement with each other. This would then suggest that one could freely change between these two when calculating degree-degree correlations. Because for ρ both the average and the variance are known upfront, it is computationally easier than $\overline{\rho}$ while the latter is easier to analyze in a non-random setting.

Finally, we notice that in the synthetic configuration model, all correlation measures are close to zero, and the difference between different realizations of the model is remakarbly small (see the values of σ). However, at this point very little can be said about statistical significance of these results because, as we proved above, r shows pathological behaviour on large power law graphs and the setting of directed graphs is very different from the setting of independent observations. This raises important and challenging questions for future research: which magnitude of degree-degree dependencies should be seen as significant and how to construct mathmatically sound statistical tests for establishing such significant dependencies.

		Pearson			Spearman uniform			Spearman average			Kendall		
			Randomized			Randomized			Randomized			Randomized	
Graph	α/β	Data	μ	σ	Data	μ	σ	Data	μ	σ	Data	μ	σ
DE wiki	+/-	-0.0552	-0.0178	0.0001	-0.1434	-0.0059	0.0002	-0.1435	-0.0059	0.0002	-0.0986	-0.0038	0.0008
	-/+	0.0154	-0.0030	0.0002	0.0481	-0.0008	0.0002	0.0484	-0.0008	0.0002	0.0.326	-0.0005	0.0001
	+/+	-0.0323	-0.0091	0.0002	-0.0640	-0.0048	0.0002	-0.0640	-0.0048	0.0002	-0.0446	-0.0006	0.0001
	-/-	-0.0123	-0.0060	0.0001	0.0119	-0.0009	0.0002	0.0120	-0.0009	0.0002	0.0074	-0.0032	0.0001
EN wiki	+/-	-0.0557	-0.0180	0	-0.1999	-0.0064	0.0001	-0.1999	-0.0064	0.0001	-0.1364	-0.0043	0.0001
	-/+	-0.0007	-0.0015	0.0001	0.0239	-0.0011	0.0001	0.0240	-0.0011	0.0001	0.0163	-0.0008	0.0001
	+/+	-0.0713	-0.0125	0.0001	-0.0855	-0.0053	0.0001	-0.0855	-0.0053	0.0001	-0.0581	-0.0035	0.0001
	-/-	-0.0074	-0.0024	0.0001	-0.0664	-0.0013	0.0001	-0.0666	-0.0013	0.0001	-0.0457	-0.0009	0.0001
ES wiki	+/-	-0.1031	-0.0336	0.0002	-0.1429	-0.0186	0.0003	-0.1429	-0.0186	0.0003	-0.0972	-0.0126	0.0002
	-/+	-0.0033	-0.0071	0.0002	-0.0407	-0.0047	0.0003	-0.0417	-00048	0.0003	-0.0294	-0.0034	0.0002
	+/+	-0.0272	-0.0201	0.0002	0.0178	-0.0125	0.0003	0.0178	-0.0125	0.0003	0.0119	-0.0084	0.0002
	-/-	-0.0262	-0.0116	0.0001	-0.1627	-0.0071	0.0003	-0.1669	-0.0072	0.0003	-0.1174	-0.0051	0.0002
FR wiki	+/-	-0.0536	-0.0252	0.0001	-0.1065	-0.0123	0.0002	-0.1065	-0.0123	0.0002	-0.0720	-0.0083	0.0002
	-/+	0.0048	-0.0031	0.0002	0.0119	-0.0016	0.0003	0.0121	-0.0016	0.0003	0.0085	-0.0011	0.0002
	+/+	-0.0512	-0.0173	0.0002	-0.0126	-0.0093	0.0002	-0.0126	-0.0090	0.0015	-0.0087	-0.0063	0.0001
	-/-	-0.0094	-0.0054	0.0001	-0.0262	-0.0021	0.0003	-0.0267	-0.0025	0.0015	-0.0186	-0.0015	0.0002
HU wiki	+/-	-0.1048	-0.0378	0.0003	-0.1280	-0.0220	0.0006	-0.1280	-0.0220	0.0006	-0.0877	-0.0148	0.0004
	-/+	0.0120	-0.0056	0.0005	0.0525	0.0002	0.0005	0.0595	0	0.0006	0.0442	0	0.0004
	+/+	-0.0579	-0.0261	0.0005	-0.0207	-0.0157	0.0005	-0.0207	-0.0157	0.0004	-0.0140	-0.0107	0.0003
	-/-	-0.0279	-0.0084	0.0004	0.0051	0.0004	0.0005	0.0060	0.0002	0.0006	0.0050	-0.0001	0.0005
IT wiki	+/-	-0.0711	-0.0319	0.0001	-0.0964	-0.0158	0.0002	-0.0964	-0.0158	0.0002	-0.0653	-0.0106	0.0002
	-/+	0.0048	-0.0031	0.0002	0.0468	-0.0013	0.0002	0.0469	-0.0013	0.0003	0.0319	-0.0009	0.0002
	+/+	-0.0704	-0.0204	0.0002	-0.0277	-0.0121	0.0002	-0.0277	-0.0122	0.0002	-0.0189	-0.0081	0.0001
	-/-	-0.0115	-0.0050	0.0001	-0.0428	-0.0016	0.0002	-0.0429	-0.0016	0.0002	-0.0296	-0.0011	0.0002
KO wiki	+/-	-0.0805	-0.0562	0.0004	-0.2696	-0.0476	0.0037	-0.2722	-0.0482	0.0038	-0.1985	-0.0328	0.0073
	-/+	0.0157	-0.0009	0.0030	0.1760	0.0019	0.0046	0.2323	0.0034	0.0046	0.1902	0.0031	0.0035
	+/+	-0.1697	-0.0357	0.0035	0.0016	-0.0267	0.0041	0.0191	-0.0272	0.0040	0.0170	0.0298	0.0415
	-/-	-0.0138	-0.0034	0.0015	-0.0493	0.0062	0.0045	-0.0618	0.0083	0.0042	-0.0463	0.0065	0.0032
NL wiki	+/-	-0.0585	-0.0346	0.0001	-0.3017	-0.0211	0.0002	-0.3018	-0.0211	0.0002	-0.2089	-0.0142	0.0002
	-/+	0.0100	-0.0025	0.0003	0.0727	-0.0007	0.0003	0.0730	-0.0007	0.0003	0.0504	-0.0004	0.0003
	+/+	-0.0628	-0.0194	0.0001	0.0016	-0.0104	0.0003	0.0016	-0.0104	0.0003	0.0015	-0.0070	0.0002
	-/-	-0.0233	-0.0091	0.0001	-0.1498	-0.0019	0.0003	-0.1505	-0.0019	0.0003	-0.1048	-0.0013	0.0002
RU wiki	+/-	-0.0911	-0.0225	0.0004	-0.1080	-0.0093	0.0015	-0.1084	-0.0093	0.0015	-0.0755	-0.0064	0.0010
	-/+	0.0398	-0.0006	0.0009	0.1977	0	0.0008	0.2200	0.0001	0.0009	0.1655	0.0001	0.0007
	+/+	0.0082	-0.0038	0.0010	0.2472	0.0002	0.0015	0.2480	0.0001	0.0015	0.1736	0.0001	0.0010
	-/-	-0.0242	-0.0030	0.0007	0.0236	0.0009	0.0011	0.0255	0.0007	0.0015	0.0187	0.0006	0.0007

