PROJECT PERIODIC REPORT

Grant Agreement number: 288956

Project acronym: NADINE

Project title: New tools and Algorithms for Directed Network analysis

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

Periodic report: 1st 2nd X

Period covered: from 1.11.2013 to 30.04.2015

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

DECLARATION BY THE SCIENTIFIC REPRESENTATIVE OF THE PROJECT COORDINATOR:

I, as scientific representative of the coordinator of this project and in line with the obligations as stated in Article II.2.3 of the Grant Agreement declare that:

- The attached periodic report represents an accurate description of the work carried out in this project for this reporting period;
- The project (tick as appropriate):

  X has fully achieved its objectives and technical goals for the period;

  □ has achieved most of its objectives and technical goals for the period with relatively minor deviations[2];

  □ has failed to achieve critical objectives and/or is not at all on schedule[3].

- The public website is up to date, if applicable.
- To my best knowledge, the financial statements which are being submitted as part of this report are in line with the actual work carried out and are consistent with the report on the resources used for the project and if applicable with the certificate on financial statement.
- All beneficiaries, in particular non-profit public bodies, secondary and higher education establishments, research organisations and SMEs, have declared to have verified their legal status. Any changes have been reported under section 5 (Project Management) in accordance with Article II.3.f of the Grant Agreement.

Name of scientific representative of the Coordinator: Dima Shepelyansky

Date:31/10/2013 : electron.. sign.

Signature of scientific representative of the Coordinator: Dima Shepelyansky

[2] If either of these boxes is ticked, the report should reflect these and any remedial actions taken.

[3] If either of these boxes is ticked, the report should reflect these and any remedial actions taken.

Table of Contents

Publishable Summary

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)

Project coordinator: Dima Shepelyansky, Lab de Phys. Theorique, CNRS Toulouse, France

Website: www.quantware.ups-tlse.fr/FETNADINE

NADINE – Summary

The central aims of this project are to develop new algorithms to facilitate classification and information retrieval from large directed networks, including PageRank and CheiRank with two-dimensional ranking proposed by partners, using newly developed Monte Carlo methods. The Google matrix formed by the links of the network is analyzed by analytical tools of Stochastic Processes, Random Matrix Theory and quantum chaos and by efficient numerical methods for large matrix diagonalization including the Arnoldi method. The investigations of real directed networks performed by the project highlight their new characteristics allowing to understand in a deeper way the hidden features of these networks. New tools and algorithms produced by the project create fundamental basis for developers of new types of search and social media services.

The consortium has interdisciplinary skills since it unites partners from different sciences including physics, mathematics and computer science.

The project has fulfilled all deliverables and milestones for the reporting period and has resulted in collaborations between all partners. In total, since the beginning of the project, 41 papers and preprints have appeared during the 2nd period within the framework of NADINE (42 in 1st period and 73 during the whole period), These works include 2 papers in PLoS ONE ([69] P4.16) and Sci. Reports ([36] P1.16). The PLOS paper of P1+P4 has been highlighted by Guardian, Independent, Le Figaro, EC CORIDS and press of about 20 countries. The Sci. Reports paper has been highlighted by Independent and MIT Thech. Rev. During the 2nd period the NADINE results have been reported on 28 international conferences (37 conferences during the 2st period, 65 in total).

Highlights in the second reporting period include work on interactions of cultures and top 100 people of Wikipedia from ranking of 24 language editions [69], Google matrix analysis of the multiproduct world trade network [39], development of Monte Carlo algorithm for quick detection of high-degree entities in large directed networks [50], results for RecSys Challenge 2014: an ensemble of binary classifiers and matrix factorization [53], construction of a weighted correlation index for rankings with ties [70] (Refs numbers are given for the whole grant period as on http://www.quantware.ups-tlse.fr/FETNADINE/pub.html)

1.1H. [69] P4.16 Young Ho Eom, Pablo Aragon, David Laniado, Andreas Kaltenbrunner, Sebastiano Vigna, and Dima L. Shepelyansky, **"Interactions of cultures and top people of Wikipedia from ranking of 24 language editions"**, PLoS ONE v.10(3), p.e0114825 (2015) (arXiv:1405.7183[cs.SI], 2014)

2.1H. [39] P1.19 L.Ermann and D.L.Shepelyansky, **"Google matrix analysis of the multiproduct world trade network"**, Eur.Phys. J. B v.88, p.84 (2015) (arXiv:1502.00584[cond-mat.dis-nn])

3.2H. [50] P2.13 K.Avrachenkov, N.Litvak, L.Ostroumova-Prokhorenkova and E.Suyargulova, **"Quick detection of high-degree entities in large directed networks"**, IEEE International Conference on Data Mining (ICDM 2014), 14-17 Dec 2014, Shenzhen, China. pp. 20-29. IEEE Computer Society (2014) (arXiv:1410.0571v2[cs.SI])

4.2H. [53] P3.9 R.Palovics, F. Ayala-Gomez, B. Csikota, B.Daroczy, L. Kocsis, D. Spadacene, A.A. Benczur, **"RecSys Challenge 2014: an ensemble of binary classifiers and matrix factorization"**, Proceedings of the 2014 Recommender Systems Challenge (p. 13) ACM (2014)

5.2H. [70] P4.17 Sebastiano Vigna, **"A weighted correlation index for rankings with ties"**, Proceedings of the 24th international conference on World Wide Web, ACM (2015) (arXiv:1404.3325[cs.SI], 2014)

Joint publications for the whole grant period are: [8] P1.8 (P1 and P3);

[49] P2.12 (P2 and P4); [60] P3.16 (P1 and P3); [61] P3.17 (P2 and P3);

[69] P4.16 (P1 and P4).

1. Project objectives for the period

The overall aim of the project NADINE is to investigate modern directed networks by new tools developed by the project, determine network specific properties and characteristics, improve the tools and algorithms using obtained results, provide generic methods for directed network analysis and extract new features of modern directed networks in various sciences.

The main efforts have been devoted to milestones delivered for the second period including: network-specific centrality measures (M5, results added), fractal Weyl law properties of networks (M6), protocols for large-scale network processing (M7), characterization of multiproduct world trade network (M8), webcrawler development and database collection (M9), Monte Carlo algorithms for centrality measures (M10), delocalization conditions for Google matrix eigenstates (M11), new protocols for social voting and recommendation (M12), characterization of ranking of Wikipedia and other networks (M13), characterization of time-evolving Web structures (M14). Milestones M1-M5 had been delivered in the first period.

All scientific objectives of the second period, guided by the deliverable and milestones, are fulfilled.

2. Work progress and achievements during the period

WP1: CheiRank versus PageRank, centrality measures and network structure

Summary

The main objective of this work package is to lay mathematical foundations for development and application of new ranking schemes such as 2DRanking, and provide fast algorithms for their computation. PageRank is widely applied for ranking of nodes in directed networks including World Wide Web and citation graph. However, up to date, very little is known about mathematical properties of the resulting PageRank vector. The results of the consortium prove that the power law behaviour of PageRank is defined by the distribution of the in-degree. However, the dependence between these two quantities is remarkably different, e.g., for Web and Wikipedia. The partners also found that correlations of PageRank and CheiRank are small in some networks (e.g. for Linux kernel software and gene networks), and large in others (e.g. Web samples and Wikipedia). We will use novel methods, proposed by the consortium, to adequately measure correlations between node parameters, and obtain analytical description for 2DRanking, where these correlations are taken into account. We will extend our analysis to new centrality measures, of which desirable properties for specific network structures and applications will be justified by a mathematical model. Finally, our objective is to develop efficient Monte Carlo algorithms for evaluating centrality measures. Our results prove that such methods are remarkably efficient if the goal is to evaluate the ranking order, and not the exact values of centrality scores. Our aim is to evaluate the required computational complexity of Monte Carlo in order to produce an informative ranking order.

Detailed exposition of tasks

Task WP 1.1. Measuring and modelling network-specific dependencies between node parameters. (**UTWE**, CNRS, UMIL) Delivered in first period

Task WP 1.2. Analytical tools for the 2DRanking distribution in directed graphs. (**CNRS**, UTWE, MTA_SZTAKI, UMIL) Delivered in first period

Task WP 1.3. Design and analysis of new model-based centrality measures. (**UTWE**, CNRS, MTA_SZTAKI, UMIL) Delivered in first period

Task WP 1.4. Design and analysis of Monte Carlo algorithms for computation of importance measures. (**UTWE**, CNRS, MTA_SZTAKI, UMIL)

Milestone M10: Monte Carlo algorithms for centrality measures *(Reporting period: M36)*

We have developed fast randomized algorithms, based on random walks and random sampling [17] P2.5 for finding nodes with large degrees if the network structure is unknown, e.g. as in Twitter. Monte Carlo methods have also been developed and applied for evaluating the alpha-current flow betweenness. This work will be continued and extended to other centrality measures and correlation measures in large directed networks.

In [50] P2.13, we address the problem of quick detection of high-degree entities in large online social networks. Practical importance of this problem is attested by a large number of companies that continuously collect and update statistics about popular entities, usually using the degree of an entity as an approximation of its popularity. We suggest a simple, efficient, and easy to implement two-stage randomized algorithm that provides highly accurate solutions to this problem. For instance, our algorithm needs only one thousand API requests in order to find the top-100 most followed users, with more than 90 percent precision, in the online social network Twitter with approximately a billion of registered users. Our algorithm significantly outperforms existing methods and serves many different purposes such as finding the most popular users or the most popular interest groups in social networks. An important contribution of this work is the analysis of the proposed algorithm using 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 an accurate prediction for the algorithm's performance and show that the number of API requests for finding the top-most popular entities is sublinear in the number of entities. Moreover, we formally show that the high variability of the entities, expressed through heavy-tailed distributions, is the reason for the algorithm's efficiency. We quantify this phenomenon in a rigorous mathematical way

Special Issue on Searching an Mining the Web and Social Networks at Internet Mathematics has been prepared by P2 and P4 [49].

Publications M10:

[17] P2.5 L. Ostroumova, K. Avrachenkov and N. Litvak. "Quick detection of popular entities in large directed networks." Submitted Computer Science Conference, Oct 2013 (reported in period 1; extended version appeared in [50]) [M10-WP1.4]

[49] P2.12 N.Litvak and S.Vigna, "Introduction to Special Issue on Searching and Mining the Web and Social Networks", Internet Mathematics, v.10(3-4), p.219-221 (2014) [M1-M10-WP1.1-WP1.4]

[50] P2.13 K.Avrachenkov, N.Litvak, L.Ostroumova-Prokhorenkova and E.Suyargulova, "Quick detection of high-degree entities in large directed networks", IEEE International Conference on Data Mining (ICDM 2014), 14-17 Dec 2014, Shenzhen, China. pp. 20-29. IEEE Computer Society (2014) (arXiv:1410.0571v2[cs.SI]) {M10-WP1.4]

Deliverable D1.2: No deviations from the initial plan for this deliverable/milestones are reported.

WP2: Network analysis through Google matrix eigenspectrum and eigenstates:

Summary

WP2 investigates spectrum of Google matrices of such real networks as WWW university networks, network of hyperlinks between Wikipedia English articles, network of links of procedure call procedures in open source software. The Arnoldi method applied to the Linux network established the validity of fractal Weyl law, found recently in systems of quantum chaotic scattering and Perron-Frobenius operators of dynamical systems. WP2 investigates the spectrum of Wikipedia network analyzed recently. The eigenmodes, with eigenvalue modulus being close to the damping factor, correspond to slow relaxation modes in networks. Such modes should be linked with specific communities hidden inside network. The Arnoldi method allows to detect such modes in an effective way thus open new possibilities for extracting of hidden communities from networks. Examples of Google matrix spectrum for the WWW of Cambridge and Oxford Universities, obtained by the Arnoldi method, show decomposition of degenerate subspaces with $\lambda=1$. The size of degenerate subspaces can be rather large (around 40000 for the case of WWW of Cambridge University of total size 200000). This Arnoldi approach will be also applied for networks of Wikipedia articles, open source software networks, university networks. Fractal dimensions of the networks will be also determined. The Arnoldi method will be also used to detect communities, linked to eigenstates with eigenvalue close to one, in the Wikipedia articles network of N=3282257 nodes extending previous results. The high efficiency of the Arnoldi method allows to handle Google matrices of very large size using modern computers available to the consortium. Delocalization properties of eigenstates will also be determined for various networks.

Detailed exposition of tasks

Task WP 2.1. Random matrix models of Google type matrices (**CNRS**, UTWE, UMIL) Delivered in first period

Task WP 2.2. Eigenspectrum and eigenfunctions of Google matrix of directed networks (**CNRS**, UTWE, MTA_SZTAKI, UMIL) Delivered in first period

Task WP 2.3. Fractal dimensions and fractal Weyl law for directed networks (**CNRS**, UTWE, MTA_SZTAKI, UMIL)

Milestone M6: Fractal Weyl law properties of networks (WP2.3) *(Reporting period: M24-36)*

The fractal Weyl law had been established for Linux Kernel network and Ulam networks of dissipative dynamical maps [35] P1.15. Investigations of validity of this property for for the citation network of Physical Review [12] P1.12 established the fractal dimension to be close to unity (Weyl exponent of growth 0.5) while for the Linux kernel network the fractal dimension is approximately 1.3 (Weyl exponent of growth 0.65).

For the Ulam networks it is shown that the Weyl exponent is equal to a half of fractal dimension of the invariant set of the strange attractor of the dynamical system.

