PROJECT PERIODIC REPORT

Grant Agreement number: 288956

Project acronym: NADINE

Project title: New tools and Algorithms for Directed Network analysis

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

Periodic report: 1st X 2nd □

Period covered: from 1.5.2012 to 31.10.2013

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

Dr. Dima Shepelyansky

Directeur de recherche au CNRS

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

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

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

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

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

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, 32 papers and preprints have now appeared within the framework of NADINE, including 2 papers in PLoS ONE, one paper of P4 had been highlighted by New York Times. Results have been reported on 37 international conferences.

Highlights in the first reporting period include work re on spectral properties of spectrum and eigenstates of Wikipedia and their links with communities detection [1], multilingual ranking of world persons with global heroes like Napoleon and Michael Jackson [2], design of new correlation measures between degrees of neighboring nodes [3], detailed analysis of Last.fm network of users [4], demonstration of four degrees of separation for the whole Facebook network of users [5].

[1] L.Ermann, K.M.Frahm and D.L.Shepelyansky, "Spectral properties of Google matrix of Wikipedia and other networks", Eur. Phys. J. B v.86, p.193 (2013), arXiv:1212.1068 [cs.IR]

[2] Y.-H.Eom and D.L.Shepelyansky, "Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles", to appear in PLoS ONE Oct 2013, arXiv:1306.6259 [cs.SI]

[3] N. Litvak, and R. van der Hofstad “Uncovering disassortativity in large scale-free networks”, Phys. Rev. E v.87, p.022801 (2013) (arXiv:1204.0266[physics.soc-ph])

[4] R. Pálovics, A.A. Benczúr. Temporal influence over the Last.fm social network. The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM 2013 Niagara Falls, Canada, August 25-28, 2013

[5] SebaL.Backstrom, P.Boldi, M.Rosa, J.Ugander, and S.Vigna. "Four degrees of separation", ACM Web Science 2012: Conference Proceedings, June 2012, pages 45-54, ACM Press (2012); best paper award, highlighted by the New York Times; (arXiv:1111.4570, 2012)

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 first period including: correlation properties of directed networks, statistical characterization of 2DRanking, eigenstate community detection, spam filter protocols, network-specific centrality measures. In parallel an extensive research is developed in the scientific directions of other milestones.

All scientific objectives of the first 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)

Milestone M1. Correlation properties of directed networks (WP1.1).
(Reporting period: M12-18)

Correlation between PageRank and CheiRank vectors is analyzed for a large variety of directed networks including Linux Kernel network, Wikipedia networks, UK university networks, gene transcription networks, Twitter network 2009, brain models. It is shown that some networks have correlator close to zero (e.g. Linux case) while in other cases it can be rather larger (e.g. Twitter, Wikipedia). Statistical distributions in PageRank-CheiRank plane are determined.

Degree-degree dependencies of neighboring nodes have been analyzed. It has been shown that the commonly used Pearson's correlation coefficient has important flaws, in particular, it converges to zero with network size under realistic assumptions on the network structure. New measures, based of rank correlations, have been proposed and analyzed on Web data, Wikipedia data, scientific citation graphs, and analytically in several random graph models, including the configuration model and the Preferential Attachment model.

Deliverables are reported in [1-4],[5-7]

Publications M1:

[1] L.Ermann, A.D.Chepelianskii and, D.L.Shepelyansky "Towards two-dimensional search engines", J. Phys. A: Math. Theor. v.45, p.275101 (2012) (arXiv:1106.6215[cs.IR])
[2] K.M.Frahm and D.L. Shepelyansky "Google matrix of Twitter", Eur. Phys. J. B v.85, p.355 (2012) (arXiv:1207.3414[cs.SI], 2012)
[3] Y.-H.Eom, K.M.Frahm, A.Benczur and D.L. Shepelyansky, "Time evolution of Wikipedia network ranking", submitted Eur. Phys. J. B (2013) (arXiv:1304.6601 [physics.soc-ph], 2013)
[4] Y.-H.Eom and D.L. Shepelyansky, "Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles", PLoS ONE v.8(10),p.e74554(2013) (arXiv:1306.6259 [cs.SI], 2013)
[5] N. Litvak, and R. van der Hofstad "Uncovering disassortativity in large scale-free networks", Phys. Rev. E v.87, p.022801 (2013) (arXiv:1204.0266[physics.soc-ph])
[6] N. Litvak, and R. van der Hofstad, "Degree-degree correlations in random graphs with heavy-tailed degrees." Accepted in Internet Mathematics. (arXiv:1202.3071 [math.PR]) (2013)
[7] P. van der Hoorn and N. Litvak, "Degree-degree correlations in directed networks with heavy-tailed degrees." Manuscript submitted in October arXiv:1310.6528[math.PR] (2013.)