Table 1: Degree-degree correlations for Wikipedia graphs.

References

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[2] 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.

[3] Paolo Boldi and Sebastiano Vigna. The webgraph framework ii: Codes for the world-wide web. In *Data Compression Conference, 2004. Proceedings. DCC 2004*, page 528. IEEE, 2004.

[4] Markus Brede and Sitabhra Sinha. Assortative mixing by degree makes a network more unstable. *arXiv preprint cond-mat/0507710*, 2005.

[5] Ningyuan Chen and Mariana Olvera-Cravioto. Directed random graphs with given degree distributions. *arXiv preprint arXiv:1207.2475*, 2012.

[6] Daren B.H. Cline. Convolution tails, product tails and domains of attraction. *Probability Theory and Related Fields*, 72(4):529–557, 1986.

[7] Sebastiano de Franciscis, Samuel Johnson, and Joaquín J. Torres. Enhancing neural-network performance via assortativity. *Physical Review E*, 83(3):036114, 2011.

[8] 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.

[9] Adrien Henry, Françoise Monéger, Areejit Samal, and Olivier C. Martin. Network function shapes network structure: the case of the arabidopsis flower organ specification genetic network. *Mol. BioSyst.*, 2013.

[10] Andreas Kaltenbrunner, Gustavo Gonzalez, Ricard Ruiz De Querol, and Yana Volkovich. Comparative analysis of articulated and behavioural social networks in a social news sharing website. *New Review of Hypermedia and Multimedia*, 17(3):243–266, 2011.

[11] Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[12] David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. When the wikipedians talk: Network and tree structure of wikipedia discussion pages. In *ICWSM*, 2011.

[13] Nelly Litvak and Remco van der Hofstad. Degree-degree correlations in random graphs with heavy-tailed degrees. *arXiv preprint arXiv:1202.3071*, 2012. To appear in Internet Mathematics.

[14] Nelly Litvak and Remco van der Hofstad. Uncovering disassortativity in large scale-free networks. *Physical Review E*, 87(2):022801, 2013.

[15] Mark E.J. Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[16] Mark E.J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[17] Mark E.J. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[18] Mahendra Piraveenan, Mikhail Prokopenko, and Albert Zomaya. Assortative mixing in directed biological networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(1):66–78, 2012.

[19] Mahendra Piraveenan, Mikhail Prokopenko, and Albert Y. Zomaya. Assortativeness and information in scale-free networks. *The European Physical Journal B*, 67(3):291–300, 2009.

[20] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

In-Core Computation of Geometric Centralities with HyperBall: A Hundred Billion Nodes and Beyond

Paolo Boldi Sebastiano Vigna
Dipartimento di Informatica, Università degli Studi di Milano, Italy

***Abstract*—Given a social network, which of its nodes are more central? This question was asked many times in sociology, psychology and computer science, and a whole plethora of *centrality measures* (a.k.a. *centrality indices*, or *rankings*) were proposed to account for the importance of the nodes of a network. In this paper, we approach the problem of computing *geometric centralities*, such as *closeness* [1] and *harmonic centrality* [2], on very large graphs; traditionally this task requires an all-pairs shortest-path computation in the exact case, or a number of breadth-first traversals for approximated computations, but these techniques yield very weak statistical guarantees on highly disconnected graphs. We rather assume that the graph is accessed in a *semi-streaming* fashion, that is, that adjacency lists are scanned almost sequentially, and that a very small amount of memory (in the order of a dozen bytes) per node is available in core memory. We leverage the newly discovered algorithms based on HyperLogLog counters [3], making it possible to approximate a number of geometric centralities at a very high speed and with high accuracy. While the application of similar algorithms for the approximation of closeness was attempted in the MapReduce [4] framework [5], our exploitation of HyperLogLog counters reduces exponentially the memory footprint, paving the way for in-core processing of networks with a hundred billion nodes using "just" 2 TiB of RAM. Moreover, the computations we describe are inherently parallelizable, and scale linearly with the number of available cores.**

I. Introduction

In the last years, there has been an ever-increasing research activity in the study of real-world complex networks. These networks, typically generated directly or indirectly by human activity and interaction, appear in a large variety of contexts and often exhibit a surprisingly similar structure.

One of the most important notions that researchers have been trying to capture in such networks is "node centrality": ideally, every node (often representing an individual) has some degree of influence or importance within the social domain under consideration, and one expects such importance to be reflected in the structure of the social network. Centrality in fact has a long history in the context of social sciences: starting from the late 1940s [1] the problem of singling out influential individuals in a social group has been a holy grail that sociologists have been trying to capture for many decades.

The authors have been supported by the EU-FET grant NADINE (GA 288956).

Among the types of centrality that have been considered in the literature (see [6] for a good survey), many have to do with the distance to other nodes. If, for instance, the sum of distances to all other nodes is large, the node is *peripheral*, which is the starting point to define Bavelas's *closeness centrality* as the reciprocal of peripherality (i.e., the reciprocal of the distances to all other nodes).