Publications

[12] P1.12 K.M.Frahm, Y.-H.Eom and D.L. Shepelyansky, "Google matrix of the citation network of Physical Review", submitted to Phys. Rev. E Oct 21, 2013 (arXiv:1310.5624 [physics.soc-ph], 2013); published Phys. Rev. E v.89, p.052814 (2014) [M6-WP2.3] [reported in period 1]

[35] P1.15 L.Ermann, K.M.Frahm and D.L.Shepelyansky, "Google matrix analysis ofdirected networks", submitted to Rev. Mod. Phys. (2014), arXiv:1409.0428 [physics.soc- ph]) [M6-WP2.3; M11-WP2.3-WP2.4]

Task WP 2.4. Localization and delocalization properties of Google matrix eigenstates (CNRS, UTWE, MTA_SZTAKI)

Milestone M11: Delocalization conditions for Google matrix eigenstates (WP2.4) *(Reporting period: M36)*

Various examples of matrices, directed networks and solid state models with the Anderson localization and the Anderson transition from localized (insulating phase) to delocalized (metalic phase) are analyzed in [35]. In [40] we introduce a number of random matrix models describing the Google matrix G of directed networks. The properties of their spectra and eigenstates are analyzed by numerical matrix diagonalization. We show that for certain models it is possible to have an algebraic decay of PageRank vector with the exponent similar to real directed networks. At the same time the spectrum has no spectral gap and a broad distribution of eigenvalues in the complex plain. The eigenstates of G are characterized by the Anderson transition from localized to delocalized states and a mobility edge curve in the complex plane of eigenvalues.

Publications

[35] P1.15 L.Ermann, K.M.Frahm and D.L.Shepelyansky, "Google matrix analysis ofdirected networks", submitted to Rev. Mod. Phys. (2014), arXiv:1409.0428 [physics.soc- ph]) [M6-WP2.3; M11-WP2.3-WP2.4]

[40] P1.20 O.V.Zhirov and D.L.Shepelyansky, "Anderson transition for Google matrix eigenstates", Ann. der Physik (Berlin) DOI 10.1002/andp.201500110 (2015) (arXiv:1501.03371[q-fin.ST]) [M11-WP2.4]

Deliverable D2.2: No deviations from the initial plan for this deliverable/milestones are reported.

WP3: Applications to voting systems in social networks

Summary

Voting is a basic decision procedure by which individuals express their preferences among a set of choices. Given the preferences of all voters (each one a permutation of the possible choices), a voting system generates a single choice. Voting theory studies how to select such a choice under certain optimization constraints. In particular, choices can be individuals that must be chosen for some purpose (e.g., to take a decision, or to represent the population). In *direct democracy*, each individual can vote any other individual. Recently, to obviate the lack of acquaintance between voter and voted individual in large populations, *liquid democracy* (a.k.a. *proxy voting*) has been introduced. In this case, a vote is given to some other individual that can keep it (and then we can perform an election just by majority) or give it away to someone else.

In social networks representing acquaintances between people (e.g., Facebook), however, we have a much more interesting scenario, as we are given from the start, for each individual, a set of users that are directly known (its neighbours in the graph). By restricting the ability to vote to acquaintances, we can obviate (even for very large networks) the problem of low representativity: if we give our vote to one of our acquaintances, we judge it apt to take a decision for us. Due to the large size of social networks (Facebook has currently more than 700 million active users), however, a direct application of liquid democracy can lead to a number of problems, most notably the loss of control of our vote: due to the small-world phenomenon, in a very small number of passages our vote can reach essentially any individual.

Recently, *viscous democracy* has been proposed for social networks by members of the consortium. Voters can only choose one of their neighbours, generating a *voting graph*—a directed graph of constant outdegree one. Each vote is passed to the chosen neighbour, but weakened by a multiplicative attenuation factor. If the vote travels too far, it is ineffective. It turns out that this is equivalent to computing Katz's index (or, in this case, due to the fixed outdegree of the graph, PageRank) on the voting graph—hence the name *spectral voting* for this kind of technique. Due to the known connection between path-based ranking and eigenvector-based ranking, the resulting scores turn out to be given by the dominant eigenvector of a suitable matrix

Detailed exposition of tasks

Task WP 3.1. Eigenvectors for spectral voting (**UMIL**, CNRS, MTA_SZTAKI) Delivered in first period

Task WP 3.2. Social voting analysis through centrality measures (**UMIL**, CNRS, UTWE, MTA_SZTAKI) Delivered in first period

Milestone M 5: Network specific centrality measures (WP1.1, WP1.3, WP3.1,WP3.2)*(Reporting period: M18)* Delivered in first period

Additional work left from period 1 on voting [M5-WP3.1-WP3.2] has been done and published in [71] P4.18. Most modern recommendation systems use the approach of collaborative filtering: users that are believed to behave alike are used to produce recommendations. In this work we describe an application (Liquid FM) taking a completely different approach. Liquid FM is a music recommendation system that makes the user responsible for the recommended items. Suggestions are the result of a voting scheme, employing the idea of viscous democracy. Liquid FM can also be thought of as the first teststbed for this voting system. In this paper we outline the design and architecture of the application, both from the theoretical and from the implementation viewpoints. The software has been developed and is now open for public via the web page https://github.com/corradomonti/fbvoting

Publications

[71] P4.18 Paolo Boldi, Corrado Monti, Massimo Santini, and Sebastiano Vigna, "Liquid FM: Recommending Music through Viscous Democracy", submitted to CoRR (2015) (arXiv:1503.08604[cs.SI], 2015) [M5-M12-WP3.1-3.2-3.3-3.4] AND [M5-WP3.1-WP3.2 promised to be finished in report 2; the software is open and available here https://github.com/corradomonti/fbvoting]

Task WP 3.3. Social network analysis through graph neighbourhood function (**MTA_SZTAKI**, UMIL, UTWE)

Milestone M 12: New protocols for social voting and recommendation (WP3.3, WP3.4)*(Reporting period: M36)*

In [41] we study a two states opinion formation model driven by PageRank node influence and report an extensive numerical study on how PageRank affects collective opinion formations in large-scale empirical directed networks. In our model the opinion of a node can be updated by the sum of its neighbor nodes' opinions weighted by the node influence of the neighbor nodes at each step. We consider PageRank probability and its sublinear power as node influence measures and investigate evolution of opinion under various conditions. First, we observe that all networks reach steady state opinion after a certain relaxation time. This time scale is decreasing with the heterogeneity of node influence in the networks. Second, we find that our model shows consensus and non-consensus behavior in steady state depending on types of networks: Web graph, citation network of physics articles, and LiveJournal social network show non-consensus behavior while Wikipedia article network shows consensus behavior. Third, we find that a more heterogeneous influence distribution leads to a more uniform opinion state in the cases of Web graph, Wikipedia, and Livejournal. However, the opposite behavior is observed in the citation network. Finally we identify that a small number of influential nodes can impose their own opinion on significant fraction of other nodes in all considered networks. Our study shows that the effects of heterogeneity of node influence on opinion formation can be significant and suggests further investigations on the interplay between node influence and collective opinion in networks.

The work [71] P4.18, described above, is also reported in this WP3.3.

In [72] we consider the following problem. Besides finding trends and unveiling typical patterns, modern information retrieval is increasingly more interested in the discovery of surprising information in textual datasets. In this work we focus on finding unexpected links in hyper- linked document corpora when documents are assigned to categories; our approach is based on the determination of a latent category matrix that explains common links; the matrix is built using a perceptron-like technique. We show that our method provides better accuracy than most existing text-based techniques, with higher efficiency and relying on a much smaller amount of information. It also provides higher precision than standard link prediction, especially at low recall levels; the two methods are in fact shown to be orthogonal and can therefore be fruitfully combined.

Publications

[41] P1.21 Young-Ho Eom and D.L.Shepelyansky, "Opinion formation driven by PageRank node influence on directed networks", submitted to Physica A Feb (2015) (arXiv:1502.02567[physics.soc-ph]) [M12-WP3.3,WP3.4]

[71] P4.18 Paolo Boldi, Corrado Monti, Massimo Santini, and Sebastiano Vigna, "Liquid FM: Recommending Music through Viscous Democracy", submitted to CoRR (2015) (arXiv:1503.08604[cs.SI], 2015) [M5-M12-WP3.1-3.4] AND [M5-WP3.1-WP3.2 promised to be finished in report 2; the software is open and available here https://github.com/corradomonti/fbvoting]

[72] P4.19 Paolo Boldi and Corrado Monti, "LlamaFur: Learning Latent Category Matrix to Find Unexpected Relations in Wikipedia", preprint submitted to CoRR (2015) [M12-WP3.3,WP3.4] available on request because of double-blind submission [M12-WP3.3,WP3.4]

Task WP 3.4. Recommendation systems in social networks (MTA_SZTAKI, UMIL, CNRS)

Milestone M 12: New protocols for social voting and recommendation (WP3.3, WP3.4)*(Reporting period: M36)*

In [53] we give our solution to the RecSys Challenge 2014. In our ensemble we use (1) a mix of binary classication methods for predicting nonzero engagement, including logistic regression and SVM; (2) regression methods for directly predicting the engagement, including linear regression and gradient boosted trees; (3) matrix factorization and factorization machines over the user-movie matrix, by using user and movie features as side information. For most of the methods, we use the GraphLab Create implementation. Our current nDCG achieves 0.877.

Properties of recommendation systems are investigated in [54]. Recommender systems often deal with a large amount of sequential data. For these scenarios, online matrix factorization techniques based on online prediction and incremental updates are often the most promising approaches. Decentralizing the system and keeping the user data on their devices is an important step in the direction of preserving user privacy. In this paper we propose a peer-to-peer online matrix factorization algorithm that stores the ratings of a user and her private data local. Additionally, the users have a local copy of the common part of the factor model and communicate with other users to advance towards a consensus on it. The algorithm is proven to converge to a set of local optima in the stationary case, while we show empirically that the algorithm performs well in the non-stationary case, both in terms of ranking performance and privacy preservation.

The recommendation systems are applied for analysis of the NOMAO data sets for voting of users (about 1 million) for spots (hotels, restaurants etc, about 20000 items) of Paris and France [60]. We explore the spectrum and the eigenvalues of a matrix containing user ratings to geolocalized items. Eigenvalues nicely map to large towns and regions but show certain level of instability as we modify the interpretation of the underlying matrix. We evaluate imputation strategies that reach improved prediction performance by reaching geographically smooth eigenvectors.

The data for Twitter network are analyzed in [61]. TO ADD

Publications

[53] P3.9 R.Palovics, F. Ayala-Gomez, B. Csikota, B.Daroczy, L. Kocsis, D. Spadacene, A.A. Benczur, "RecSys Challenge 2014: an ensemble of binary classifiers and matrix factorization", Proceedings of the 2014 Recommender Systems Challenge (p. 13) ACM (2014) [M12-WP3.3,WP3.4]

[54] P3.10 Andrea N. Ban, Levente Kocsis, Robert Palovics, "Peer-to-peer Online Collaborative Filtering", preprint (2015) [M12-WP3.3,WP3.4]

[60] P3.16 Robert Palovics, Balint Daroczym Andras A. Benczur, Julia Pap, Leonardo Ermann, Samuel Phan, Alexei D. Chepelianskii, Dima L. Shepelyansky, "Statistical analysis of NOMAO customer votes for spots of France", preprint arXiv (2015) [M12-WP3.3,WP3.4]

[61] Andras A. Benczur, Nelly Litwak et al., "Analysis of Twitter network ", preprint arXiv:1504.XXXX (2015) [M12-WP3.3,WP3.4]

Deliverable D3.2: No deviations from the initial plan for this deliverable/milestones are reported.

WP4: Applications of new tools and algorithms to real-world network structures:

Summary

Methods of WP1-WP3 are implemented in large scale applications based on real data collected in WP5. Achievements in this WP are measured in terms of: the size of the data processed, with WP targets at Web scale, billions of objects; another benchmark is the approximation error of the fingerprinting and lazy update procedures, with the target to keep the error below the limit of notice in a user application. Special distributed network technologies will be developed to reach such goals. Spam filtering protocols will also be developed and tested. Using these tools and those of WP1-WP3, statistical analysis will be done for several types of important networks including Wikipedia in English, French, German, Italian and Spanish at different moments of time evolution; open software procedure networks, genes and other networks. Applications of centrality measures to game theory will also be developed. We will generalize recent results for the Google matrix of world trade network to the case of multiproduct trade for which the matrix size is increased by two or more orders of magnitude.

Detailed exposition of tasks

Task WP 4.1. Distributed network processing technologies.(**MTA_SZTAKI**, UMIL, UTWE)

Milestone M 7: Protocols for large-scale network processing (WP4.1, WP5.2)*(Reporting period: M24-36)*

The size of available networks pushes towards new algorithms (typically, approximate or distributed) and new computational frameworks (e.g., MapReduce, NoSQL and streaming data). In our experiments, algorithms over large graphs that cannot be fit into the internal memory be solved using algorithms with three different distributed computing paradigms: Distributed key-value stores, Map-Reduce and Bulk Synchronous Parallel.