Milestone M5: Network specific centrality measures *(Reporting period: M18)*

We have measured the differences between networks using our newly developed rank correlation measures in Wikipedia graphs for nine different languages, and compared the results to directed configuration models with same degree sequences and randomized connections. The results are reported in [1]. We have proposed a new centrality measure – alpha-current flow betweenness centrality. The original shortest-path betweenness centrality is based on counting shortest paths which go through a node or an edge, and it ignores the important paths that are just one or two hops longer than the shortest paths. Our new measure rectifies this shortcoming, plus it has an important advantage that it can be computed efficiently on large graphs, while other betweennees centrality measures are known to have a prohibitive computational complexity. The results are reported in [2]. We have made a significant progress in analyzing a newly posed mathematical problem: the analysis of a PageRank and CheiRank distribution in random directed graphs. We have obtained the coupling of random graphs with trees and derived the behavior of PageRank and CheiRank. The results have been reported at the INFORMS APS 2013 conference. The preprint is expected at the beginning of 2014.

Deliverables are reported in [1,2].

Publications M5:

[1] P. van der Hoorn and N. Litvak, "Degree-degree correlations in directed networks with heavy-tailed degrees." Manuscript submitted in October 2013.
[2] K. Avrachenkov, N. Litvak, V. Medyanikov, and M. Sokol, "Alpha current flow betweenness centrality", Accepted In:10th Workshop on Algorithms and Models for the Web Graph, WAW2013, 15-16 December, 2013, Harvard University (arXiv:1308.2591v1 [cs.SI], 2013)

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

Milestone M2: Statistical characterization of 2DRanking (WP1.2; WP2.1; WP4.3) *(Reporting period: M12-18)*

We analyze [1] the statistical properties of various directed networks using ranking of their nodes based on the dominant vectors of the Google matrix known as PageRank and CheiRank. On average PageRank orders nodes proportionally to a number of ingoing links, while CheiRank orders nodes proportionally to a number of outgoing links. In this way the ranking of nodes becomes two-dimensional that paves the way for development of two-dimensional search engines of new type. Statistical properties of information flow on PageRank-CheiRank plane are analyzed for networks of British, French and Italian Universities, Wikipedia, Linux Kernel, gene regulation and other networks. A special emphasis is laid on British Universities networks using the large database publicly available at UK. Methods of spam links control are also analyzed. This 2DRanking analysis is extended to Twitter network [2], Wikipedia networks at different years 2003-2012 [3] and Wikipedia editions in 9 languages [4]. Deliverables are published in [1-3].

Publications M2:

[1] L.Ermann, A.D.Chepelianskiiand, D.L.Shepelyansky “Towards two-dimensional search engines”, J. Phys. A: Math. Theor. v.45, p.275101 (2012) (arXiv:1106.6215[cs.IR])
[2] K.M.Frahm and D.L. Shepelyansky "Google matrix of Twitter", Eur. Phys. J. B v.85, p.355 (2012) (arXiv:1207.3414[cs.SI], 2012)
[3] Y.-H.Eom, K.M.Frahm, A.Benczur and D.L. Shepelyansky, "Time evolution of Wikipedia network ranking", submitted Eur. Phys. J. B (2013) (arXiv:1304.6601 [physics.soc-ph], 2013)
[4] Y.-H.Eom and D.L. Shepelyansky, "Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles", PLoS ONE v.8910), p.e74554 (2013) (arXiv:1306.6259 [cs.SI], 2013)

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

Milestone M5: Network specific centrality measures *(Reporting period: M18)*

We developed a new class of centrality measures – alpha-current flow betweenness centrality, that solve the long-standing problem of the high computational complexity of betweenness centrality measures, in particular, the current flow betweenness centrality introduced by Newman. We showed that our new measures identify the nodes that keep the network connected [1]. Compared to the known betweenness measures, our new measures can be computed efficiently on large graphs, while the computations of other betweenness centrality measures are merely infeasible. The results are reported in [1].

We have developed and analyzed a new randomized ranking algorithm that finds most popular entities in large directed graphs (such as the most followed users in Twitter or most popular interest groups). We obtained accurate results in a very small number of steps, for example, starting with zero information on the Twitter graph, we need only 1000 requests to find top-100 most followed Twitter users with more than 90% accuracy. We have applied this algorithm in Twitter graph, Web graphs, and the Russian social network VKontakte (more than 200M users).

We prove that the highly skewed in-degree distribution is crucial for the performance of the algorithm. The results are reported in [2, 3].

Deliverables are reported in [1-3]..