Interestingly, many of these indices can be recast in terms of suitable calculations using the sizes of the balls of varying radius around a node. In a previous work [3] we presented HyperANF, a tool that can compute the distance distribution of very large graphs. HyperANF has been used, for instance, to show that Facebook has just four "degrees of separation" [7]. The goal of this paper is to extends the HyperANF approach to compute a number of centrality indices based on distances.

Beside large-scale experiment using the full ClueWeb09 graph (almost five billion nodes), we provide an empirical evaluation of the accuracy of our method through a comparison with the exact centrality values on a snapshot of Wikipedia (on larger graphs the exact computation would be infeasible). We also provide comparisons with a MapReduce-based [4] approach [5], showing that a careful combination of HyperLogLog counters, compression and succinct data structure can provide a speedup of two orders of magnitude, and in fact, comparing costs, more scalability. We also show how to extend our techniques to a class of weighted graphs with a tiny loss in space.

The Java software implementing the algorithms described in this paper is distributed as free software within the WebGraph framework.[1] Moreover, all dataset we use are publicly available.

Using our Java tool we are able, for the first time, to approximate distance-based centrality indices on graphs with billions of nodes using a standard workstation.

II. Notation

In this paper, we use the following notation: $G = (V, E)$ is a directed graph with $n = |V|$ nodes and $m = |E|$ arcs; we write $x \to y$ as a shortcut for $(x, y) \in E$. The length of the shortest path from x to y is denoted by $d(x, y)$ and called the *distance* between x and y; we let $d(x, y) = \infty$ if there is no

[1] `http://webgraph.di.unimi.it/`

directed path from x to y. The nodes *reachable* from x are the nodes y such that $d(x, y) < \infty$. The nodes *coreachable* from x are the nodes y such that $d(y, x) < \infty$. We let G^T be the *transpose* of G (i.e., the graph obtained by reverting all arc directions in G). The ball of radius r around x is

$$\mathscr{B}_G(x, r) = \{\, y \mid d(x, y) \leq t \,\}.$$

III. Geometric centralities

We call *geometric* those centrality measures[2] whose basic assumption is that importance depends on some function of the distances. These are actually some of the oldest measures defined in the literature.

A. Closeness centrality

Closeness was introduced by Bavelas in the late forties [8]; the closeness of x is defined by

$$\frac{1}{\sum_y d(y, x)}. \tag{1}$$

The intuition behind closeness is that nodes with a large sum of distances are *peripheral*. By reciprocating the sum, nodes with a smaller denominator obtain a larger centrality. We remark that for the above definition to make sense, the graph needs be strongly connected. Lacking that condition, some of the denominators will be ∞, resulting in a rank of zero for all nodes which cannot coreach the whole graph.

In fact, it was not probably in Bavelas's intentions to apply the measure to non-connected graphs, but nonetheless the measure is sometimes "patched" by simply not including pairs with infinite distance, that is,

$$\frac{1}{\sum_{d(y,x)<\infty} d(y, x)};$$

for the sake of completeness, one further assumes that nodes with an empty coreachable set have centrality 0 by definition. These apparently innocuous adjustments, however, introduce a strong bias toward nodes with a small coreachable set.

B. Lin's centrality

Nan Lin [9] tried to patch the definition of closeness for graphs with infinite distances by weighting closeness using the square of the number of coreachable nodes; his definition for the centrality of a node x with a nonempty coreachable set is

$$\frac{\left|\{y \mid d(y, x) < \infty\}\right|^2}{\sum_{d(y,x)<\infty} d(y, x)}.$$

Nodes with an empty coreachable set have centrality 1 by definition.

The rationale behind this definitions is the following: first, we consider closeness not the inverse of a sum of distances, but rather the inverse of the *average* distance, which entails a first multiplication by the number of coreachable nodes. This change normalizes closeness across the graph. Now, however, we want nodes with a larger coreachable set to be more important, given that the average distance is the same, so we multiply again by the number of coreachable nodes.

Lin's index was somewhat surprisingly ignored in the following literature. Nonetheless, it seems to provide a reasonable solution for the problems caused by the definition of closeness.

C. Harmonic centrality

As we noticed, the main problem with closeness lies in the presence of pairs of unreachable nodes. In [2], we have proposed to replace the reciprocal of the sum of distances in the definition of closeness with the sum of reciprocals of distances. Conceptually, this corresponds to replacing the reciprocal of a denormalized average of distances with the the reciprocal of a denormalized *harmonic* mean of distances, analogously to what Marchiori and Latora proposed to do with the notion of average distance [10]. The harmonic mean has the useful property of handling ∞ cleanly (assuming, of course, that $\infty^{-1} = 0$).

We thus obtain the *harmonic centrality* of x:

$$\sum_{y \neq x} \frac{1}{d(y, x)} = \sum_{d(y,x)<\infty, y \neq x} \frac{1}{d(y, x)}. \tag{2}$$

The difference with (1) might seem minor, but actually it is a radical change. Harmonic centrality is strongly correlated to closeness centrality in simple networks, but naturally also accounts for nodes y that cannot reach x. Thus, it can be fruitfully applied to graphs that are not strongly connected.

IV. HyperBall

In this section, we present *HyperBall*, a general framework for computations that depend on the number of nodes at distance at most t or exactly t from a node. HyperBall uses the same dynamic programming scheme of algorithms that approximate neighborhood functions, such as ANF [11] or HyperANF [3], but instead of aggregating at each step the information about all nodes into a single output value (the neighbourhood function at t) HyperBall makes it possible to perform a different set of operations (for example, depending on the centrality to be computed). We have tried to make the treatment self-contained, albeit a few details will be only sketched here, when they can be deduced from the description of HyperANF [3].

[2] Most centrality measures proposed in the literature were actually described only for undirected, connected graphs. Since the study of web graphs and online social networks has posed the problem of extending centrality concepts to networks that are directed, and possibly not strongly connected, in the rest of this paper we consider measures depending on the *incoming* arcs of a node, so distances will be taken from all nodes to a fixed node. If necessary, these measures can be called "negative", as opposed to the "positive" versions obtained by taking the transpose of the graph.

A. HyperLogLog counters

HyperLogLog counters, as described in [12] (which is based on [13]), are used to count approximately the number of distinct elements in a stream. For the purposes of the present paper, we need to recall briefly their behaviour. Essentially, these probabilistic counters are a sort of *approximate set representation* to which, however, we are only allowed to pose questions about the (approximate) size of the set.