Initial results have been reported in period 1 in [23]. The research have been continued in updated analysis presented in [63]. The computation of a peeling order in a randomly generated hypergraph is the most time-consuming step in a number of constructions, such as perfect hashing schemes, random r-SAT solvers, error-correcting codes, and approximate set encodings. While there exists a straightforward linear time algorithm, its poor I/O performance makes it impractical for hypergraphs whose size exceeds the available internal memory. We show how to reduce the computation of a peeling order to a small number of sequential scans and sorts, and analyze its I/O complexity in the cache-oblivious model. The resulting algorithm requires O(sort(n)) I/Os and O(n log n) time to peel a random hypergraph with n edges. We experimentally evaluate the performance of our implementation of this algorithm in a real-world scenario by using the construction of minimal perfect hash functions (MPHF) as our test case: our algorithm builds a MPHF of 7.6 billion keys in less than 21 hours on a single machine. The resulting data structure is both more space-efficient and faster than that obtained with the current state-of-the-art MPHF construction for large-scale key sets.

In [64] we analyse a network model characterized by a latent attribute structure with competition. The quest for a model that is able to explain, describe, analyze and simulate real-world complex networks is of uttermost practical as well as theoretical interest. In this paper we introduce and study a network model that is based on a latent attribute structure: each node is characterized by a number of features and the probability of the existence of an edge between two nodes depends on the features they share. Features are chosen according to a process of Indian-Bu et type but with an additional random “fitness" parameter attached to each node, that determines its ability to transmit its own features to other nodes. As a consequence, a node's connectivity does not depend on its age alone, so also “young" nodes are able to compete and succeed in acquiring links. One of the advantages of our model for the latent bipartite “node-attribute" network is that it depends on few parameters with a straightforward interpretation. We provide some theoretical, as well experimental, results regarding the power-law behavior of the model and the estimation of the parameters. By experimental data, we also show how the proposed model for the attribute structure naturally captures most local and global properties (e.g., degree distributions, connectivity and distance distributions) real networks exhibit.

In [65] we study entity-linking via graph-distance minimization. Entity-linking is a natural-language–processing task that consists in identifying the entities mentioned in a piece of text, linking each to an appropriate item in some knowledge base; when the knowledge base is Wikipedia, the problem comes to be known as wikification in this case, items are wikipedia articles). One instance of entity-linking can be formalized as an optimization problem on the underlying concept graph, where the quantity to be optimized is the average distance between chosen items. Inspired by this application, we define a new graph problem which is a natural variant of the Maximum Capacity Representative Set. We prove that our problem is NP-hard for general graphs; nonetheless, under some restrictive assumptions, it turns out to be solvable in linear time. For the general case, we propose two heuristics: one tries to enforce the above assumptions and another one is based on the notion of hitting distance; we show experimentally how these approaches perform with respect to some baselines on a real-world dataset.

An additional large part of results for M7 is presented in MP5.

Publications

[23] P3.5 Real-time streaming mobility analytics. Andras Garzo, Andras A. Benczur, Csaba Istvan Sidlo, Daniel Tahara, Erik Francis Wyatt. Conference IEEE Big Data 2013 [M7-WP4.1] [reported period 1]

[63] P4.10 Djamal Belazzougui, Paolo Boldi, Giuseppe Ottaviano, Rossano Venturini, and Sebastiano Vigna, "Cache-oblivious peeling of random hypergraphs", 2014 Data Compression Conference (DCC 2014), IEEE pp.352-361. (2014) [M7-WP4.1]

[64] P4.11 Paolo Boldi, Irene Crimaldi, and Corrado Monti, "A network model characterized by a latent attribute structure with competition", submitted CoRR (2014), (arXiv:1407.7729[cs.SI], 2014 [M7-WP4.1]

[65] P4.12 Roi Blanco, Paolo Boldi, and Andrea Marino, **"**Entity-linking via graph-distance minimization**"**, Proceedings 3rd Workshop on GRAPH Inspection and Traversal

Engineering, GRAPHITE 2014, Grenoble, France, 5th April 2014., pp.30-43 (2014) [M7-WP4.1]

Task WP 4.2. Network quality and trust classification and spam filtering (**MTA_SZTAKI**, UMIL, UTWE) Delivered in first period

Task WP 4.3. 2DRanking and centrality measures of Wikipedia, open software and other networks(CNRS, UMIL, UTWE, MTA_SZTAKI**)**

Milestone M13: Characterization of ranking of Wikipedia and other networks (WP4.3-WP5.2)*(Reporting period: M36)*

There are a large number of directed networks investigated in M13.

In [33] we study the structural properties of the neural network of the C.elegans (worm) from a directed graph point of view. The Google matrix analysis is used to characterize the neuron connectivity structure and node classifications are discussed and compared with physiological properties of the cells. Our results are obtained by a proper definition of neural directed network and subsequent eigenvector analysis which recovers some results of previous studies. Our analysis highlights particular sets of important neurons constituting the core of the neural system. The applications of PageRank, CheiRank and ImpactRank to characterization of interdependency of neurons are discussed.

In [34] we use the methods of quantum chaos and Random Matrix Theory for analysis of statistical fluctuations of PageRank probabilities in directed networks. In this approach the effective energy levels are given by a logarithm of PageRank probability at a given node. After the standard energy level unfolding procedure we establish that the nearest spacing distribution of PageRank probabilities is described by the Poisson law typical for integrable quantum systems. Our studies are done for the Twitter network and three networks of Wikipedia editions in English, French and German. We argue that due to absence of level repulsion the PageRank order of nearby nodes can be easily interchanged. The obtained Poisson law implies that the nearby PageRank probabilities fluctuate as random independent variables

In the review paper [35] the overview of the Google matrix analysis of directed networks is presented. This review describes the Google matrix analysis of directed complex networks demonstrating its eciency on various examples including World Wide Web, Wikipedia, software architecture, world trade, social and citation networks, brain neural networks, DNA sequences and Ulam networks. The analytical and numerical matrix methods used in this analysis originate from the fields of Markov chains, quantum chaos and Random Matrix theory.

The friendship paradox states that your friends have on average more friends than you have. The detailed study of thie phenomenon is presented in [36,37]. Does the paradox ‘‘hold’’ for other individual characteristics like income or happiness? To address this question, we generalize the friendship paradox for arbitrary node characteristics in complex networks. By analyzing two co-authorship networks of

Physical Review journals and Google Scholar profiles, we find that the generalized friendship paradox (GFP) holds at the individual and network levels for various characteristics, including the number of coauthors, the number of citations, and the number of publications. The origin of the GFP is shown to be rooted in positive correlations between degree and characteristics. As a fruitful application of the GFP, we suggest effective and efficient sampling methods for identifying high characteristic nodes in large-scale networks. Our study on the GFP can shed lights on understanding the interplay between network structure and node characteristics in complex networks. The paper [36] is highlighted by Independent and MIT Tech. Rev.

In [38] we analyze the game of go from the point of view of complex networks. We construct three different directed networks of increasing complexity, defining nodes as local patterns on plaquettes of increasing sizes, and links as actual successions of these patterns in databases of real games. We discuss the peculiarities of these networks compared to other types of networks. We explore the ranking vectors and community structure of the networks and show that this approach enables to extract groups of moves with common strategic properties. We also investigate different networks built from games with players of different levels or from different phases of the game. We discuss how the study of the community structure of these networks may help to improve the computer simulations of the game. More generally, we believe such studies may help to improve the understanding of human decision process.

In [44] we introduce, and analyze, three measures for degree-degree dependencies, also called degree assortativity, in directed random graphs, based on Spearman’s rho and Kendall’s tau. We proof statistical consistency of these measures in general random graphs and show that the directed Configuration Model can serve as a null model for our degree-degree dependency measures. Based on these results we argue that the measures we introduce should be preferred over Pearson’s correlation coefficients, when studying degree-degree dependencies, since the latter has several issues in the case of large networks with scale-free degree distribution.

In [45] analysis of degree-degree dependencies in complex networks, and their impact on processes on networks requires null models, i.e. models that generate uncorrelated scale-free networks. Most models to date however show structural negative dependencies, caused by finite size effects. We analyze the behavior of these structural negative degree-degree dependencies, using rank based correlation measures, in the directed Erased Configuration Model. We obtain expressions for the scaling as a function of the exponents of the distributions. Moreover, we show that this scaling undergoes a phase transition, where one region exhibits scaling related to the natural cut-off of the network while another region has scaling similar to the structural cut-off for uncorrelated networks. By establishing the speed of convergence of these structural dependencies we are able to asses statistical significance of degree-degree dependencies on finite complex networks when compared to networks generated by the directed Erased Configuration Model.

In [46] we model user behaviour in Twitter to capture the emergence of trending topics. For this purpose, we first extensively analyse tweet datasets of several different events. In particular, for these datasets, we construct and investigate the retweet graphs. We find that the retweet graph for a trending topic has a relatively dense largest connected component (LCC). Next, based on the insights obtained from the analyses of the datasets, we design a mathematical model that describes the evolution of a retweet graph by three main parameters. We then quantify, analytically and by simulation, the in influence of the model parameters on the basic characteristics of the retweet graph, such as the density of edges and the size and density of the LCC. Finally, we put the model in practice, estimate its parameters and compare the resulting behavior of the model to our datasets.

In [62] we revisit the graph structure in the Web. Knowledge about the general graph structure of the World Wide Web is important for understanding the social mechanisms that govern its growth, for designing ranking methods, for devising better crawling algorithms, and for creating accurate models of its structure. In this paper, we describe and analyse a large, publicly accessible crawl of the web that was gathered by the Common Crawl Foundation in 2012 and that contains over3.5 billion web pages and 128.7 billion links. This crawl makes it possible to observe the evolution of the underlying structure of the World Wide Web within the last 10 years: we analyse and compare, among other features, degree distributions, connectivity, average distances, and the structure of weakly/strongly connected components. Our analysis shows that, as evidenced by previous research, some of the features previously observed by Broder et al. Are very dependent on artefacts of the crawling process, whereas other appear to be more structural. We confirm the existence of a giant strongly connected component; we however find, as observed by other researchers, very different proportions of nodes that can reach or that can be reached from the giant component, suggesting that the “bow-tie structure” as described previously is strongly dependent on the crawling process, and to the best of our current knowledge is not a structural property of the web. More importantly, statistical testing and visual inspection of size-rank plots show that the distributions of indegree, outdegree and sizes of strongly connected components are not power laws, contrarily to what was previously reported for much smaller crawls, although they might be heavy tailed. We also provide for the first time accurate measurement of distance-based features, using recently introduced algorithms that scale to the size of our crawl.

Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions is analysed in [69] highlighted by Guardian, Washington post, EC CORDIS and other press of about 20 countries (see press.htm on the NADINE web page) and Wikipedia article “Top 100 historical figures of Wikipedia” [43]). In [69] , we apply methods of Markov chainsand Google matrix for the analysis of the hyperlink networks of 24 Wikipedia language editions, and rank all their articles by PageRank, 2DRank and CheiRank algorithms. Using automatic extraction of people names, we obtain the top 100 historical figures, for each edition and for each algorithm. We investigate their spatial, temporal, and gender distributions in dependence of their cultural origins. Our study demonstrates not only the existence of skewness with local figures, mainly recognized only in their own cultures, but also the existence of global historical figures appearing in a large number of editions. By determining the birth time and place of these persons, we perform an analysis of the evolution of such figures through 35 centuries of human history for each language, thus recovering interactions and entanglement of cultures over time. We also obtain the distributions of historical figures over world countries, highlighting geographical aspects of cross-cultural links. Considering historical figures who appear in multiple editions as interactions between cultures, we construct a network of cultures and identify the most influential cultures according to this network. This research line is in competition with two groups at MIT (e.g. Pantheon project of Hidalgo) and one group at Stony-Brook NY S.Skiena).

Understanding the correlation between two different scores for the same set of items is a common problem in information retrieval, and the most commonly used statistics that quantifies this correlation is Kendall’s. However, the standard definition fails to capture that discordances between items with high rank are more important than those between items with low rank. Recently, a new measure of correlation based on average precision has been proposed to solve this problem, but like many alternative proposals in the literature it assumes that there are no ties in the scores. This is a major deficiency in a number of contexts, and in particular while comparing centrality scores on large graphs, as the obvious baseline, indegree, has a very large number of ties in web and social graphs. We propose [70] to extend Kendall’s definition in a natural way to take into account weights in the presence of ties. We prove a number of interesting mathematical properties of our generalization and describe an O(n log n) algorithm for its computation. We also validate the usefulness of our weighted measure of correlation using experimental data.

Local ranking problem on the BrowseGraph is analysed in [73]. The Local Ranking Problem" (LRP) is related to the computation of a centrality-like rank on a local graph, where the scores of the nodes could significantly differ from the ones computed on the global graph. Previous work has studied LRP on the hyperlink graph but never on the BrowseGraph, namely a graph where nodes are webpages and edges are browsing transitions. Recently, this graph has received more and more attention in many different tasks such as ranking, prediction and recommendation. However, a webserver has only the browsing traffic performed on its pages (local BrowseGraph) and, as a consequence, the local computation can lead to estimation errors, which hinders the increasing number of applications in the state of the art. Also, although the divergence between the local and global ranks has been measured, the possibility of estimating such divergence using only local knowledge has been mainly over-looked. These aspects are of great interest for online service providers who want to gauge their ability to correctly assess the importance of their resources only based on their local knowledge, and by taking into account real user browsing fluxes that better capture the actual user interest than the static hyperlink network. We study the LRP problem on a BrowseGraph from a large news provider, considering as subgraphs the aggregations of browsing traces of users coming from different domains. We show that the distance between rankings can be accurately predicted based only on structural information of the local graph, being able to achieve an average rank correlation as high as 0.8.