Publications M5:

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

[2] K. Avrachenkov, N. Litvak, M. Sokol, and D.Towsley, “Quick detection of nodes with large degrees.” 9th International Workshop on Algorithms and Models for the Web Graph, WAW 2012, 22-23 June 2012, Halifax, NS, Canada. pp. 54-65. Lecture Notes in Computer Science 7323. Springer Verlag. (arXiv:1202.3261v1 [cs.DS], 2012)

[3] L. Ostroumova, K. Avrachenkov and N. Litvak. “Quick detection of popular entities in large directed networks.” Submitted before Nov 2013. The paper has not been not published on arXiv for the purpose of the blind review.

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 [1] and random sampling [2] 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 [3]. This work will be continued and extended to other centrality measures and correlation measures in large directed networks.

In [4], we approach the problem of computing geometric centralities, such as closeness and harmonic centrality, on very large graphs; traditionally this task requires an all- pairs shortest-path computation in the exact case, or a number of breadth-first traversals for approximated computations, but these techniques yield very weak statistical guarantees on highly disconnected graphs. We rather assume that the graph is accessed in a semi-streaming fashion, that is, that adjacency lists are scanned almost sequentially, and that a very small amount of memory (in the order of a dozen bytes) per node is available in core memory. We leverage the newly discovered algorithms based on probabilistic HyperLogLog counters, making it possible to approximate a number of geometric centralities at a very high speed and with high accuracy. While the application of similar algorithms for the approximation of closeness was attempted in the MapReduce framework, our exploitation of HyperLogLog counters reduces exponentially the memory footprint, paving the way for in-core processing of networks with a hundred billion nodes using “just" 2TiB of RAM. Moreover, the computations we describe are inherently parallelizable, and scale linearly with the number of available cores.

Publications M10:

[1]K. Avrachenkov, N. Litvak, M. Sokol, and D.Towsley, “Quick detection of nodes with large degrees.” 9th International Workshop on Algorithms and Models for the Web Graph, WAW 2012, 22-23 June 2012, Halifax, NS, Canada. pp. 54-65. Lecture Notes in Computer Science 7323. Springer Verlag. (arXiv:1202.3261v1 [cs.DS], 2012)

[2] L. Ostroumova, K. Avrachenkov and N. Litvak. “Quick detection of popular entities in large directed networks.” Submitted. The paper has not been not published on arXiv for the purpose of the blind review.
[3] K. Avrachenkov, N. Litvak, V. Medyanikov, and M. Sokol, “Alpha current flow betweenness centrality”, Accepted In:10th Workshop on Algorithms and Models for the Web Graph, WAW2013, 15-16 December, 2013, Harvard University (arXiv:1308.2591v1 [cs.SI], 2013)
[4] P. Boldi, S. Vigna. "In-core computation of geometric centralities with HyperBall: A hundred billion nodes and beyond", To appear in the Proceedings of 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW 2013). (arXiv:1308.2144, 2013)

Deliverable D1.1: 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)

Milestone M 2: Statistical characterization of 2DRanking (WP1.2; WP2.1; WP4.3) *(Reporting period: M12-18)*

For Hermitian and unitary matrices the Random Matrix Theory invented by Wigner and Dyson captures the universal properties of many real physical systems including complex nuclei, atoms and molecules, chaotic billiards, systems of quantum chaos. It would be very useful to find such universal models for Markov chains. However, in our studies we established that the well known models of complex networks (e.g. Albert-Barabasi model, randomized link model) have a large spectral gap and do not correctly reproduce the spectrum of Google matrix of real networks [1,2].

Thus for UK universities there is a large fraction (20 percent) of states with degenerate unit eigenvalue at unit damping factor [3]. We developed a theory for network of integers with nilpotent Google matrix [4] and citation network of Phys Rev [5]. Certain similarities are established with small size N=3,4 random ortostochastic matrices [3,6,7]. For Google matrix of DNA sequences we find new algebraic statistical distribution of matrix elements, however a random matrix with this fixed distribution of matrix elements does not reproduce the PageRank slow decay with index [8]. Deliverables are published in [4-9].

2DRanking of the citation network of Physical Review for the whole period 1893 – 2009 is analyzed in detail in [5]. We also discuss the properties of random matrix models of Perron-Frobenius operators [5]. Deliverables are published in [1-8].