Let $\mathscr{D}$ be a fixed domain and $h : \mathscr{D} \to 2^\infty$ be a fixed hash function mapping each element of $\mathscr{D}$ into an infinite binary sequence. For a given $x \in 2^\infty$, let $h_t(x)$ denote the sequence made by the leftmost t bits of $h(x)$, and $h^t(x)$ be the sequence of remaining bits of x; h_t is identified with its corresponding integer value in the range $\{0, 1, \ldots, 2^t - 1\}$. Moreover, given a binary sequence w, we let $\rho^+(w)$ be the number of leading zeroes in w plus one (e.g., $\rho^+(00101) = 3$). Unless otherwise specified, all logarithms are in base 2.

Algorithm 1 The Hyperloglog counter as described in [12]: it allows one to count (approximately) the number of distinct elements in a stream. α_p is a constant whose value depends on p and is provided in [12]. Some technical details have been simplified.

```
0   h : D → 2^∞, a hash function from the domain of items
1   M[−] the counter, an array of p = 2^b registers
2      (indexed from 0) and set to −∞
3
4   function add(M: counter, x: item)
5   begin
6      i ← h_b(x);
7      M[i] ← max{M[i], ρ+(h^b(x))}
8   end; // function add
9
10  function size(M: counter)
11  begin
12     Z ← (Σ_{j=0}^{p−1} 2^{−M[j]})^{−1};
13     return E = α_p p^2 Z
14  end; // function size
15
16  foreach item x seen in the stream begin
17     add(M,x)
18  end;
19  print size(M)
```

The value E printed by Algorithm 1 is [12][Theorem 1] an asymptotically almost[3] unbiased estimator for the number n of distinct elements in the stream; for $n \to \infty$, the *relative standard deviation* (that is, the ratio between the standard deviation of E and n) is at most $\beta_p/\sqrt{p} \leq 1.06/\sqrt{p}$, where β_p is a suitable constant. Moreover, even if the size of the registers (and of the hash function) used by the algorithm is unbounded, one can limit it to $\log\log(n/p) + \omega(n)$ bits obtaining almost certainly the same output ($\omega(n)$ is a function going to infinity arbitrarily slowly); overall, the algorithm requires $(1 + o(1)) \cdot p \log\log(n/p)$ bits of space (this is the reason why these counters are called HyperLogLog). Here and in the rest of the paper we tacitly assume that $p \geq 16$ and that registers are made of $\lceil \log\log n \rceil$ bits.

[3] For the purposes of this paper, in the following we will consider in practice the estimator as it if was unbiased, as suggested in [12].

B. Estimating balls

The basic idea used by algorithms such as ANF [11] and HyperANF [3] is that that $\mathscr{B}_G(x, r)$, the ball of radius r around node x, satisfies

$$\mathscr{B}_G(x, 0) = \{x\}$$
$$\mathscr{B}_G(x, r+1) = \bigcup_{x \to y} \mathscr{B}_G(y, r) \cup \{x\}.$$

We can thus compute $\mathscr{B}_G(x, r)$ iteratively using sequential scans of the graph (i.e., scans in which we go in turn through the successor list of each node). One obvious drawback of this solution is that during the scan we need to access randomly the sets $\mathscr{B}_G(x, r-1)$ (the sets $\mathscr{B}_G(x, r)$ can be just saved on disk on an *update file* and reloaded later). For this to be possible, we need to store the (approximated) balls in a data structure that can be fit in the core memory: here is where probabilistic counters come into play; to be able to use them, though, we need to endow counters with a primitive for the union. Union can be implemented provided that the counter associated with the stream of data AB can be computed from the counters associated with A and B; in the case of HyperLogLog counters, this is easily seen to correspond to maximising the two counters, register by register.

Algorithm 2, named *HyperBall*, describes our strategy to compute centralities. We keep track of one HyperLogLog counter for each node; at the t-th iteration of the main loop, the counter $c[v]$ is in the same state as if it would have been fed with $\mathscr{B}_G(v, t)$, and so its expected value is $|\mathscr{B}_G(v, t)|$. During the execution of the loop, when we have finished examining node v the counter a is in the same state as if it would have been fed with $\mathscr{B}_G(v, t+1)$, and so its value will be $|\mathscr{B}_G(v, t+1)|$ in expectation.

This means, in particular, that it is possible to compute an approximation of

$$|\{y \mid d(x, y) = t\}|$$

(the number of nodes at distance t from x) by evaluating

$$|\mathscr{B}_G(v, t+1)| - |\mathscr{B}_G(v, t)|.$$

The computation would be exact if the algorithm had actually kept track of the set $\mathscr{B}_G(x, t)$ for each node, something that is obviously not possible; using probabilistic counters makes this feasible, at the cost of tolerating some approximation in the computation of cardinalities.

The idea of using differences between ball sizes to estimate the number of nodes at distance t appeared also in [14],

where it was used with a different kind of counter (Martin–Flajolet) to estimate the 90% percentile of the distribution of distances from each node. An analogous technique, always exploiting Martin–Flajolet counters, was adopted in [5] to approximate closeness. In both cases the implementations were geared towards MapReduce [4]. A more sophisticated approach, which can be implemented using breadth-first visits or dynamic programming, uses *all-distances sketches* [15]: it provides better error bounds, but it requires also significantly more memory.

Algorithm 2 HyperBall in pseudocode. The algorithm uses, for each node $v \in n$, an initially empty HyperLogLog counter $c[v]$. The function union$(-,-)$ maximises two counters register by register. At line 19, one has the estimate of $|\mathscr{B}_G(v,t)|$ from $c[v]$ and the estimate of $|\mathscr{B}_G(v,t+1)|$ from a.

```
0   c[−], an array of n HyperLogLog counters
1
2   function union(M: counter, N: counter)
3     foreach i < p begin
4       M[i] ← max(M[i], N[i])
5     end
6   end; // function union
7
8   foreach v ∈ n begin
9     add(c[v], v)
10  end;
11  t ← 0;
12  do begin
13    foreach v ∈ n begin
14      a ← c[v];
15      foreach v → w begin
16        a ← union(c[w], a)
17      end;
18      write ⟨v, a⟩ to disk
19      do something with a and c[v]
20    end;
21    Read the pairs ⟨v, a⟩ and update the array c[−]
22    t ← t + 1
23  until no counter changes its value.
```

HyperBall is run until all counters stabilise (e.g., the last iteration must leave all counters unchanged). As shown in [3], any alternative termination condition may lead to arbitrarily large mistakes on pathological graphs.