Publications M13:

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

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

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

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

[37] P1.17 Hong-Hyun Jo and Young-Ho Eom, "Generalized friendship paradox in networks with tunable degree-attribute correlation", Phys. Rev. E v.90, p.022809 (20124) [M13-WP4.3]

[38] P1.18 V.Kandiah, B.Georgeot and O.Giraud, "More ordering and communities in complex networks describing the game of go", Eur. Phys. J. B v.87, p.246 (20124) [M13- WP4.3]

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

[44] P2.7 P. van der Hoorn and N. Litvak, "Convergence of rank based degree-degree correlations in random directed networks", to appear in Moscow Journal of Combinatorics (2015) (arXiv:1407.7662[math.PR], 2014) [M13-WP4.3]

[45] P2.8 P. van der Hoorn and N. Litvak, "Phase transitions for scaling of structural

correlations in directed networks", (arXiv:1504.01535[physics.soc-ph], 2015 [M13-WP4.3]

[46] P2.9 M. Ten Thij, T. Ouboter, D. Worm, N. Litvak, J.L. van den Berg and S. Bhulai, Modelling of trends in Twitter using retweet graph dynamics, Proceedings 11th International Workshop Algorithms and Models for the Web Graph, WAW 2014, 17-18 Dec 2014, Beijing, China. pp. 132-147; Lecture Notes in Computer Science 2014 (8882), Springer (2014), (arXiv:1502.00166[cs.SI], 2015) [M13-WP4.3]

[62] P4.9 Robert Meusel, Sebastiano Vigna, Oliver Lehmberg, and Christian Bizer, "Graph structure in the web - Revisited, or a trick of the heavy-tail", WWW'14 Companion, pp.427-432, International World Wide Web Conferences Steering Committee, 2014; a revised version is to appear in the Journal of Web Science (2015) [M13-WP4.3]

[69] P4.16 Young Ho Eom, Pablo Aragon, David Laniado, Andreas Kaltenbrunner, Sebastiano Vigna, and Dima L. Shepelyansky, "Interactions of cultures and top people of Wikipedia from ranking of 24 language editions", PLoS ONE v.10(3), p.e0114825 (2015) (arXiv:1405.7183[cs.SI], 2014) [M13-WP4.3-WP5.2]

[70] P4.17 Sebastiano Vigna, "A weighted correlation index for rankings with ties", Proceedings of the 24th international conference on World Wide Web, ACM (2015) (arXiv:1404.3325[cs.SI], 2014) [M13-WP4.3-WP5.2]

[73] P4.20 Michele Trevisio, Luca Maria Aiello, Paolo Boldi and Roi Blanco, "Local Ranking Problem on the BrowseGraph", accepted for publication in SIGIR (2015) [M13-WP4.3-WP5.2]

Task WP 4.4. Analysis of Google matrix of multiproduct world trade network (**CNRS**, UTWE, MTA_SZTAKI)

Milestone M8: Characterization of multiproduct world trade network (WP4.4)*(Reporting period: M24-36)*

The Google matrix analysis of the world trade network from UN COMTRADE was first performed by P1 in 2011. The extention of this analysis to multiproduct trade meets certain mathematical difficulties which are resolved in this milestone.

The ecological analysis of the world trade is developed in [2] (particlly reported in 1st period). Ecological systems have a high complexity combined with stability and rich biodiversity. The analysis of their properties uses a concept of mutualistic networks and provides a detailed understanding of their features being linked to a high nestedness of these networks. Using the United Nations COMTRADE database we show that a similar ecological analysis gives a valuable description of the world trade: countries and trade products are analogous to plants and pollinators, and the whole trade network is characterized by a high nestedness typical for ecological networks. Our approach provides new mutualistic features of the world trade.

In [39] Using the United Nations COMTRADE database [United Nations Commodity Trade Statistics Database, available at http://comtrade.un.org/db/ we construct the Google matrix G of multiproduct world trade between the UN countries and analyze the properties of trade flows on this network for years 1962-2010. This construction, based on Markov chains, treats all countries on equal democratic grounds independently of their richness and at the same time it considers the contributions of trade products proportionally to their trade volume. We consider the trade with 61 products for up to 227 countries. The obtained results show that the trade contribution of products is asymmetric: some of them are export oriented while others are import oriented even if the ranking by their trade volume is symmetric in respect to export and import after averaging over all world countries. The construction of the Google matrix allows to investigate the sensitivity of trade balance in respect to price variations of products, e.g. petroleum and gas, taking into account the world connectivity of trade links. The trade balance based on PageRank and CheiRank probabilities highlights the leading role of China and other BRICS countries in the world trade in recent years. We also show that the eigenstates of G with large eigenvalues select specific trade communities.

In [42] Using the new data from the OECD-WTO world network of economic activities we construct the Google matrix G of this directed network and perform its detailed analysis. The network contains 58 countries and 37 activity sectors for years 1995 and 2008. The construction of G, based on Markov chain transitions, treats all countries on equal democratic grounds while the contribution of activity sectors is proportional to their exchange monetary volume. The Google matrix analysis allows to obtain reliable ranking of countries and activity sectors and to determine the sensitivity of CheiRank-PageRank commercial balance of countries in respect to price variations and labor cost in various countries. We demonstrate that the developed approach takes into account multiplicity of network links with economy interactions between countries and activity sectors thus being more efficient compared to the usual export-import analysis. The spectrum and eigenstates of G are also analyzed being related to specific activity communities of countries. This research is done in collaboration with the World Trade Organization at Geneve (H.Escaith).

Publications M8:

[2] P1.2 L.Ermann and D.L.Shepelyansky, "Ecological analysis of world trade", Phys. Lett. A v.377, p.250-256 (2013), arXiv:1201.3584[q-fin.GN] [M8-WP4.4 – reported in period 1]

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

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

Deliverable D4.2: No deviations from the initial plan for this deliverable/milestones are reported.

WP5: Database development of real-world networks

Summary

WP5 develops efficient protocols for large scale network analysis and generates database collections that will be treated by the methods developed in WP1-WP3. To this aim specific skilful crawlers will be developed to collect information from modern enormous data bases. Data sets evolving in time will be analyzed by specially developed protocols.

Detailed exposition of tasks

Task WP 5.1. Crawler development and database collection (**UMIL**, MTA_SZTAKI, CNRS)

Milestone M 9: Webcrawler development and database collection (WP5.1)*(Reporting period: M24-36)*

This task and milestone was reported in period 1 in [32] P4.8.

Task WP 5.2. Internet Scale Data Management (MTA_SZTAKI, UMIL, UTWE)

Milestone M 7: Protocols for large-scale network processing (WP4.1, WP5.2)*(Reporting period: M24-36)*

In [47] we analyze the distribution of PageRank on a directed configuration model and show that as the size of the graph grows to infinity it can be closely approximated by the PageRank of the root node on an appropriately constructed tree, This tree approximation is in turn related to the solution of a linear stochastic fixed point equation that has been thoroughly studied in the recent literature.

The paper [48] studies the distribution of a family of rankings, which includes Google's PageRank, on a directed configuration model. In particular, it is shown that the distribution of the rank of a randomly chosen node in the graph converges in distribution to a finite random variable R* that can be written as a linear combination of i.i.d. copies of the endogenous solution to a stochastic fixed point equation of a specific form. Moreover, we provide precise asymptotics for the limit R*, which when the in-degree distribution in the directed configuration model has a power law imply a power law distribution for R* with the same exponent.

At the RecSys 2014 Workshop on Large Scale Recommender Systems [51], we presented our idea on broadcasting identical node values to all graph neighbors to speed up distributed algorithms with communication patterns as simple as PageRank to as complex as Alternating Least Squares. SZTAKI PI will be co-organizer for this workshop in 2015.

In [52] we analyse Similarity Kernel learning. Kernel methods are popular in machine learning tasks. For Support Vector Machine classification or Support Vector Regression, the central question is the selection of the appropriate kernel. The task is dicult in particular if the data points have complex or multimodal attributes such as time series or visual content enhanced with geographic, numeric or text metadata. Unlike earlier approaches of the so-called Multiple Kernel Learning problem, where a large number of kernels are fused by wrapper methods as part of the optimization process, in this paper we mathematically derive an optimal kernel for the data set in question. We begin with selecting appropriate distances for the appropriate modalities, for example dynamic time warping distance for time series and Jensen-Shannon distance for the bag of words text representation. Our kernel is defined, without needs of wrapper methods, by considering the distances as attributes generated by a Markov Random Field. For the Markov Random Field, the natural kernel is based on the Fisher information matrix and its exact form can be computed from the data. We experiment with the above similarity kernel over a wide variety of data sets, including: 64-channel EEG data; General time series data sets; Images with text annotations; Web documents; Gene expression level. Over the complex, multimodal or multiple time series classification tasks, our method outperforms the state of the art while reaching identical performance even over the simple unimodal problems as well, hence our method seems applicable under very general settings.

In [66] step-asynchronous successive overrelaxation updates the values contained in a single vector using the usual Gau\ss-Seidel-like weighted rule, but arbitrarily mixing old and new values, the only constraint being temporal coherence: you cannot use a value before it has been computed. We show that given a nonnegative real matrix A, a $\sigma\geq\rho(A)$ and a vector $\boldsymbol{w}>0$ such that $A\boldsymbol{w}\leq\sigma\boldsymbol{w}$, every iteration of step-asynchronous successive overrelaxation for the problem $(sI-A)\boldsymbol{x}=\boldsymbol{b}$, with $s>\sigma$, reduces geometrically the $\boldsymbol{w}$-norm of the current error by a factor that we can compute explicitly. Then, we show that given a $\sigma>\rho(A)$ it is in principle always possible to compute such a $\boldsymbol{w}$. This property makes it possible to estimate the supremum norm of the absolute error at each iteration without any additional hypothesis on A, even when A is so large that computing the product $A\boldsymbol{x}$ is feasible, but estimating the supremum norm of $(sI-A)-1$ is not.

In [67] : Marsaglia proposed recently xorshift generators as a class of very fast, good-quality pseudorandom number generators. Subsequent analysis by Panneton and L'Ecuyer has lowered the expectations raised by Marsaglia's paper, showing several weaknesses of such generators, verified experimentally using the TestU01 suite. Nonetheless, many of the weaknesses of xorshift generators fade away if their result is scrambled by a non-linear operation (as originally suggested by Marsaglia). In this paper we explore the space of possible generators obtained by multiplying the result of a xorshift generator by a suitable constant. We sample generators at 100 equispaced points of their state space and obtain detailed statistics that lead us to choices of parameters that improve on the current ones. We then explore for the first time the space of high-dimensional xorshift generators, following another suggestion in Marsaglia's paper, finding choices of parameters providing periods of length 21024−1 and 24096−1. The resulting generators are of extremely high quality, faster than current similar alternatives, and generate long-period sequences passing strong statistical tests using only eight logical operations, one addition and one multiplication by a constant.

In [68]: xorshift* generators are a variant of Marsaglia's xorshift generators that eliminate linear artifacts typical of generators based on **Z**/2**Z**-linear operations using multiplication by a suitable constant. Shortly after high-dimensional xorshift* generators were introduced, Saito and Matsumoto suggested a different way to eliminate linear artifacts based on addition in **Z**/232**Z**, leading to the XSadd generator. Starting from the observation that the lower bits of XSadd are very weak, as its reverse fails systematically several statistical tests, we explore xorshift+, a variant of XSadd using 64-bit operations, which leads, in small dimension, to extremely fast high-quality generators.

Publications

[47] P2.10 N.Chen, N.Litvak and M.Olvera-Cravioto, "PageRank in scale-free random graphs", Proceedings 11th International Workshop Algorithms and Models for the Web

Graph, WAW 2014, 17-18 Dec 2014, Beijing, China pp. 120-131, Lecture Notes in Computer Science 2014 (8882), Springer (2014). (arXiv:1408.3610[math.PR], 2014 [M7- WP5.2]

[48] P2.11 N.Chen, N.Litvak and M.Olvera-Cravioto, "Ranking algorithms on directed configuration networks", Submitted to Random Structures and Algorithms (2014) (arXiv:1409.7443v2[math.PR], 2014) [M7-WP5.2]

[51] P3.7 Marton Balassi, Robert Palovics and Andras A. Benczur, "Distributed Frameworks for Alternating Least Squares (Poster presentation)", Large-Scale Recommender Systems in conjunction with RecSys, Foster City, Silicon Valley, USA, 6th-10th October 2014

[52] P3.8 Balint Daroczy, Krisztian Buza, Andras A. Benczur, "Similarity Kernel Learning", preprint (2015) [M7-WP5.2]

[66] P4.13 Sebastiano Vigna, "Supremum-norm convergence for step-asynchronous successive overrelaxation on M-matrices", submitted to CoRR (2014) (arXiv:1404.3327[cs.DS], 2014) [M7-WP5.2]

[67] P4.14 Sebastiano Vigna, "An experimental exploration of Marsaglia's xorshift generators, scrambled", submitted to CoRR (2014) (arXiv:1402.6246v2[cs.DS], 2014) [M7-WP5.2]

[68] P4.15 Sebastiano Vigna, "Further scramblings of Marsaglia's xorshift generators", submitted to CoRR (2014) (arXiv:1403.0930[cs.NI], 2014) [M7-WP5.2]

Milestone M13: Characterization of ranking of Wikipedia and other networks (WP4.3-WP5.2)*(Reporting period: M36)*

Publications on this milestone are described in WP4.3, they include [69], [70], [73].