Publications

[1] O.Giraud, B.Georgeot and D.L.Shepelyansky, "Delocalization transition for the Google matrix", Phys. Rev. E v.80, p.026107 (2009)

[2] B.Georgeot, O.Giraud and D.L.Shepelyansky, "Spectral properties of the Google matrix of the World Wide Web and other directed networks", Phys. Rev. E v.81, p.056109 (2010)

[3] K.M.Frahm, B.Georgeot and D.L.Shepelyansky, "Universal emergence of PageRank", J. Phys, A: Math. Theor. v.44, p.465101 (2011)

[4] K.M.Frahm, A.D.Chepelianskii and D.L.Shepelyansky, "PageRank of integers", J. Phys. A: Math. Theor. v.45, p.405101 (2012), arXiv:1205.6343[cs.IR]

[5] K.M.Frahm, Y.-H.Eom, D.L.Shepelyansky, :"Google matrix of the citation network of Physical Review", submitted Phys. Rev. E, arXiv:1310.5624[physics.soc-ph] (2013)

[6] L.Ermann, K.M.Frahm and D.L.Shepelyansky, "Spectral properties of Google matrix of Wikipedia and other networks", Eur. Phys. J. B v.86, p.193 (2013), arXiv:1212.1068 [cs.IR]

[7] Y.-H.Eom, K.M.Frahm, A.Benczur and D.L.Shepelyansky, "Time evolution of Wikipedia network ranking", submitted EPJB 24 April 2013, arXiv:1304.6601 [physics.soc-ph]

[8] V.Kandiah and D.L.Shepelyansky, "Google matrix analysis of DNA sequences", PLOS One v.8(5), p. e61519 (2013), (arXiv:1301.1626[q-bio.GN]

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

Milestone M3: Eigenstate community detection (WP2.2; WP3.1)

(Reporting period: M12-18)

The general properties of Google matrix spectrum of real networks (UK universities, Wikipedia, Twitter etc) are established in [1,2,3,4]. The spectrum has strongly degenerate unit eigenvalue (at unit damping factor). This is related to isolated subspaces, the core component may have eigenvalues being exponentially close to unit eigenvalue. The states other than PageRank are composed from relatively small number of nodes [2]. Possibilities of Pagerank delocalization, similar to the Anderson metal-insulator transition in disordered solids is discussed [3]. Data for

Twitter and DNA matrices indicate on delocalization tendency with the increase of number of links per node [1,2,3,5]. It is shown that eigenstates with a relatively significant eigenvalue modulus correspond to well defined communities [2].

In [6] we study the statistical properties of spectrum and eigenstates of he Google matrix of the citation network of Physical Review for the period 1893 - 2009. Practically the whole range of complex eigenvalues is determined numerically using the computational precision with up to 16384 binary digits, that allows to resolve hard numerical problems for small eigenvalues. The nearly nilpotent matrix structure allows to obtain semi-analytical computation of eigenvalues. We find that the spectrum is characterized by the fractal Weyl law with a fractal dimension approximately 1. It is found that the majority of eigenvectors are located in a localized phase. The statistical distribution of articles in the PageRank-CheiRank plane is established providing a better understanding of information flows on the network. The concept of ImpactRank is proposed to determine an influence domain of a given article.

Deliverables are published in [1-6].

Publications M3:

[1] K.M.Frahm, B.Georgeot and D.L.Shepelyansky, "Universal emergence of PageRank", J. Phys, A: Math. Theor. v.44, p.465101 (2011)

[2] L.Ermann, K.M.Frahm and D.L.Shepelyansky, "Spectral properties of Google matrix of Wikipedia and other networks", Eur. Phys. J. B v.86, p.193 (2013), arXiv:1212.1068 [cs.IR]

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

[4] Y.-H.Eom, K.M.Frahm, A.Benczur and D.L.Shepelyansky, "Time evolution of Wikipedia network ranking", submitted EPJB 24 April 2013, arXiv:1304.6601 [physics.soc-ph]

[5] V.Kandiah and D.L.Shepelyansky, "Google matrix analysis of DNA sequences", PLOS One v.8(5), p. e61519 (2013), arXiv:1301.1626[q-bio.GN]

[6] K.M.Frahm, Y.-H.Eom, D.L.Shepelyansky, “Google matrix of the citation network of Physical Review”, arXiv:1310.5624 (2013)

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

Milestone M 6: Fractal Weyl law properties of networks (WP2.3) *(Reporting period: M24-36)*

The work is in progress. The fractal Weyl law had been established for Linux Kernel network [1] and Ulam networks of dissipative dynamical maps [2,3]. Investigations of validity of this property for other networks are in progress. The fractal Weyl law is shown to be valid for the citation network of Physical Review [4].