V. Estimating centralities

It should be clear that exactly three ingredients for each node x are necessary to compute closeness, harmonic, and Lin's centrality:

- the sum of the distances to x;
- the sum of the reciprocals of the distances to x;
- the size of the coreachable set of x.

The last quantity is simply the value of each counter $c[v]$ in HyperBall at the end of the computation on G^T. The other quantities can be easily computed in a cumulative fashion nothing that

$$\sum_y d(y,x) = \sum_{t>0} t|\{\, y \mid d(y,x) = t \,\}| = \sum_{t>0} t\big(|\mathscr{B}_{G^T}(x,t)| - |\mathscr{B}_{G^T}(x,t-1)|\big),$$

and

$$\sum_{y\neq x} \frac{1}{d(y,x)} = \sum_{t>0} \frac{1}{t}|\{\, y \mid d(y,x) = t \,\}| = \sum_{t>0} \frac{1}{t}\big(|\mathscr{B}_{G^T}(x,t)| - |\mathscr{B}_{G^T}(x,t-1)|\big).$$

We can thus obtain estimators for the first two ingredients by storing a single floating point value per node, and cumulating the values for each node during the execution of HyperBall. Note that we have to run the algorithm on the *transpose* of G, since we need to estimate the distances *to* x, rather than *from* x.

If we accept the minimum possible precision (16 registers per HyperLogLog counter), the core memory necessary for running HyperBall is just 16 bytes per node (assuming $n \leq 2^{64}$), plus four booleans per node to keep track of modifications, and ancillary data structures that are orders of magnitude smaller. A machine with 2 TiB of core memory could thus compute centralities on networks with more than a hundred billion nodes, prompting the title of this paper.

Note that even if we use a small number of registers per HyperLogLog counter, by executing HyperBall multiple times we can increase the confidence in the computed value for each estimator, leading to increasingly better approximations.

As in the case of the average distance [3], the theoretical bounds are quite ugly, but actually the derived values we compute are very precise, as shown by the concentration of the values associated several runs. Multiple runs in this case are very useful, as they make it possible to compute the empirical standard deviation.

A. Representing and scanning the graph

In the previous section we have estimated the core memory usage of HyperBall without taking the graph size into account. However, representing and accessing the graph is a nontrivial problem, in particular during the last phases of the computation, where we can keep track of the few nodes that are modifying their counter, and propagate new values only when necessary.

Here we exploit two techniques: *compression*, to represent the graph as a bit stream in a small amount of disk space, so that we are able to access it from disk efficiently using memory mapping; and *succint data structures*, to access quickly the bitstream in a random fashion.

In particular, for compression we use the WebGraph framework [16], which is a set of state-of-the-art algorithms and codes to compress web and social graphs. WebGraph represents a graph as a bitstream, with a further 64-bit pointer for each node if random access is necessary. To store the pointers in memory, we use a succinct encoding based on a broadword implementation [17] of the Elias-Fano representation of monotone sequences [18]. This way, the cost of a pointer is logarithmic in the average length per node of the bitstream, and in real-world graphs this means about one byte of core memory per node, which is an order of magnitude less than the memory used by HyperBall.

B. Error bounds

The estimate $\hat{\mathscr{B}}_G(x,t)$ for $|\mathscr{B}_G(x,t)|$ obtained by HyperBall follow the bounds given in Section IV-A. Nonetheless, as soon as we consider the differences $\hat{\mathscr{B}}_G(x,t+1)-\hat{\mathscr{B}}_G(x,t)$, the bounds on the error become quite ugly. A similar problem occurs when estimating the distance distribution and its statistics: by taking the difference between points of the cumulative distribution, the bound on the relative standard deviation is lost [3].

Note that in part this is an intrinsic problem: HyperBall essentially runs in quasi-linear expected time $O(pm \log n)$ [15], and due to known bounds on the approximation the diameter [19] it is unlikely that it can provide in all cases a good approximation of the differences (which would imply a good approximation of the eccentricity of each node, and in the end a good approximation of the diameter).

Nonetheless, for a number of reasons the estimates of the differences on real-world graphs turn out to be very good. First of all, for very small numbers the HyperLogLog counters compute a different estimator (not shown in Algorithm 1) that is much more accurate. Second, on social and web graphs (and in general, for small-world graphs) the function $|\mathscr{B}_G(x,t)|$ grows very quickly for small values of t, so the magnitude of the difference is not far from the magnitude of the ball size, which makes the relative error on the ball size small with respect to the difference. Third, once most of the nodes in the reachable set are contained in $\mathscr{B}_G(x,t)$, the error of the HyperLogLog counter tends to stabilise, so the bound on the relative standard deviation "transfers" to the differences.

We thus expect (and observe) that the estimation of the size of the nodes at distance t to be quite accurate, in spite of the difficulty of proving a theoretical error bound.

From a practical viewpoint, the simplest way of controlling the error is generating multiple samples, and computing the empirical standard deviation. This is, for example, the way in which the results for the "degrees of separation" in [7] were reported. By generating several samples, we can restrict the confidence interval for the computed values.

In Section VIII we report experiments on a relatively small graph on which centralities could be computed exactly to show that the precision obtained on the final values is very close to the theoretical prediction for a single counter.

VI. Computing with weights on the nodes

It is very natural, in a number of contexts, to have *weights* on the nodes that represent their importance. Centrality measures should then be redefined taking into account weights in the obvious way: the sum of distances should become

$$\sum_y w(y)d(y,x),$$

the sum of inverse distances should become

$$\sum_y \frac{w(y)}{d(y,x)},$$

and the size of the coreachable set should become

$$\sum_{d(y,x)<\infty} w(y).$$

There is no direct way to incorporate weights in the dynamic programming algorithm, but weights can be easily simulated if they are integers. Suppose that the weighting function is $w : V \to \{1,\dots,W\}$, and assume that each node $x \in V$ is associated with a set $\mathscr{R}(x) = \{x_1,\dots,x_{w(x)}\}$ of replicas of the node (with the proviso that distinct nodes have disjoint replicas).