Publications

[69] P4.16 Young Ho Eom, Pablo Aragon, David Laniado, Andreas Kaltenbrunner, Sebastiano Vigna, and Dima L. Shepelyansky, "Interactions of cultures and top people of Wikipedia from ranking of 24 language editions", PLoS ONE v.10(3), p.e0114825 (2015) (arXiv:1405.7183[cs.SI], 2014) [M13-WP4.3-WP5.2]

[70] P4.17 Sebastiano Vigna, "A weighted correlation index for rankings with ties", Proceedings of the 24th international conference on World Wide Web, ACM (2015) (arXiv:1404.3325[cs.SI], 2014) [M13-WP4.3-WP5.2]

[73] P4.20 Michele Trevisio, Luca Maria Aiello, Paolo Boldi and Roi Blanco, "Local Ranking Problem on the BrowseGraph", accepted for publication in SIGIR (2015) [M13-WP4.3-WP5.2]

Task WP 5.3. Cross-data and temporal Web analytics (MTA_SZTAKI, UMIL, CNRS**)**

Milestone M 14: Characterization of time evolving Web structures (WP5.3) *(Reporting period: M24-36)*

In paper [55] we give methods for time-aware music recommendation in a social media service with the potential of exploiting immediate temporal influences between users. We consider events when a user listens to an artist the first time and this event follows some friend listening to the same artist short time before. We train a blend of matrix factorization methods that model the relation of the in influencer, the in influenced and the artist, both the individual factor decompositions and their weight learned by variants of stochastic gradient descent (SGD). Special care is taken since events of in influence form a subset of the positive implicit feedback data and hence we have to cope with two different definitions of the positive and negative implicit training data. In addition, in the time-aware setting we have to use online learning and evaluation methods. While SGD can easily be trained online, evaluation is cumbersome by traditional measures since we will have potentially di erent top recommendations at different times. Our experiments are carried over the two-year “scrobble" history of 70,000 Last.fm users and show a 5percent increase in recommendation quality by predicting temporal in influences.

In [56] we compare machine learning methods to predict quality aspects of the C3 dataset collected as a part of the Reconcile project. We give methods for automatically assessing the credibility, presentation, knowledge, intention and completeness by extending the attributes in the C3 dataset by the page textual content. We use Gradient Boosted Trees and recommender methods over the evaluator, site, evaluation triplets and their metadata and combine with text classifiers. In our experiments best results can be reached by the theoretically justified normalized SVM kernel. The normalization can be derived by using the Fisher information matrix of the text content. As the main contribution, we describe the theory of the Fisher matrix and show that SVM may be particularly suitable for difficult text classification tasks.

The temporally evolving models for dynamic networks is analysed in [57]. The research of complex networks and large graphs generated a wide variety of stochastic graph models that try to capture the properties of these complex systems. Most of the well-known models can describe a static graph extracted from a real-world dataset. They are capable of generating an ensemble of graphs, in which all graph instances are similar in terms of specific statistics to the original one. For example, models that capture the power-law degree distribution of real-world networks such as the Albert-Barabasi one are dynamic but do not attempt to model the actual temporal evolution of large graphs. Our goal is to give temporal stochastic graph model for the temporal dynamics of these complex systems. Our models address the link prediction problem introduced by Liben-Nowell and Kleinberg, in a temporal setting. More specifically, we try to predict accurately each new link in the graph at the time when it is created in the network. This experimental setting is similar to our method introduced for recommender systems. We explain this setup in case of dynamic graphs. For baseline algorithm, we apply online matrix factorization on temporal network data. Various node centrality measures capture the “importance” of a node by using the structural properties of the graph. While these metrics are widely investigated, few is known about the evolution of graph centrality in temporal graphs. In our work, we investigate the applicability of node centrality metrics in temporal graphs by examining their temporal behavior and computational complexity. We also use these metrics as side features in our matrix factorization models.

Temporal Twitter prediction by content and network are developed in [58]. In recent years Twitter became the social network for information sharing and spreading. By retweeting, users spreading information and build cascades of information pathways. In this paper we investigate the possibility of predicting the future popularity of emerging retweet cascades immediately after the message appears. We introduce a supervised machine learning approach which employs a rich feature set utilizing the textual content of the messages along with the retweet networks of the users. We also propose a temporal evaluation framework focusing on user level predictions in time.

In the paper [59] we model the properties of growing communities in social networks. Our main result is that small communities have higher edge density compared to random sub- graphs and their edge number follows power law in the number of nodes. In other words, the smaller the community, the larger the relative density. Our observation resembles the densification law of Leskovec, Kleinberg and Faloutsos who show that the average degree increases super-linearly as the size of the network grows. In our settings, however, densification is natural since the average degree of a random subgraph grows linearly. In contrary, sublinear growth translates to increased relative density in smaller subgraphs. Our experiments are carried over Twitter retweets and hashtags as well as a detailed music consumption log from Last.fm. In addition to the social network of Twitter followers and Last.fm friends, key in our experiments is that community subgraphs are defined by media use. We give theoretical results and simulations to explain our findings. The observed edge density can be explained by a mixture of epidemic growth that infects a uniform random neighbor of the community and a low probability selection of a completely new, isolated element. We also explore the relation of graph densification and subgraph sparsification by simulations over graphs of the Stanford Large Network Dataset Collection.

Publications

[55] P3.11 R.Palovics, A.A.Benczur, L.Kocsis, T.Kiss, E.Frigo, "Exploiting temporal influence in online recommendation", Proceedings of the 8th ACM Conference on Recommender systems (pp. 273-280), ACM (2015) [M14-WP5.3]

[56] P3.12 Balint Daroczy, David Siklosi, Robert Palovics, Andras A. Benczur, "Text Classification Kernels for Quality Prediction over the C3 Data Set", preprint, WebQuality 2015 in conjunction with WWW 2015 [M14-WP5.3]

[57] P3.13 Frederick Ayala, Robert Palovics, Andras A. Benczur, "Temporally Evolving Models for Dynamic Networks", accepted poster presentation at the International Conference on Computational Social Science, Helsinki, June 2015 [M14-WP5.3]

[58] P3.14 Balint Daroczy, Robert Palovics, Vilmos Wieszner, Richard Farkas, Andras A. Benczur, "Temporal Twitter prediction by content and network", preprint (2015) [M14-WP5.3]

[59] P3.15 Robert Palovics, Andras A. Benczur, "Modeling Community Growth: densifying graphs or sparsifying subgraphs? ", preprint (2015) [M14-WP5.3]

Deliverable D5.2: No deviations from the initial plan for this deliverable/milestones are reported.

3. Deliverables and milestones tables

Deliverables and milestones (for the 1st and 2th reporting period)

Del. no.	**Milestone/Deliverable name**	**WP no.**	**Nature**	**Disse mi-nation level**	**Delivery date**	**Delivered**
1	Correlation properties of directed networks	1	R	PU	18	yes
2	Statistical characterization of 2DRanking	1,2, 4	R	PU	18	yes
3	Eigenstate community detection	2,3	R	PU	18	yes
4	Spam filter protocols	4	R	PU	18	yes
5	Network specific centrality measures	1,3	R	PU	18	yes
6	Fractal Weyl law properties of networks	2	R	PU	36	yes
7	Protocols for large-scale network processing	4,5	R	PU	36	yes
8	Characterization of multiproduct world trade network	4	R	PU	36	yes
9	Webcrawler development and database collection	5	R	PU	36	yes
10	Monte Carlo algorithms for centrality measures	1	R	PU	36	yes
11	Delocalization conditions for Google matrix eigenstates	2	R	PU	36	yes
12	New protocols for social voting and recommendation	3	R	PU	36	yes
13	Characterization of ranking of Wiki-pedia and other networks	4	R	PU	36	yes
14	Characterization of time evolving Web structures	5	R	PU	36	yes

	Project website	1-5	O	PU	6	yes
	1st report	1-5	R	RE	18	yes
	2nd report	6-14	R	RE	36	yes
	Final report	1-14	R	RE	36	yes
	Final plan for using and disseminating knowledge	1-14	R	RE	36	yes

4. Project management

Project progress

The NADINE project activities as laid down in the WPs and in Section 1 above have progressed according to plan. All milestones of the second and whole project period have been reached.

Project web site is operating from the first day at www.quantware.ups-tlse.fr/FETNADINE

Jobs:

P1 hired post-doctoral fellow Young-Ho Eom (PhD at KAIST, S.Korea); hired period: 1 Oct 2012 - 26 November 2014 (26 months); PhD student Vivek Kandiah also participates in the project being supported by CNRS-Region-Midi-Pyreneesm he defended his thesis “Application of the Google matrix methods for characterization of directed networks” on 13 Oct 2014. Both left the group now.

P2 hired PhD student W.L.F. (Pim) van der Hoorn hired from 1 Oct 2012 for a period of 4 years (up to 3 years are covered by FET NADINE), he continues his PhD.

P3 hired post-doctoral fellows Zsolt Fekete, Csaba Sidlo, Istvan Petras and doctoral students Julianna Gobolos-Szabo, Robert Palovics, Andras Garzo supported in part from the NADINE project (in total of 24 months covered by FET NADINE)

P4 hired post-doctoral fellow Andrea Marino (PhD in Computer Science at Universita di Firenze); hired period: 1 March 2013 - 28 February 2015 (24 months)

All financial resources had been used according to the initial plan except Partner2 who had hired PhD student not from 1 May 2012 but from 1 Oct 2012 that gave a reduction of total NADINE budget for the fist 18 months less than 10 percent; also post-doc at Partner1 Y.-H.Eom left the group 4 months earlier, that was compensated by working months of premanent staff of P1. Planified and actual working person-months are reported in Annex person-month status table.

List of project meetings

In the second period the NADINE partners have met in various constellations at the following meetings, mainly to discuss physics, mathematics and computer science and to brainstorm on further research ideas within the NADINE framework:

- workshop Directed networks days was organized at Data Mining Research Group, Informatics Laboratory at the Computer and Automation Research Institute, Hungarian Academy of Sciences, Budapest, 8-11 May 2014; web page http://www.quantware.ups-tlse.fr/FETNADINE/dnd2014/

- Summer School at Ecole des Sciences Avancess de Luchon “Network analysis and applications”, June 21 – July 5, 2014 was organized by NADINE with 50 participants from 15 countries; web page http://www.quantware.ups-tlse.fr/ecoleluchon2014/

List of conference talks of NADINE partners: oral presentations of NADINE members are listed at www.quantware.ups-tlse.fr/FETNADINE/ (line Conferences); in total there are 65 (28 in second period) oral presentations on international conferences.

Publications in pdf format are available at the same NADINE web page (line publications); data sets and software are available at corresponding lines on the project web page.

5 publications crucial for the project during the second reporting period

[1] - [35]-P1.15 **Spectral properties of Google matrix of Wikipedia and other networks**

Authors: L.Ermann, K.M.Frahm and D.L.Shepelyansky
Journal: Eur. Phys. J. B v.86, p.193 (2013) (arXiv:1212.1068 [cs.IR], 2012)

Abstract: This review paper analyzes the properties of eigenvalues and eigenvectors of the Google matrix of the Wikipedia articles hyperlink network and other real networks. With the help of the Arnoldi method we analyze the distribution of eigenvalues in the complex plane and show that eigenstates with significant eigenvalue modulus are located on well defined network communities.

[2] - [40]-P1.20 **Delocalization transition for Google matrix eigenstates**

Authors: O.V.Zhirov and D.L. Shepelyansky
Journal: Ann. Der Physik (Berlin) DOI 10.1002/andp.201500110 (2015)

Abstract: We introduce a number of random matrix models describing the Google matrix G of directed networks. The properties of their spectra and eigenstates are analyzed by numerical matrix diagonalization. We show that for certain models it is possible to have an algebraic decay of PageRank vector with the exponent similar to real directed networks. At the same time the spectrum has no spectral gap and a broad distribution of eigenvalues in the complex plain. The eigenstates of G are characterized by the Anderson transition from localized to delocalized states and a mobility edge curve in the complex plane of eigenvalues.

[3] - [45]-P2.8 **Phase transitions for scaling of structural correlations in directed networks**

Authors: P. Van der Hoorn and N. Litvak,
Journal: preprint arXiv:1504.01535 [physics.soc-ph], 2015)

Abstract: Here the analysis of degree-degree dependencies in complex networks, and their impact on processes on networks requires null models, i.e. models that generate uncorrelated scale-free networks. Most models to date however show structural negative dependencies, caused by finite size effects. We analyze the behavior of these structural negative degree-degree dependencies, using rank based correlation measures, in the directed Erased Configuration Model. We obtain expressions for the scaling as a function of the exponents of the distributions. Moreover, we show that this scaling undergoes a phase transition, where one region exhibits scaling related to the natural cut-off of the network while another region has scaling similar to the structural cut-off for uncorrelated networks. By establishing the speed of convergence of these structural dependencies we are able to asses statistical significance of degree-degree dependencies on finite complex networks when compared to networks generated by the directed Erased Configuration Model.

[4] - [56]-P3.12 **Text Classification K ernels for Quality Prediction over the C3 Data Set**

Authors: Balint Daroczy, David Siklosi, Robert Palovics, Andras A. Benczur
Conference: preprint, WebQuality 2015 in conjunction with WWW 2015 (2015)

Abstract: We compare machine learning methods to predict quality aspects of the C3 dataset collected as a part of the Reconcile project. We give methods for automatically assessing the credibility, presentation, knowledge, intention and completeness by extending the attributes in the C3 dataset by the page textual content. We use Gradient Boosted Trees and recommender methods over the evaluator, site, evaluation triplets and their metadata and combine with text classifiers. In our experiments best results can be reached by the theoretically justified normalized SVM kernel. The normalization can be derived by using the Fisher information matrix of the text content. As the main contribution, we describe the theory of the Fisher matrix and show that SVM may be particularly suitable for difficult text classification tasks.