Publications

[1] L.Ermann, A.D.Chepelianskii and D.L.Shepelyansky, "Fractal Weyl law for Linux Kernel Architecture", Eur. Phys. J. B v.79, p.115-120 (2011)

[2] D.L.Shepelyansky and O.V.Zhirov, "Google matrix, dynamical attractors and Ulam networks", Phys. Rev. E v.81, p.036213 (2010)

[3] L.Ermann and D.L.Shepelyansky, "Ulam method and fractal Weyl law for Perron-Frobenius operators", Eur. Phys. J. B v.75, p.299-304 (2010)

[4] K.M.Frahm, Y.-H.Eom, D.L.Shepelyansky, “Google matrix of the citation network of Physical Review”, arXiv:1310.5624[physics.soc-ph] (2013)

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

Milestone M 11: Delocalization conditions for Google matrix eigenstates (WP2.4) *(Reporting period: M36)*

Work is in progress. The results presented in [1,2,3] indicate on possible delocalization of PageRank when the number of links per node becomes suffuciently large. In [3] we extend the Google matrix analysis to DNA sequences. For DNA sequences of various species we construct the Google matrix G of Markov transitions between nearby words composed of several letters. The statistical distribution of matrix elements of this matrix is shown to be described by a power law with the exponent being close to those of outgoing links in such scale-free networks as the World Wide Web (WWW). At the same time the sum of ingoing matrix elements is characterized by the exponent being significantly larger than those typical for WWW networks. This results in a slow algebraic decay of the PageRank probability determined by the distribution of ingoing elements. The spectrum of G is characterized by a large gap leading to a rapid relaxation process on the DNA sequence networks. We introduce the PageRank proximity correlator between different species which determines their statistical similarity from the view point of Markov chains. The properties of other eigenstates of the Google matrix are also discussed. Our results establish scale-free features of DNA sequence networks showing their similarities and distinctions with the WWW and linguistic networks. We find that the decay exponent of PageRank of DNA sequencies is by factor 4 smaller than its typical value for WWW.

In [4] we study the Ulam networks in symplectic dynamical maps with absorption. We study numerically the statistics of Poincare recurrences for the Chirikov standard map and the separatrix map at parameters with a critical golden invariant curve. The properties of recurrences are analyzed with the help of a generalized Ulam method. This method allows to construct the corresponding Ulam matrix whose spectrum and eigenstates are analyzed by the powerful Arnoldi method. We also develop a new survival Monte Carlo method which allows us to study recurrences on times changing by ten orders of magnitude. We show that the recurrences at long times are determined by trajectory sticking in a vicinity of the critical golden curve and secondary resonance structures. The values of Poincare exponents of recurrences are determined for the two maps studied. We also discuss the localization properties of eigenstates of the Ulam matrix and their relation with the Poincare recurrences. This study allows to understand the properties of Markov chains in a better way.

Publications

[1] D.L.Shepelyansky and O.V.Zhirov, "Google matrix, dynamical attractors and Ulam networks", Phys. Rev. E v.81, p.036213 (2010)

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

[3] V.Kandiah and D.L.Shepelyansky, "Google matrix analysis of DNA sequences", PLOS One v.8(5), p. e61519 (2013), (arXiv:1301.1626[q-bio.GN]

[4] K.M.Frahm and D.L.Shepelyansky, "Poincare recurrences and Ulam method for the Chirikov standard map ", Eur. Phys. J. B v.86, p.322 (2013), arXiv:1302.2761 [nlin.CD]

Deliverable D2.1: 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)

Milestone M3: Eigenstate community detection (WP2.2; WP3.1) *(Reporting period: M12-18)*

For Wikipedia networks it is shown that the Arnoldi method can efficiently compute the eigenvalues and eigenstates with significant modulus of eigenvalues, these eigenstates correspond to well defined communities [1].

Community detection in social networks is a topic of central importance in modern graph mining, and the existence of overlapping communities has recently given rise to new interest in arc clustering. In [2], we propose the notion of triangular random walk as a way to unveil arc-community structure in social graphs: a triangular walk is a random process that insists differently on arcs that close a triangle. We prove that triangular walks can be used effectively, by translating them into a standard weighted random walk on the line graph; the dominant eigenvector of the associated process provides eigenweights that are used to enhance arc communities.

Our experiments show that the eigenweights so defined are in fact very helpful in determining the similarity between arcs and yield high-quality clustering. Even if our technique gives a weighting scheme on the line graph and can be combined with any node-clustering method in the final phase, to make our approach more scalable we also propose an algorithm (ALP) that produces the clustering directly without the need to build the weighted line graph explicitly. Our experiments show that ALP, besides providing the largest accuracy, it is also the fastest and most scalable among all arc-clustering algorithms we are aware of.

The relation between eigenvectors and communities is also studied for the citation network of Physical Review [3].