Then the *weighted ball of radius r around x* can be defined recursively as:

$$\begin{aligned}\mathscr{W}_G(x,0) &= \mathscr{R}(x)\\ \mathscr{W}_G(x,r+1) &= \mathscr{R}(x) \cup \bigcup_{x\to y} \mathscr{W}_G(y,r).\end{aligned}$$

It is easy to see that

$$|\mathscr{W}_G(x,r+1)| - |\mathscr{W}_G(x,r)| = \sum_{y:d(x,y)=r} w(y).$$

Attention must be paid, of course, to the sizing of the counters in this case. Instead of $\log\log n$ bits, counters with

$$\log\log \sum_x w(x) \le \log\log(Wn) = \log(\log n + \log W)$$

bits will have to be used. We note, however, that since the increase factor $\sum_x w(x)/n$ passes through two logarithms, it is unlikely that more than 6 or at most 7 bits will be ever necessary.

VII. Computing with discount functions

If we look at harmonic centrality from a more elementary perspective, we can see that when measuring the centrality of a node we start by considering its (in)degree, that is, how many neighbours it has at distance one. Unsatisfied by this raw measure, we continue and take into consideration nodes at distance two. However, their number is not as important as the degree, so before adding it to the degree we *discount* its importance it by $1/2$. The process continues with nodes at distance three, discounted by $1/3$ until all coreachable nodes have been considered.

The essence of this process is that we are counting nodes at larger and larger distances from the target, discounting their number based on their distance. One can generalize this idea to a *family* of centrality measures. The idea, similar to the definition of *discounted cumulative gain* in information retrieval [20], is that with each coreachable node we gain some importance. However, the importance given by the node is *discounted* by a quantity depending on the distance that, in the case of harmonic centrality, is the reciprocal $1/d$. Another reasonable choice is a *logarithmic* discount $1/\log(d+1)$, which attenuates even more slowly the importance of far nodes, or a *quadratic* discount $1/d^2$. More generally, the centrality of x based on a non-increasing discount function $f : \mathbf{N} \to \mathbf{R}$ is

$$\sum_{d(y,x)<\infty, y\neq x} f(d(y,x)).$$

It can be approximated by HyperBall nothing that

$$\begin{aligned}\sum_{d(y,x)<\infty, y\neq x} f(d(y,x)) &= \sum_{t>0} f(t)|\{\, y \mid d(y,x) = t \,\}| \\ &= \sum_{t>0} f(t)\big(|\mathscr{B}_{G^T}(x,t)| - |\mathscr{B}_{G^T}(x,t-1)|\big).\end{aligned}$$

We are proposing relatively mild discount functions, in contract with the *exponential* decay used, for example, in Katz's index [21]. This is perfectly reasonable, since Katz's index is based on *paths*, which are usually infinite. Discount-based centralities are necessarily given by finite summations, so there is no need for a rapid decay. Actually, by choosing a constant discount function we would estimate the importance of each node just by the number of nodes it can coreach (i.e., in the undirected case, by the size of its connected component).

Combining this observation and that of Section VI, we conclude that HyperBall can compute a class of centralities that could be called *discounted-gain centralities*:[4]

$$\sum_{d(y,x)<\infty, y\neq x} w(y) f(d(y,x)).$$

VIII. Experiments

We decided to perform three kinds of experiments:

- A small-scale experiment on the same graphs for which explicit timings are reported in [5], to compare the absolute speed of a MapReduced-based approach using the Hadoop open-source implementation and of an in-core approach. Note that the graphs involved are extremely unrealistic (e.g., they have all diameter 2 and are orders of magnitude denser than typical web or social graphs). This experiment was run using $p = 64$ registers per HyperLogLog counter, corresponding to a relative standard deviation of 13.18%, which is slightly better than the one used in [5] (13.78%, as communicated by the authors), to make a comparison of the execution times possible.

[4]These are called *spatially decaying* in [22].

TABLE I. Comparative timings per iteration between the Hadoop implementation described in [5] running on 50 machines and HyperBall on a MacBook Pro laptop (2.6 GHz Intel i7, 8 GiB RAM, 8 cores) and on a 32-core, 64 GiB RAM workstation using 2.3 GHz AMD Opteron 6276 processors. Timings for the Hadoop implementation were deducted from Figure 4(b) of [5]. Note that the better processor and the SSD disk of the MacBook Pro make it almost twice faster (per core) than the workstation.

Size (nodes/arcs)	Hadoop [5]	MacBook	32 cores
20 K / 40 M	250 s	2 s	1 s
59 K / 282 M	1750 s	10 s	4 s
177 K / 1977 M	2875 s	70 s	23 s

- A medium-size experiment to verify the convergence properties of our computations. For this purpose, we had to restrict ourselves to a graph for which exact values could be computed using n breadth-first visits. We focused on a public snapshot of Wikipedia[5]. This graph consists of 4 206 785 nodes and 101 355 853 arcs (with average degree 24 and the largest strongly connected component spanning about 89% of the nodes). We performed 100 computations using $p = 4096$ registers per counters, corresponding to a theoretical relative standard deviation of 1.62% for each computation. The exact computation of the centralities required a few days using 40 cores.
- A large-scale experiment using the largest ClueWeb09[6] graph; ClueWeb09 is, at the time of this writing, the largest web graph publicly available, one order of magnitude larger that previous efforts in terms of nodes. It contains 4 780 950 903 nodes and 7 939 647 896 arcs. The purpose of this experiment was to show our methods in action on a very large dataset.[7]

In Table I we report the timings for an iteration on the same set of Kronecker graphs used in [5]. A standard workstation with 32 cores using HyperBall is at least 150 times faster than a Hadoop-based implementation using using 50 machines; even a MacBook Pro with 8 cores is at least 50 times faster.

In Figure 1 we report the results of the second set of experiments, which fully confirm our empirical observations on the behaviour of the difference estimator: on average, the relative error on the computed centrality indices is very close to the theoretical prediction for each single HyperLogLog counter, and, in fact, almost always significantly smaller.