[5] – [71]-P4.18 **Liquid FM: Recommending Music through Viscous Democracy**

Authors: Paolo Boldi, Corrado Monti, Massimo Santini, and Sebastiano Vigna
Journal: submitted to CoRR (2015) (arXiv:1503.08604[cs.SI], 2015)

Abstract: Most modern recommendation systems use the approach of collaborative filtering: users that are believed to behave alike are used to produce recommendations. In this work we describe an application (Liquid FM) taking a completely different approach. Liquid FM is a music recommendation system that makes the user responsible for the recommended items. Suggestions are the result of a voting scheme, employing the idea of viscous democracy. Liquid FM can also be thought of as the first teststbed for this voting system. In this paper we outline the design and architecture of the application, both from the theoretical and from the implementation viewpoints.

The software has been developed and is now open for public via the web page
https://github.com/corradomonti/fbvoting

Dissemination: List and abstracts of papers and preprints appeared during the 2nd reporting period within the framework of NADINE (numbering continues from 1st period)

[33] **Google matrix analysis of C.elegans neural network**

Authors: V.Kandiah and D.L.Shepelyansky
Journal: Phys. Lett. A v.378, p.1932 (2014) (arXiv:1311.2013[physics.soc-ph])

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

[34] **Poisson statistics of PageRank probabilities of Twitter and Wikipedia networks**

Authors: K.M.Frahm and D.L.Shepelyansky
Journal: Eur. Phys. J. B v.87, p. 93 (2014) (arXiv:1402.5839[physics.soc-ph])

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

[35] **Google matrix analysis of directed networks**

Authors: L.Ermann, K.M.Frahm and D.L.Shepelyansky
Journal: submitted to Rev. Mod. Phys. (2014) (arXiv:1409.0428[physics.soc-ph])

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

[36] **Generalized friendship paradox in complex networks: the case of scientific collaboration**

Authors: Young-Ho Eom and Hong-Hyun Jo
Journal: Scientific Reports v.4, p.4603 (2014

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

[37] **Generalized friendship paradox in networks with tunable degree-attribute correlation**

Authors: Hong-Hyun Jo and Young-Ho Eom
Journal: Phys. Rev. E v.90, p.022809 (20124)

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

[38] **More ordering and communities in complex networks describing the game of go**

Authors: V.Kandiah, B.Georgeot and O.Giraud
Journal: Eur. Phys. J. B v.87, p.246 (20124)

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

[39] **Google matrix analysis of the multiproduct world trade network**

Authors: L.Ermann, and D.L.Shepelyansky
Journal: Eur.Phys. J. B v.88, p.84 (2015) (arXiv:1502.00584[cond-mat.dis-nn])

Abstract: Using the United Nations COMTRADE database [United Nations Commodity Trade Statistics Database, available at http://comtrade.un.org/db/ we construct the Google matrix G of multiproduct world trade between the UN countries and analyze the properties of trade flows on this network for years 1962-2010. This construction, based on Markov chains, treats all countries on equal democratic grounds independently of their richness and at the same time it considers the contributions of trade products proportionally to their trade volume. We consider the trade with 61 products for up to 227 countries. The obtained results show that the trade contribution of products is asymmetric: some of them are export oriented while others are import oriented even if the ranking by their trade volume is symmetric in respect to export and import after averaging over all world countries. The construction of the Google matrix allows to investigate the sensitivity of trade balance in respect to price variations of products, e.g. petroleum and gas, taking into account the world connectivity of trade links. The trade balance based on PageRank and CheiRank probabilities highlights the leading role of China and other BRICS countries in the worldtrade in recent years. We also show that the eigenstates of G with large eigenvalues select specific trade communities.

[40] **Anderson transition for Google matrix eigenstates**

Authors: O.V.Zhirov and D.L.Shepelyansky
Journal: Ann. der Physik (Berlin) DOI 10.1002/andp.201500110 (2015) (arXiv:1501.03371[q-fin.ST]

Abstract: We introduce a number of random matrix models describing the Google matrix G of directed networks. The properties of their spectra and eigenstates are analyzed by numerical matrix diagonalization. We show that for certain models it is possible to have an algebraic decay of PageRank vector with the exponent similar to real directed networks. At the same time the spectrum has no spectral gap and a broad distribution of eigenvalues in the complex plain. The eigenstates of G are characterized by the Anderson transition from localized to delocalized states and a mobility edge curve in the complex plane of eigenvalues.

[41] **Opinion formation driven by PageRank node influence on directed networks**

Authors: Y.-H.Eom and D.L.Shepelyansky
Journal: submitted to Physica A Feb (2015) (arXiv:1502.02567[physics.soc-ph])

Abstract: We study a two states opinion formation model driven by PageRank node influence and report an extensive numerical study on how PageRank affects collective opinion formations in large-scale empirical directed networks. In our model the opinion of a node can be updated by the sum of its neighbor nodes' opinions weighted by the node influence of the neighbor nodes at each step. We consider PageRank probability and its sublinear power as node influence measures and investigate evolution of opinion under various conditions. First, we observe that all networks reach steady state opinion after a certain relaxation time. This time scale is decreasing with the heterogeneity of node influence in the networks. Second, we find that our model shows consensus and non-consensus behavior in steady state depending on types of networks: Web graph, citation network of physics articles, and LiveJournal social network show non-consensus behavior while Wikipedia article network shows consensus behavior. Third, we find that a more heterogeneous influence distribution on the network has a more polarized opinion state for Web graph, Wikipedia, and Livejournal. However, the opposite behavior is observed in the citation network. Finally we identify that a small number of influential nodes can impose their own opinion on significant fraction of other nodes in all considered networks. Our study shows that the effects of heterogeneity of node influence on opinion formation can be significant and suggests further investigations on the interplay between node influence and collective opinion in networks.

[42] **Google matrix of the world network of economic activities**

Authors: V.Kandiah, H.Escaith and D.L. Shepelyansky
Journal: submitted to Eur. Phys. J. B April (2015) (arXiv:1504.XXXX[q-fin.ST])

Abstract: Using the new data from the OECD-WTO world network of economic activities we construct the Google matrix G of this directed network and perform its detailed analysis. The network contains 58 countries and 37 activity sectors for years 1995 and 2008. The construction of G, based on Markov chain transitions, treats all countries on equal democratic grounds while the contribution of activity sectors is proportional to their exchange monetary volume. The Google matrix analysis allows to obtain reliable ranking of countries and activity sectors and to determine the sensitivity of CheiRank-PageRank commercial balance of countries in respect to price variations and labor cost in various countries. We demonstrate that the developed approach takes into account multiplicity of network links with economy interactions between countries and activity sectors thus being more efficient compared to the usual export-import analysis. The spectrum and eigenstates of G are also analyzed being related to specific activity communities of countries.

[43] **Top 100 historical figures of Wikipedia**

Authors: D.L. Shepelyansky and other Wikipedia authors
Journal: Wikipedia article, 2015

Abstract: The top 100 historical figures of Wikipedia were determined by researchers from the University of Toulouse in France using mathematical and statistical methods from the Wikipedia database, and published in two scientific papers. In the statistical respects this top 100 list is of differs from the historical, cultural and other type arguments used by such historians like Michael H. Hart. The various mathematical methods and results obtained by different groups are described below. In spite or the mathematical and statistical grounds of those approaches they have overlap of about 43 percent with the top 100 list of Hart.

[44] **Convergence of rank based degree-degree correlations in random directed networks**

Authors: P2.8 P. van der Hoorn and N. Litvak
Journal: to appear in Moscow Journal of Combinatorics (2015)
(arXiv:1407.7662[math.PR], 2014)

Abstract: We introduce, and analyze, three measures for degree-degree dependencies, also called degree assortativity, in directed random graphs, based on Spearman's rho and Kendall's tau. We proof statistical consistency of these measures in general random graphs and show that the directed Configuration Model can serve as a null model for our degree-degree dependency measures. Based on these results we argue that the measures we introduce should be preferred over Pearson's correlation coefficients, when studying degree-degree dependencies, since the latter has several issues in the case of large networks with scale-free degree distribution.

[45] **Phase transitions for scaling of structural correlations in directed networks**

Authors: P. van der Hoorn and N. Litvak

Journal: arXiv:1504.01535[physics.soc-ph], 2015

Abstract: Analysis of degree-degree dependencies in complex networks, and their impact on processes on networks requires null models, i.e. models that generate uncorrelated scale-free networks. Most models to date however show structural negative dependencies, caused by finite size effects. We analyze the behavior of these structural negative degree-degree dependencies, using rank based correlation measures, in the directed Erased Configuration Model. We obtain expressions for the scaling as a function of the exponents of the distributions. Moreover, we show that this scaling undergoes a phase transition, where one region exhibits scaling related to the natural cut-off of the network while another region has scaling similar to the structural cut-off for uncorrelated networks. By establishing the speed of convergence of these structural dependencies we are able to asses statistical significance of degree-degree dependencies on finite complex networks when compared to networks generated by the directed Erased Configuration Model.

[46] **Modeling of trends in Twitter using retweet graph dynamics**

Authors: M. Ten Thij, T. Ouboter, D. Worm, N. Litvak, J.L. van den Berg and S. Bhulai
Journal: Proceedings 11th International Workshop Algorithms and Models for the Web

Graph, WAW 2014, 17-18 Dec 2014, Beijing, China. pp. 132-147; Lecture Notes in Computer Science 2014 (8882), Springer (2014), (arXiv:1502.00166[cs.SI], 2015)

Abstract: We model user behaviour in Twitter to capture the emergence of trending topics. For this purpose, we first extensively analyse tweet datasets of several different events. In particular, for these datasets, we construct and investigate the retweet graphs. We find that the retweet graph for a trending topic has a relatively dense largest connected component (LCC). Next, based on the insights obtained from the analyses of the datasets, we design a mathematical model that describes the evolution of a retweet graph by three main parameters. We then quantify, analytically and by simulation, the in influence of the model parameters on the basic characteristics of the retweet graph, such as the density of edges and the size and density of the LCC. Finally, we put the model in practice, estimate its parameters and compare the resulting behavior of the model to our datasets.

[47] **PageRank in scale-free random graphs**

Authors: N.Chen, N.Litvak and M.Olvera-Cravioto
Journal: Proceedings 11th International Workshop Algorithms and Models for the Web Graph, WAW 2014, 17-18 Dec 2014, Beijing, China pp. 120-131, Lecture Notes in Computer Science 2014 (8882), Springer (2014). (arXiv:1408.3610[math.PR], 2014)

Abstract: We analyze the distribution of PageRank on a directed configuration model and show that as the size of the graph grows to infinity it can be closely approximated by the PageRank of the root node on an appropriately constructed tree, This tree approximation is in turn related to the solution of a linear stochastic fixed point equation that has been thoroughly studied in the recent literature.

[48] **Ranking algorithms on directed configuration networks**

Authors: N.Chen, N.Litvak and M.Olvera-Cravioto
Journal: Submitted to Random Structures and Algorithms (2014) (arXiv:1409.7443v2[math.PR], 2014)

Abstract: This paper studies the distribution of a family of rankings, which includes Google's PageRank, on a directed configuration model. In particular, it is shown that the distribution of the rank of a randomly chosen node in the graph converges in distribution to a finite random variable R* that can be written as a linear combination of i.i.d. copies of the endogenous solution to a stochastic fixed point equation of a specific form. Moreover, we provide precise asymptotics for the limit R*, which when the in-degree distribution in the directed configuration model has a power law imply a power law distribution for R* with the same exponent.

[49] **Introduction to Special Issue on Searching and Mining the Web and Social Network**

Authors: N.Litvak and S.Vigna
Journal: Internet Mathematics, v.10(3-4), p.219-221 (2014)

Abstract: This issue of Internet Mathematics, titled `Searching and mining the Web and social networks', was born out of the interest of the editors for the problem of searching and analyzing not only the web, but also social networks in a broad sense. In particular, we aimed to publish a collection of papers that take a rigorous mathematical viewpoint on problems most important and common in network applications. The general topics represented in this special issue cover ranking of the nodes, network measurements, and adversarial behavior. Each of these topics received a large attention in the literature. We believe however that the originality of the papers presented in this volume is in a high level of mathematical rigor.