Publications M3:

[1] L.Ermann, K.M.Frahm and D.L.Shepelyansky, "Spectral properties of Google matrix of Wikipedia and other networks", Eur. Phys. J. B v.86, p.193 (2013), arXiv:1212.1068 [cs.IR]

[2] P.Boldi, M.Rosa, "Arc-Community Detection via Triangular Random Walks", LA-WEB 2012: 48-56 (2012)

[3] K.M.Frahm, Y.-H.Eom, D.L.Shepelyansky, :Google matrix of the citation network of Physical Review”, arXiv:1310.5624[physics.soc-ph] (2013)

Milestone M5: Network specific centrality measures (WP1.1, WP1.3, WP3.1,WP3.2) *(Reporting period: M18)*

In [1], we provide a mathematically sound survey of the most important classic centrality measures known from the literature and propose an axiomatic approach to establish whether they are actually doing what they have been designed for. Our axioms suggest some simple, basic properties that a centrality measure should exhibit. Surprisingly, only a new simple measure based on distances, harmonic centrality, turns out to satisfy all axioms; essentially, harmonic centrality is a correction to Bavelas’s classic closeness centrality designed to take unreachable nodes into account in a natural way. Our results suggest that centrality measures based on distances, which have been neglected in information retrieval in favor of spectral centrality measures in the last years, are of very high quality; moreover, harmonic centrality pops up as an excellent general-purpose centrality index for arbitrary directed graphs.

In [2,3] we report the results of the first world-scale social-network graph-distance computations, using the entire Facebook network of active users (approx 721 million users, approx 69 billion friendship links). The average distance we observe is 4.74, corresponding to 3.74 intermediaries or "degrees of separation", showing that the world is even smaller than we expected, and prompting the title of this paper. More generally, we study the distance distribution of Facebook and of some interesting geographic subgraphs, looking also at their evolution over time. The networks we are able to explore are almost two orders of magnitude larger than those analysed in the previous literature. We report detailed statistical metadata showing that our measurements (which rely on probabilistic algorithms) are very accurate.

Our studies [4] highlight deep differences in the structure of social networks and web graphs, show significant limitations of previous experimental results, and at the same time reveal clustering by label propagation as a new and very effective way of locating nodes that are important from a structural viewpoint.

Unfortunately, publishing social-network graphs is considered an ill-advised practice due to privacy concerns. To alleviate this problem, several anonymization methods have been proposed, aiming at reducing the risk of a privacy breach on the published data, while still allowing to analyze them and draw relevant conclusions. In [5] we introduce a new anonymization approach that is based on injecting uncertainty in social graphs and publishing the resulting uncertain graphs. While existing approaches obfuscate graph data by adding or removing edges entirely, we propose using a finer-grained perturbation that adds or removes edges partially: this way we can achieve the same desired level of obfuscation with smaller changes in the data, thus maintaining higher utility. Our experiments on real-world networks confirm that at the same level of identity obfuscation our method provides higher usefulness than existing randomized methods that publish standard graphs.

Deliverables are reported in [1-5].

Publications

[1] P. Boldi, S. Vigna. "Axioms for centrality", accepted for publication on Internet Mathematics. (arXiv:1308.2140, 2013)

[2] L.Backstrom, P.Boldi, M.Rosa, J.Ugander, and S.Vigna. **"Four degrees of separation"**, ACM Web Science 2012: Conference Proceedings, pages 45-54, ACM Press (2012); best paper award, highlighted by New York Timse; (arXiv:1111.4570, 2012)

[3] P.Boldi and S.Vigna. "Four degrees of separation, really" 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE, 2012: 1222-1227 (arXiv:1205.5509, 2012)

[4] P.Boldi, M.Rosa, S.Vigna, "Robustness of social and web graphs to node removal", Social Network Analysis and Mining, Springer: 1-14 (2012)

[5] P.Boldi, F.Bonchi, A.Gionis, T.Tassa, **"Injecting Uncertainty in Graphs for Identity Obfuscation"**, PVLDB 5(11): 1376-1387 (2012)

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

Milestone M 5: Network specific centrality measures (WP1.1, WP1.3, WP3.1,WP3.2)

(Reporting period: M18)

In [1] we propose the PageRank model of opinion formation and investigate its rich properties on real directed networks of Universities of Cambridge and Oxford, LiveJournal and Twitter. In this model the opinion formation of linked electors is weighted with their PageRank probability. We find that the society elite, corresponding to the top PageRank nodes, can impose its opinion to a significant fraction of the society. However, for a homogeneous distribution of two opinions there exists a bistability range of opinions which depends on a conformist parameter characterizing the opinion formation. We find that LiveJournal and Twitter networks have a stronger tendency to a totalitar opinion formation. We also analyze the Sznajd model generalized for scale-free networks with the weighted PageRank vote of electors. The case of Ulam networks is analyzed in [2], here the bistability is much less pronounced. Deliverables are published in [1,2].

Deliverables are reported in [1,2].