It is interesting to observe that the estimation on the number of coreachable nodes (depending on the value of a single

[5]Available at `http://law.di.unimi.it/`

[6]A dataset gathered in 2009 within the U.S. National Science Foundation's Cluster Exploratory (CluE) program. The ClueWeb12 graph will be even larger, but it is presently still under construction. See http://lemurproject.org/clueweb09/

[7]We remark that due to the way in which the graph has been collected (e.g., probably starting from a large seed) the graph is actually significantly less dense than a web graph obtained by breadth-first sampling or similar techniques. Moreover, the graph contains the whole set of discovered nodes, even if only about 1.2 billion pages were actually crawled. As a result, many statistics are off scale: the harmonic diameter [10], [23] is ≈ 15131 (typical values for breadth-first web snapshots are ≈ 20) and the giant component is just 0.6% of the whole graph.

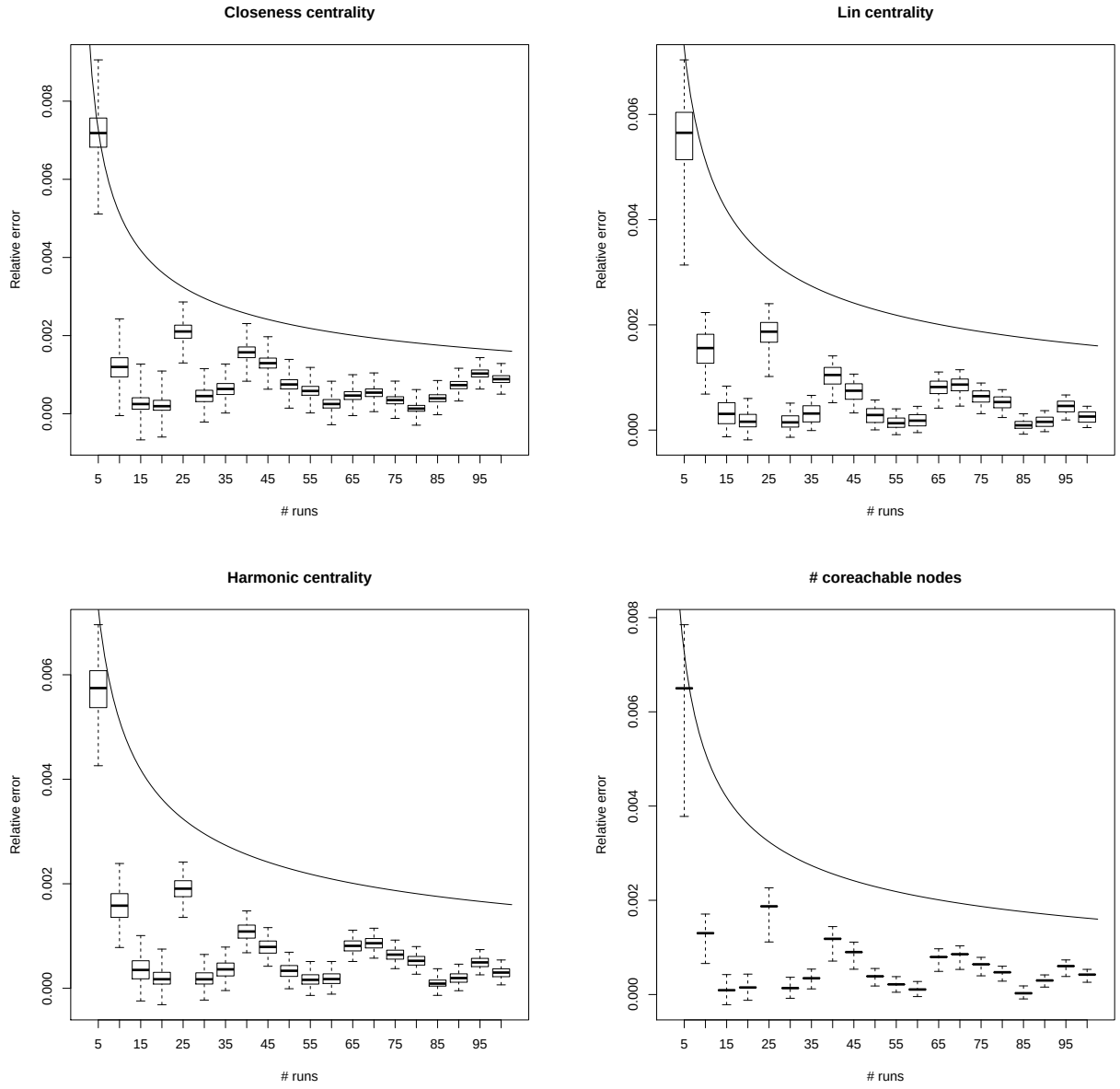


Fig. 1. Relative errors in the computation of centrality measures on Wikipedia: we averaged the values computed in 5, 10, 15, ..., 100 runs and computed the relative error with respect to the real value (the latter were obtained by running an exact implementation). The boxes represent the 1st (lower edge), 2nd (i.e., the median; midline) and 3rd (upper edge) quartile; the whiskers correspond to an interval of length 2σ around the mean. For comparison, each plot contains the curve of the theoretical relative standard deviation for each single HyperLogLog counter over the given number of samples.

counter at the end of the computation) is extremely more concentrated. This is due both to the lack of differences, which reduces the error, and to the fact that most nodes (89%) lie in the giant strongly connected component, so their coreachable set is identical, and this induces a collapse of the quartiles of the error on the median value.

On the same dataset, Table II reports figures showing that increasing the number of cores leaves essentially unmodified the time per arc per core (i.e., linear scalability). The only significant (30%) increase happen at 32 cores, and it is likely to be caused by the nonlinear cost of caching.