[50] **Quick detection of high-degree entities in large directed networks**

Authors: K.Avrachenkov, N.Litvak, L.Ostroumova-Prokhorenkova and E.Suyargulova
Journal/Conference: IEEE International Conference on Data Mining (ICDM 2014), 14-17 Dec 2014, Shenzhen, China. pp. 20-29. IEEE Computer Society (2014) (arXiv:1410.0571v2[cs.SI])

Abstract: we address the problem of quick detection of high-degree entities in large online social networks. Practical importance of this problem is attested by a large number of companies that continuously collect and update statistics about popular entities, usually using the degree of an entity as an approximation of its popularity. We suggest a simple, efficient, and easy to implement two-stage randomized algorithm that provides highly accurate solutions to this problem. For instance, our algorithm needs only one thousand API requests in order to find the top-100 most followed users, with more than 90 percent precision, in the online social network Twitter with approximately a billion of registered users. Our algorithm significantly outperforms existing methods and serves many different purposes such as finding the most popular users or the most popular interest groups in social networks. An important contribution of this work is the analysis of the proposed algorithm using 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 an accurate prediction for the algorithm's performance and show that the number of API requests for finding the top-most popular entities is sublinear in the number of entities. Moreover, we formally show that the high variability of the entities, expressed through heavy-tailed distributions, is the reason for the algorithm's efficiency. We quantify this phenomenon in a rigorous mathematical way

[51] **Distributed Frameworks for Alternating Least Squares (Poster presentation)**

Authors: Marton Balassi, Robert Palovics and Andras A. Benczur
Journal/Conference: Large-Scale Recommender Systems in conjunction with RecSys, Foster City, Silicon Valley, USA, 6th-10th October 2014

Abstract: At the RecSys 2014 Workshop on Large Scale Recommender Systems [51], we presented our idea on broadcasting identical node values to all graph neighbors to speed up distributed algorithms with communication patterns as simple as PageRank to as complex as Alternating Least Squares. SZTAKI PI will be co-organizer for this workshop in 2015.

[52] **Similarity Kernel Learning**

Authors: Balint Daroczy, Krisztian Buza, Andras A. Benczu

Conference: preprint (2015)

Abstract: Kernel methods are popular in machine learning tasks. For Support Vector Machine classification or Support Vector Regression, the central question is the selection of the appropriate kernel. The task is dicult in particular if the data points have complex or multimodal attributes such as time series or visual content enhanced with geographic, numeric or text metadata. Unlike earlier approaches of the so-called Multiple Kernel Learning problem, where a large number of kernels are fused by wrapper methods as part of the optimization process, in this paper we mathematically derive an optimal kernel for the data set in question. We begin with selecting appropriate distances for the appropriate modalities, for example dynamic time warping distance for time series and Jensen-Shannon distance for the bag of words text representation. Our kernel is defined, without needs of wrapper methods, by considering the distances as attributes generated by a Markov Random Field. For the Markov Random Field, the natural kernel is based on the Fisher information matrix and its exact form can be computed from the data. We experiment with the above similarity kernel over a wide variety of data sets, including: 64-channel EEG data; General time series data sets; Images with text annotations; Web documents; Gene expression level. Over the complex, multimodal or multiple time series classification tasks, our method outperforms the state of the art while reaching identical performance even over the simple unimodal problems as well, hence our method seems applicable under very general settings.

[53] **RecSys Challenge 2014: an ensemble of binary classifiers and matrix factorization**

Authors: R.Palovics, F. Ayala-Gomez, B. Csikota, B.Daroczy, L. Kocsis, D. Spadacene, A.A. Benczur

Journal/conference: Proceedings of the 2014 Recommender Systems Challenge (p. 13) ACM (2014)

Abstract: We give our solution to the RecSys Challenge 2014. In our ensemble we use (1) a mix of binary classication methods for predicting nonzero engagement, including logistic regression and SVM; (2) regression methods for directly predicting the engagement, including linear regression and gradient boosted trees; (3) matrix factorization and factorization machines over the user-movie matrix, by using user and movie features as side information. For most of the methods, we use the GraphLab Create implementation. Our current nDCG achieves 0.877.

[54] **Peer-to-peer Online Collaborative Filtering**

Authors: Andrea N. Ban, Levente Kocsis, Robert Palovics

Journal: preprint (2015)

Abstract: Recommender systems often deal with a large amount of sequential data. For these scenarios, online matrix factorization techniques based on online prediction and incremental updates are often the most promising approaches. Decentralizing the system and keeping the user data on their devices is an important step in the direction of preserving user privacy. In this paper we propose a peer-to-peer online matrix factorization algorithm that stores the ratings of a user and her private data local.

Additionally, the users have a local copy of the common part of the factor model and communicate with other users to advance towards a consensus on it. The algorithm is proven to converge to a set of local optima in the stationary case, while we show empirically that the algorithm performs well in the non-stationary case, both in terms of ranking performance and privacy preservation.

[55] **Exploiting temporal influence in online recommendation**

Authors: R.Palovics, A.A.Benczur, L.Kocsis, T.Kiss, E.Frigo

Conference: Proceedings of the 8th ACM Conference on Recommender systems (pp. 273-280), ACM (2015)

Abstract: We give methods for time-aware music recommendation in a social media service with the potential of exploiting immediate temporal influences between users. We consider events when a user listens to an artist the first time and this event follows some friend listening to the same artist short time before. We train a blend of matrix factorization methods that model the relation of the in influencer, the in influenced and the artist, both the individual factor decompositions and their weight learned by variants of stochastic gradient descent (SGD). Special care is taken since events of in influence form a subset of the positive implicit feedback data and hence we have to cope with two different definitions of the positive and negative implicit training data. In addition, in the time-aware setting we have to use online learning and evaluation methods. While SGD can easily be trained online, evaluation is cumbersome by traditional measures since we will have potentially different top recommendations at different times. Our experiments are carried over the two-year “scrobble" history of 70,000 Last.fm users and show a 5percent increase in recommendation quality by predicting temporal in influences.

[56] **Text Classification Kernels for Quality Prediction over the C3 Data Set**

Authors: Balint Daroczy, David Siklosi, Robert Palovics, Andras A. Benczur

Conference: preprint, WebQuality 2015 in conjunction with WWW 2015 (2015)

Abstract: We compare machine learning methods to predict quality aspects of the C3 dataset collected as a part of the Reconcile project. We give methods for automatically assessing the credibility, presentation, knowledge, intention and completeness by extending the attributes in the C3 dataset by the page textual content. We use Gradient Boosted Trees and recommender methods over the evaluator, site, evaluation triplets and their metadata and combine with text classifiers. In our experiments best results can be reached by the theoretically justified normalized SVM kernel. The normalization can be derived by using the Fisher information matrix of the text content. As the main contribution, we describe the theory of the Fisher matrix and show that SVM may be particularly suitable for difficult text classification tasks.

[57] **Temporally Evolving Models for Dynamic Networks**

Authors: Frederick Ayala, Robert Palovics, Andras A. Benczur

Conference: accepted poster presentation at the International Conference on Computational Social Science, Helsinki, June 2015

Abstract: The research of complex networks and large graphs generated a wide variety of stochastic graph models that try to capture the properties of these complex systems. Most of the well-known models can describe a static graph extracted from a real-world dataset. They are capable of generating an ensemble of graphs, in which all graph instances are similar in terms of specific statistics to the original one. For example, models that capture the power-law degree distribution of real-world networks such as the Albert- Barabasi one are dynamic but do not attempt to model the actual temporal evolution of large graphs. Our goal is to give temporal stochastic graph model for the temporal dynamics of these complex systems. Our models address the link prediction problem introduced by Liben-Nowell and Kleinberg, in a temporal setting. More specifically, we try to predict accurately each new link in the graph at the time when it is created in the network. This experimental setting is similar to our method introduced for recommender systems. We explain this setup in case of dynamic graphs. For baseline algorithm, we apply online matrix factorization on temporal network data. Various node centrality measures capture the “importance” of a node by using the structural properties of the graph. While these metrics are widely investigated, few is known about the evolution of graph centrality in temporal graphs. In our work, we investigate the applicability of node centrality metrics in temporal graphs by examining their temporal behavior and computational complexity. We also use these metrics as side features in our matrix factorization models.

[58] **Temporal Twitter prediction by content and network**

Authors: Balint Daroczy, Robert Palovics, Vilmos Wieszner, Richard Farkas, Andras A. Benczur

Conference: preprint (2015)

Abstract: In recent years Twitter became the social network for information sharing and spreading. By retweeting, users spreading information and build cascades of information pathways. In this paper we investigate the possibility of predicting the future popularity of emerging retweet cascades immediately after the message appears. We introduce a supervised machine learning approach which employs a rich feature set utilizing the textual content of the messages along with the retweet networks of the users. We also propose a temporal evaluation framework focusing on user level predictions in time.

[59] **Modeling Community Growth: densifying graphs or sparsifying subgraphs?**

Authors: Robert Palovics, Andras A. Benczur

Conference: preprint (2015)

Abstract: We model the properties of growing communities in social networks. Our main result is that small communities have higher edge density compared to random sub- graphs and their edge number follows power law in the number of nodes. In other words, the smaller the community, the larger the relative density. Our observation resembles the densification law of Leskovec, Kleinberg and Faloutsos who show that the average degree increases super-linearly as the size of the network grows. In our settings, however, densification is natural since the average degree of a random subgraph grows linearly. In contrary, sublinear growth translates to increased relative density in smaller subgraphs. Our experiments are carried over Twitter retweets and hashtags as well as a detailed music consumption log from Last.fm. In addition to the social network of Twitter followers and Last.fm friends, key in our experiments is that community subgraphs are defined by media use. We give theoretical results and simulations to explain our findings. The observed edge density can be explained by a mixture of epidemic growth that infects a uniform random neighbor of the community and a low probability selection of a completely new, isolated element. We also explore the relation of graph densification and subgraph sparsification by simulations over graphs of the Stanford Large Network Dataset Collection.

[60] **Statistical analysis of NOMAO customer votes for spots of France**

Authors: Robert Palovics, Balint Daroczym Andras A. Benczur, Julia Pap, Leonardo Ermann, Samuel Phan, Alexei D. Chepelianskii, Dima L. Shepelyansky,

Journal: submitted to Eur. Phys. J. B, preprint arXiv (2015)

Abstract: The recommendation systems are applied for analysis of the NOMAO data sets for voting of users (about 1 million) for spots (hotels, restaurants etc, about 20000 items) of Paris and France. We explore the spectrum and the eigenvalues of a matrix containing user ratings to geolocalized items. Eigenvalues nicely map to large towns and regions but show certain level of instability as we modify the interpretation of the underlying matrix. We evaluate imputation strategies that reach improved prediction performance by reaching geographically smooth eigenvectors.

[61] **Analysis of Twitter network**

Authors: Andras A. Benczur, Nelly Litwak et al.

Journal: preprint (2015)

Abstract: Recent network data of retweets on Twitter are analyzed by computer science tools.

[62] **Graph structure in the web - Revisited, or a trick of the heavy-tail**

Authors: Robert Meusel, Sebastiano Vigna, Oliver Lehmberg, and Christian Bizer

Journal: WWW'14 Companion, pp.427-432, International World Wide Web Conferences Steering Committee, 2014; a revised version is to appear in the Journal of Web Science (2015)

Abstract: We revisit the graph structure in the Web. Knowledge about the general graph structure of the World Wide Web is important for understanding the social mechanisms that govern its growth, for designing ranking methods, for devising better crawling algorithms, and for creating accurate models of its structure. In this paper, we describe and analyse a large, publicly accessible crawl of the web that was gathered by the Common Crawl Foundation in 2012 and that contains over3.5 billion web pages and 128.7 billion links. This crawl makes it possible to observe the evolution of the underlying structure of the World Wide Web within the last 10 years: we analyse and compare, among other features, degree distributions, connectivity, average distances, and the structure of weakly/strongly connected components. Our analysis shows that, as evidenced by previous research, some of the features previously observed by Broder et al. Are very dependent on artefacts of the crawling process, whereas other appear to be more structural. We confirm the existence of a giant strongly connected component; we however find, as observed by other researchers, very different proportions of nodes that can reach or that can be reached from the giant component, suggesting that the "bow-tie structure" as described previously is strongly dependent on the crawling process, and to the best of our current knowledge is not a structural property of the web. More importantly, statistical testing and visual inspection of size-rank plots show that the distributions of indegree, outdegree and sizes of strongly connected components are not power laws, contrarily to what was previously reported for much smaller crawls, although they might be heavy tailed. We also provide for the first time accurate measurement of distance-based features, using recently introduced algorithms that scale to the size of our crawl.

[63] **Cache-oblivious peeling of random hypergraphs**

Authors: Djamal Belazzougui, Paolo Boldi, Giuseppe Ottaviano, Rossano Venturini, and Sebastiano Vigna

Conference: 2014 Data Compression Conference (DCC 2014), IEEE pp.352-361. (2014)

Abstract: The computation of a peeling order in a randomly generated hypergraph is the most time-consuming step in a number of constructions, such as perfect hashing schemes, random r-SAT solvers, error-correcting codes, and approximate set encodings. While there exists a straightforward linear time algorithm, its poor I/O performance makes it impractical for hypergraphs whose size exceeds the available internal memory. We show how to reduce the computation of a peeling order to a small number of sequential scans and sorts, and analyze its I/O complexity in the cache-oblivious model. The resulting algorithm requires O(sort(n)) I/Os and O(n log n) time to peel a random hypergraph with n edges. We experimentally evaluate the performance of our implementation of this algorithm in a real-world scenario by using the construction of minimal perfect hash functions (MPHF) as our test case: our algorithm builds a MPHF of 7.6 billion keys in less than 21 hours on a single machine. The resulting data structure is both more space-efficient and faster than that obtained with the current state-of-the-art MPHF construction for large-scale key sets.