Publications

[1] V.Kandiah and D.L.Shepelyansky, "PageRank model of opinion formation on social networks", Physica A v.391, p.5779-5793 (2012), arXiv:1204.3806v1 [physics.soc-ph]

[2]L.Chakhmakhchyan and D.L.Shepelyansky, "PageRank model of opinion formation on Ulam networks", submitted to Phys. Lett. A , arXiv:1305.7395 [nlin.CD] (2013)

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)*

Twitter, a mixture of a social network and a news media, has in the past years became the largest medium where users may spread information along their social contacts. In our research we investigated retweeting, the typical Twitter method for re-sharing and spreading the information. Retweets form cascades and information pathways. We investigate [1] the possibility of predicting the future popularity of emerging retweet cascades. We evaluate our algorithms on four different datasets that contain four different sets of tweets and the underlying follower network of the tweeting users.

Our results measure the influence of messages sent over Twitter. Cha et al. [M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, “Measuring user influence in twitter: The million follower fallacy,” in 4th International AAAI Conference on Weblogs and Social Media (ICWSM), 2010.] define influence as ``... the power of capacity of causing an effect in indirect intangible ways...". Our goal is to predict the success, on the individual message level. We analyze how certain messages may reach out to a large number of Twitter users. In contrast to a earlier investigation for analyzing the influence of users, we investigate each tweet by taking both the source user and the content into account. Our findings justify the importance of the user itself as well as content, most notable the hashtags used in the message. In our experiments we use the data set of Andreas Kaltenbrunner [V. Gómez, H. J. Kappen, and A. Kaltenbrunner, “Modeling the structure and evolution of discussion cascades,” in Proceedings of the 22nd ACM conference on Hypertext and hypermedia, pp. 181–190, ACM, 2011].

Publications

[1] A.Benczur et al. Article is in preparation by P3

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)*

Several results show the influence of friends and contacts to spread obesity, loneliness, alcohol consumption, religious belief and many similar properties in social networks. Others question the methodology of these experiments by proposing that the measured effects may be due to homophily, the fact that people tend to associate with others like themselves, and a shared environment also called confounding or contextual influence.

Our goal was to exploit the timely information gathered by the Last.fm service on users with public profile to investigate how members of the social network may influence their friends' taste. Last.fm's service is unique in that we may obtain a detailed timeline and catch immediate effects by comparing the history of friends in time and comparing to pairs of random users instead of friends.

Our results confirm the existence of influence through the social network as opposed to the pure similarity of taste between friends. We disproved the opinion that homophily could be the reason for friends listening to the same music or behave similarly by constructing a baseline that takes homophily and temporal effects into account. Over the baseline recommender, we achieved a 4% improvement in recommendation accuracy when presenting artists from friends' past scrobbles that the given user had never seen before. Our system has very strong time awareness: when we recommend, we look back in the near past and combine friends' scrobbles with the baseline methods. The influence from a friend at a given time is certain function of the observed influence in the past and the time elapsed since the friend scrobbled the given artist. Results are reported in [1].

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 [2].

Publications

[1] Róbert Pálovics, András Benczúr. Temporal influence over the Last.fm social network. The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM 2013 Niagara Falls, Canada, August 25-28, 2013

[2] R. Pálovics, L.Ermann, A.D.Chepelianskii, A.Benczúr, D.L.Shepelyansky, "Recommendation systems for users and spots of Paris, France via NOMAO data sets", in preparation

Deliverable D3.1: 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.

Distributed key-value stores provide random access to the graph, this solution is however the slowest of the alternatives. Hadoop is apparently a mature Map-Reduce infrastructure capable of efficiently implementing graph algorithms such as PageRank or connected components. The Stratosphere project offers improved efficiency for iterative MapReduce operations as needed these graph algorithms. For Bulk Synchronous Parallel algorithms, open source infrastructures are yet less mature. We tested HAMA, an incubatory project as well as GraphLab that offers both an easy-to-use multi-core external memory and an under progress no-shared-memory distributed version. Summaries of our findings will be submitted in 2013.

In our research [1] we also experimented with distributed streaming environments, and achieved remarkably high throughput with low latency using a properly designed streaming architecture.

Publications

[1] Real-time streaming mobility analytics. Andras Garzo, Andras A. Benczur, Csaba Istvan Sidlo, Daniel Tahara, Erik Francis Wyatt. IEEE Big Data 2013

Task WP 4.2. Network quality and trust classification and spam filtering (MTA_SZTAKI, UMIL, UTWE)

Milestone M 4: Spam filter protocols for (WP4.2)*(Reporting period: M18)*

In [1] we presented comprehensive Web classification experimentation based on content, link as well as temporal features, both new and recently published. Our spam filtering baseline classification procedures are collected by analyzing the results of the Web Spam Challenges and the ECML/PKDD Discovery Challenge 2010.