Finally, we ran HyperBall on ClueWeb09 using a workstation with 40 Intel Xeon E7-4870 at 2.40 GHz and 1 TiB of RAM (with the same hardware, we could have analysed a graph with 50 billion nodes using $p = 16$). We report the results in Table III. We performed three experiments with different levels of precision, and in the one with the highest precision we fully utilized the in-core memory: the timings show that increasing the precision scales even better than linearly, which is to be expected, because the cost of scanning

TABLE II. TIME PER ARC PER CORE OF A HYPERBALL ITERATION, TESTED ON THE WIKIPEDIA GRAPH WITH $p = 4096$.

cores	Time per arc per core
1	906 ns
2	933 ns
4	967 ns
8	1018 ns
16	1093 ns
32	1389 ns

TABLE III. TIMINGS FOR A FULL 40-CORE COMPUTATION (≈ 200 ITERATIONS) ON CLUEWEB09 USING A DIFFERENT NUMBER p OF REGISTERS PER HYPERLOGLOG COUNTER. THE AMOUNT OF MEMORY DOES NOT INCLUDE 7.2 GIB OF SUCCINCT DATA STRUCTURES THAT STORE POINTERS TO THE MEMORY-MAPPED ON-DISK BITSTREAMS REPRESENTING THE GRAPH AND ITS TRANSPOSE.

p	Memory	Overall time	Per iteration (avg.)
16	73 GiB	96 m	27 s
64	234 GiB	141 m	40 s
256	875 GiB	422 m	120 s

the graph is constant whereas the cost of computing with greater precision grows linearly with the number of registers per HyperLogLog counter. Thus, for a fixed desired precision a greater amount of in-core memory translates into higher speed.

IX. CONCLUSIONS AND FUTURE WORK

We have described HyperBall, a framework for in-core approximate computation of centralities based on the number of (possibly weighted) nodes at distance exactly t or at most t from each node x of a graph. With 2 TiB of memory, HyperBall makes it possible to compute accurately and quickly harmonic centrality for graphs up to a hundred billion nodes. We obtain our results with a mix of approximate set representations (by HyperLogLog counters), efficient compressed graph handling, and succinct data structures to represent pointers (that make it possible to access quickly the memory-mapped graph representation).

We provide experiments on a 4.8 billion node dataset, which should be contrasted with previous literature: the largest dataset in [5] contains 25 million nodes, and the dataset of [14] contains 1.4 billion nodes. Moreover, both papers provide timings only for a small, $\approx 177\,000$-nodes graph, whereas we report timings for all our datasets.

REFERENCES

[1] A. Bavelas, "A mathematical model for group structures," *Human Organization*, vol. 7, pp. 16–30, 1948.

[2] P. Boldi and S. Vigna, "Axioms for centrality," *CoRR*, vol. abs/1308.2140, 2013.

[3] P. Boldi, M. Rosa, and S. Vigna, "HyperANF: Approximating the neighbourhood function of very large graphs on a budget," in *Proceedings of the 20th international conference on World Wide Web*, S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, and R. Kumar, Eds. ACM, 2011, pp. 625–634.

[4] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in *OSDI '04: Sixth Symposium on Operating System Design and Implementation*, 2004, pp. 137–150.

[5] U. Kang, S. Papadimitriou, J. Sun, and H. Tong, "Centralities in large networks: Algorithms and observations," in *Proceedings of the Eleventh SIAM International Conference on Data Mining*. SIAM / Omnipress, 2011, pp. 119–130.

[6] S. P. Borgatti, "Centrality and network flow," *Social Networks*, vol. 27, no. 1, pp. 55–71, 2005.

[7] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, "Four degrees of separation," in *ACM Web Science 2012: Conference Proceedings*. ACM Press, 2012, pp. 45–54, best paper award.

[8] A. Bavelas, "Communication patterns in task-oriented groups." *Journal of the Acoustical Society of America*, 1950.

[9] N. Lin, *Foundations of Social Research*. New York: McGraw-Hill, 1976.

[10] M. Marchiori and V. Latora, "Harmony in the small-world," *Physica A: Statistical Mechanics and its Applications*, vol. 285, no. 3-4, pp. 539 – 546, 2000.

[11] C. R. Palmer, P. B. Gibbons, and C. Faloutsos, "Anf: a fast and scalable tool for data mining in massive graphs," in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2002, pp. 81–90.

[12] P. Flajolet, É. Fusy, O. Gandouet, and F. Meunier, "HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm," in *Proceedings of the 13th conference on analysis of algorithm (AofA 07)*, 2007, pp. 127–146.

[13] M. Durand and P. Flajolet, "Loglog counting of large cardinalities (extended abstract)," in *Algorithms - ESA 2003, 11th Annual European Symposium, Budapest, Hungary, September 16-19, 2003, Proceedings*, ser. Lecture Notes in Computer Science, G. D. Battista and U. Zwick, Eds., vol. 2832. Springer, 2003, pp. 605–617.

[14] U. Kang, C. E. Tsourakakis, A. P. Appel, C. Faloutsos, and J. Leskovec, "HADI: Mining radii of large graphs," *ACM Transactions on Knowledge Discovery from Data*, vol. 5, no. 2, pp. 8:1–8:24, 2011.

[15] E. Cohen, "All-distances sketches, revisited: Scalable estimation of the distance distribution and centralities in massive graphs," *CoRR*, vol. abs/1306.3284, 2013.

[16] P. Boldi and S. Vigna, "The WebGraph framework I: Compression techniques," in *Proc. of the Thirteenth International World Wide Web Conference (WWW 2004)*. Manhattan, USA: ACM Press, 2004, pp. 595–601.

[17] S. Vigna, "Broadword implementation of rank/select queries," in *Experimental Algorithms. 7th International Workshop, WEA 2008*, ser. Lecture Notes in Computer Science, C. C. McGeoch, Ed., no. 5038. Springer–Verlag, 2008, pp. 154–168.

[18] P. Elias, "Efficient storage and retrieval by content and address of static files," *J. Assoc. Comput. Mach.*, vol. 21, no. 2, pp. 246–260, 1974.

[19] L. Roditty and V. V. Williams, "Fast approximation algorithms for the diameter and radius of sparse graphs," in *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, D. Boneh, T. Roughgarden, and J. Feigenbaum, Eds. ACM, 2013, pp. 515–524.

[20] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.

[21] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[22] E. Cohen and H. Kaplan, "Spatially-decaying aggregation over a network," *Journal of Computer and System Sciences*, vol. 73, no. 3, pp. 265–288, 2007.

[23] P. Boldi and S. Vigna, "Four degrees of separation, really," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 2012, pp. 1222–1227.