[64] **A network model characterized by a latent attribute structure with competition**

Authors: Paolo Boldi, Irene Crimaldi, and Corrado Monti

Confrerence: submitted CoRR (2014), (arXiv:1407.7729[cs.SI], 2014)

Abstract: we analyse a network model characterized by a latent attribute structure with competition. The quest for a model that is able to explain, describe, analyze and simulate real-world complex networks is of uttermost practical as well as theoretical interest. In this paper we introduce and study a network model that is based on a latent attribute structure: each node is characterized by a number of features and the probability of the existence of an edge between two nodes depends on the features they share. Features are chosen according to a process of Indian-Bu et type but with an additional random “fitness" parameter attached to each node, that determines its ability to transmit its own features to other nodes. As a consequence, a node's connectivity does not depend on its age alone, so also “young" nodes are able to compete and succeed in acquiring links. One of the advantages of our model for the latent bipartite “node-attribute" network is that it depends on few parameters with a straightforward interpretation. We provide some theoretical, as well experimental, results regarding the power-law behavior of the model and the estimation of the parameters. By experimental data, we also show how the proposed model for the attribute structure naturally captures most local and global properties (e.g., degree distributions, connectivity and distance distributions) real networks exhibit.

[65] **Entity-linking via graph-distance minimization**

Authors: Roi Blanco, Paolo Boldi, and Andrea Marino

Confrerence: Proceedings 3rd Workshop on GRAPH Inspection and Traversal Engineering, GRAPHITE 2014, Grenoble, France, 5th April 2014., pp.30-43 (2014

Abstract: Entity-linking is a natural-language–processing task that consists in identifying the entities mentioned in a piece of text, linking each to an appropriate item in some knowledge base; when the knowledge base is Wikipedia, the problem comes to be known as wikification in this case, items are wikipedia articles). One instance of entity-linking can be formalized as an optimization problem on the underlying concept graph, where the quantity to be optimized is the average distance between chosen items. Inspired by this application, we define a new graph problem which is a natural variant of the Maximum Capacity Representative Set. We prove that our problem is NP-hard for general graphs; nonetheless, under some restrictive assumptions, it turns out to be solvable in linear time. For the general case, we propose two heuristics: one tries to enforce the above assumptions and another one is based on the notion of hitting distance; we show experimentally how these approaches perform with respect to some baselines on a real-world dataset.

[66] **Supremum-norm convergence for step-asynchronous successive overrelaxation on M-matrices**

Authors: Sebastiano Vigna

Confrerence: submitted to CoRR (2014) (arXiv:1404.3327[cs.DS], 2014)

Abstract: step-asynchronous successive overrelaxation updates the values contained in a single vector using the usual Gau\ss-Seidel-like weighted rule, but arbitrarily mixing old and new values, the only constraint being temporal coherence: you cannot use a value before it has been computed. We show that given a nonnegative real matrix A, a $\sigma\geq\rho(A)$ and a vector $\boldsymbol{w}>0$ such that $A\boldsymbol{w}\leq\sigma\boldsymbol{w}$, every iteration of step-asynchronous successive overrelaxation for the problem $(sI-A)\boldsymbol{x}=\boldsymbol{b}$, with $s>\sigma$, reduces geometrically the $\boldsymbol{w}$-norm of the current error by a factor that we can compute explicitly. Then, we show that given a $\sigma>\rho(A)$ it is in principle always possible to compute such a $\boldsymbol{w}$. This property makes it possible to estimate the supremum norm of the absolute error at each iteration without any additional hypothesis on A, even when

A is so large that computing the product $A\boldsymbol{x}$ is feasible, but estimating the supremum norm of $(sI-A)-1$ is not.

[67] **An experimental exploration of Marsaglia's xorshift generators, scrambled**

Authors: Sebastiano Vigna

Confrerence: submitted to CoRR (2014) (arXiv:1404.3327[cs.DS], 2014)

Abstract: Marsaglia proposed recently xorshift generators as a class of very fast, good-quality pseudorandom number generators. Subsequent analysis by Panneton and L'Ecuyer has lowered the expectations raised by Marsaglia's paper, showing several weaknesses of such generators, verified experimentally using the TestU01 suite. Nonetheless, many of the weaknesses of xorshift generators fade away if their result is scrambled by a non-linear operation (as originally suggested by Marsaglia). In this paper we explore the space of possible generators obtained by multiplying the result of a xorshift generator by a suitable constant. We sample generators at 100 equispaced points of their state space and obtain detailed statistics that lead us to choices of parameters that improve on the current ones. We then explore for the first time the space of high-dimensional xorshift generators, following another suggestion in Marsaglia's paper, finding choices of parameters providing periods of length 21024−1 and 24096−1. The resulting generators are of extremely high quality, faster than current similar alternatives, and generate long-period sequences passing strong statistical tests using only eight logical operations, one addition and one multiplication by a constant.

[68] **Further scramblings of Marsaglia's xorshift generators**

Authors: Sebastiano Vigna

Confrerence: submitted to CoRR (2014) (arXiv:1403.0930[cs.NI], 2014)

Abstract: xorshift* generators are a variant of Marsaglia's xorshift generators that eliminate linear artifacts typical of generators based on $\mathbf{Z}/2\mathbf{Z}$-linear operations using multiplication by a suitable constant. Shortly after high-dimensional xorshift* generators were introduced, Saito and Matsumoto suggested a different way to eliminate linear artifacts based on addition in $\mathbf{Z}/232\mathbf{Z}$, leading to the XSadd generator. Starting from the observation that the lower bits of XSadd are very weak, as its reverse fails systematically several statistical tests, we explore xorshift+, a variant of XSadd using 64-bit operations, which leads, in small dimension, to extremely fast high-quality generators.

[69] **Interactions of cultures and top people of Wikipedia from ranking of 24 language editions**

Authors: Young Ho Eom, Pablo Aragon, David Laniado, Andreas Kaltenbrunner, Sebastiano Vigna, and Dima L. Shepelyansky,

Journal: PLoS ONE v.10(3), p.e0114825 (2015) (arXiv:1405.7183[cs.SI], 2014)

Abstract: We apply methods of Markov chainsand Google matrix for the analysis of the hyperlink networks of 24 Wikipedia language editions, and rank all their articles by PageRank, 2DRank and CheiRank algorithms. Using automatic extraction of people names, we obtain the top 100 historical figures, for each edition and for each algorithm. We investigate their spatial, temporal, and gender distributions in dependence of their cultural origins. Our study demonstrates not only the existence of skewness with local figures, mainly recognized only in their own cultures, but also the existence of global historical figures appearing in a large number of editions. By determining the birth time and place of these persons, we perform an analysis of the evolution of such figures through 35 centuries of human history for each language, thus recovering interactions and entanglement of cultures over time. We also obtain the distributions of historical figures over world countries, highlighting geographical aspects of cross-cultural links. Considering historical figures who appear in multiple editions as interactions between cultures, we construct a network of cultures and identify the most influential cultures according to this network.

[70] **A weighted correlation index for rankings with ties**

Authors: Sebastiano Vigna

Journal/Conference: Proceedings of the 24th international conference on World Wide Web, ACM (2015) (arXiv:1404.3325[cs.SI], 2014)

Abstract: Understanding the correlation between two different scores for the same set of items is a common problem in information retrieval, and the most commonly used statistics that quantifies this correlation is Kendall's τ. However, the standard definition fails to capture that discordances between items with high rank are more important than those between items with low rank. Recently, a new measure of correlation based on average precision has been proposed to solve this problem, but like many alternative proposals in the literature it assumes that there are no ties in the scores. This is a major deficiency in a number of contexts, and in particular while comparing centrality scores on large graphs, as the obvious baseline, indegree, has a very large number of ties in web and social graphs. We propose to extend Kendall's definition in a natural way to take into account weights in the presence of ties. We prove a number of interesting mathematical properties of our generalization and describe an $O(n\log n)$ algorithm for its computation. We also validate the usefulness of our weighted measure of correlation using experimental data.

[71] **Liquid FM: Recommending Music through Viscous Democracy**

Authors: Paolo Boldi, Corrado Monti, Massimo Santini, and Sebastiano Vigna

Journal/Conference: submitted to CoRR (2015) (arXiv:1503.08604[cs.SI], 2015)

Abstract: Most modern recommendation systems use the approach of collaborative filtering: users that are believed to behave alike are used to produce recommendations. In this work we describe an application (Liquid FM) taking a completely different approach. Liquid FM is a music recommendation system that makes the user responsible for the recommended items. Suggestions are the result of a voting scheme, employing the idea of viscous democracy. Liquid FM can also be thought of as the first testbed for this voting system. In this paper we outline the design and architecture of the application, both from the theoretical and from the implementation viewpoints.

[72] **LlamaFur: Learning Latent Category Matrix to Find Unexpected Relations in Wikipedia**

Authors: Paolo Boldi and Corrado Monti

Journal/Conference: preprint submitted to CoRR (2015)

Abstract: We consider the following problem. Besides finding trends and unveiling typical patterns, modern information retrieval is increasingly more interested in the discovery of surprising information in textual datasets. In this work we focus on finding unexpected links in hyper- linked document corpora when documents are assigned to categories; our approach is based on the determination of a latent category matrix that explains common links; the matrix is built using a perceptron-like technique. We show that our method provides better accuracy than most existing text-based techniques, with higher efficiency and relying on a much smaller amount of information. It also provides higher precision than standard link prediction, especially at low recall levels; the two methods are in fact shown to be orthogonal and can therefore be fruitfully combined.

[73] **Local Ranking Problem on the BrowseGraph**

Authors: Michele Trevisio, Luca Maria Aiello, Paolo Boldi and Roi Blanco

Journal/Conference: accepted for publication in SIGIR (2015

Abstract: The Local Ranking Problem" (LRP) is related to the computation of a centrality-like rank on a local graph, where the scores of the nodes could significantly differ from the ones computed on the global graph. Previous work has studied LRP on the hyperlink graph but never on the BrowseGraph, namely a graph where nodes are webpages and edges are browsing transitions. Recently, this graph has received more and more attention in many different tasks such as ranking, prediction and recommendation. However, a webserver has only the browsing traffic performed on its pages (local BrowseGraph) and, as a consequence, the local computation can lead to estimation errors, which hinders the increasing number of applications in the state of the art. Also, although the divergence between the local and global ranks has been measured, the possibility of estimating such divergence using only local knowledge has been mainly over-looked. These aspects are of great interest for online service providers who want to gauge their ability to correctly assess the importance of their resources only based on their local knowledge, and by taking into account real user browsing fluxes that better capture the actual user interest than the static hyperlink network. We study the LRP problem on a BrowseGraph from a large news provider, considering as subgraphs the aggregations of browsing traces of users coming from different domains. We show that the distance between rankings can be accurately predicted based only on structural information of the local graph, being able to achieve an average rank correlation as high as 0.8.

5. Explanation of the use of the resources

Please provide an explanation of personnel costs, subcontracting and any major direct costs incurred by each beneficiary, such as the purchase of important equipment, travel costs, large consumable items, etc. linking them to work packages.

There is no standard definition of "major direct cost items". Beneficiaries may specify these, according to the relative importance of the item compared to the total budget of the beneficiary, or as regards the individual value of the item.

(The rest of this template will not be part of the report but be submitted independently via the online application NEF)

Financial statements – Form C and Summary financial report

Remark: This section will not be part of the scientific reporting and should be filled in via the online applicaiont NEF (Simply refer in the scientific report to the online application)

Please submit a separate financial statement from each beneficiary (if Special Clause 10 applies to your Grant Agreement, please include a separate financial statement from each third party as well) together with a summary financial report which consolidates the claimed Community contribution of all the beneficiaries in an aggregate form, based on the information provided in Form C (Annex VI) by each beneficiary.

When applicable, certificates on financial statements shall be submitted by the concerned beneficiaries according to Article II.4.4 of the Grant Agreement.

IMPORTANT:

Form C varies with the funding scheme used. Please make sure that you use the correct form corresponding to your project. Templates for Form C are provided in Annex VI of the Grant Agreement. An example for collaborative projects is enclosed hereafter. A Web-based online tool for completing and submitting the forms C is under preparation. If you have to submit forms C before the tool becomes available, please ask your Commission project officer for an Excel version of the form.

If some beneficiaries in security research have two different rates of funding (part of the funding may reach 75% in reference with Article 33.1 of the EC rules for participation - REGULATION (EC) No 1906/2006) then two separate financial statements should be filled by the concerned beneficiaries and two lines should be entered for these beneficiaries in the summary financial report.

CERTIFICATES

Remark: This section will not be part of the scientific reporting.

A copy of each duly signed certificate (depending on whether Expenditure threshold is reached such a certificate will be necessary or not).on the financial statements (Form C) or on the methodology should be included in this section, according to the table above (signed originals to be sent in parallel by post).

Audit certificates that should be send in one package.

Person-Month Status Table

Work package[1]	**WP1**		**WP2**		**WP3**		**WP4**		**WP5**		**WP6**	**WP7**	**TOTAL per Beneficiary**	
	Actual WP total	Planned WP total	Actual WP total	Planned WP total	Actual WP total	Planned WP total	Actual WP total	Planned WP total	Actual WP total	Planned WP total	Actual/ Planned total	Actual/ Planned total	Actual total	Planned total
Coordinator P1 CNRS	9	9	17	17	9	9	3	3	5	5	3/ 3	2 / 2	48	48
Beneficiary P2 UTWE	16	16	9	9	5	5	6	6	4	4	0 / 0	2 / 2	42	42
Beneficiary P3 MTA_SZTAKI	5	5	4	4	4	4	16	16	11	11	0 / 0	2/ 2	42	42
Beneficiary P4 UMIL	5	5	4	4	13	13	6	6	14	14	0 / 0	2 / 2	44	44

[1] Please indicate in the table the number of person months over the whole duration for the planned work, for each workpackage by each beneficiary