It has already been known from the early results on text classification that obtaining classification labels is expensive. This is especially true in a European project where multilingual collections have to be produced, needing manual assessment for each language in question, or techniques of cross-lingual information retrieval or machine translation have to be used. While several earlier results focus on cross-lingual classification of general text corpora, most of them use heavy natural language processing that is out of reach for large collections.

In [2] we test methods for classifying non-English Web collections based on English labelled collections only. We observe that hosts with a mix of English and national language content, likely translations, yield a very powerful resource for cross-lingual classification. Some of our methods work even without using dictionaries, not to mention without more complex tools of natural language processing. The bag-of-words representation together with appropriate machine learning techniques is the strongest method for crosslingual Web classification.

Deliverables are reported in [1,2].

Publications

[1] Miklos Erdelyi, Andras A. Benczur. Balint Daroczy, Andras Garzo, Tamas Kiss, David Siklosi. The classification power of Web features. Internet Mathematics, to appear

[2] András Garzó, Bálint Daróczy, Tamás Kiss, Dávid Siklósi, András A. Benczúr. Cross-lingual web spam classification. The 3rd Joint WICOW/AIRWeb Workshop on Web Quality Rio de Janeiro, Brasil. May 13, 2013. Proceedings of the 22nd international conference on World Wide Web companion

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)*(Reporting period: M36)*

Work is in progress. In [1] we ask how different cultures evaluate a person. Is an important person in one culture also important in another culture? We address these questions via ranking of multilingual Wikipedia articles. With three ranking algorithms based on network structure of Wikipedia, we assign ranks to all articles in 9 multilingual editions of Wikipedia and investigate general ranking structure of PageRank, CheiRank and 2DRank. In particular, we focus on articles related to persons, identify top 30 persons for each rank among different editions and analyze distinctive properties of their distributions over activity fields such as politics, art, science, religion, sport for each edition. We find that local heroes are dominant but also global heroes exist and create an effective network representing entanglement of cultures. The Google matrix analysis of network of cultures shows signs of the Zipf law distribution. This approach allows to examine diversity and shared characteristics of knowledge organization between cultures. The developed computational, data driven approach highlights cultural interconnections in a new perspective. The network data for 9 Wikipedia editions are collected by S.Vigna (P4).

The properties of eigenstates of complex network of game Go are investigated in [2].

Publications M13:

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

[2] V.Kandiah, B.Georgeot, O.Giraud, “Communities of moves in the complex network of game Go”, in preparation

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 research in this line is in progress. The Google matrix analysis of the world trade network from UN COMTRADE was first performed in [1]. The extention of this analysis to multiproduct trade meets certain mathemetical difficulties. For example, import and export are symmetric and non-democratic in products. Our study [2] indicates the breaking of this symmetry in multiproduct trade. In [2] we apply the concepts from ecological systems. Ecological systems have a high level of complexity combined with stability and rich biodiversity. Recently, the analysis of their properties and evolution has been pushed forward on a basis of concept of mutualistic networks that provides a detailed understanding of their features being linked to a high nestedness of these networks. It was shown that the nestedness architecture of mutualistic networks of plants and their pollinators minimizes competition and increases biodiversity. Here, using the United Nations COMTRADE database for years 1962 - 2009, we show that a similar ecological analysis gives a valuable description of the world trade. In fact the countries and trade products are analogous to plants and pollinators, and the whole trade network is characterized by a low nestedness temperature which is typical for the ecological networks. This approach provides new mutualistic features of the world trade highlighting new significance of countries and trade products for the world trade. At present we make advancements in multiproduct trade analysis.

Publications M8:

[1] L.Ermann and D.L.Shepelyansky, "Google matrix of the world trade network", Acta Physica Polonica A v.120(6A), pp. A158-A171 (2011), arxiv:1103.5027[physics.soc-ph]

[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]

Deliverable D4.1: 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)*

Although web crawlers have been around for twenty years by now, there is virtually no freely available, open-source crawling software that guarantees high throughput, overcomes the limits of single-machine tools and at the same time scales linearly with the amount of resources available. The work within this mileston aims at filling this gap.

In the first eighteen months of the project we have developed BUbiNG, our next-generation web crawler built upon our experience with UbiCrawler and on the last ten years of research on the topic. BUbiNG is an open-source Java fully distributed crawler (no central coordination), and single BUbiNG agents using sizeable hardware can crawl several thousands pages (per agent) per second respecting strict politeness constraints, both host- and IP-based. Unlike existing open-source distributed crawlers that rely on batch techniques (like MapReduce), BUbiNG job distribution is based on modern high-speed protocols so to achieve very high throughput.

BubiNG will be used in the remaining months of the project to collect large-scale web datasets. A first 1 billion pages dataset (graph) of the European web will be made available within M20.

At UTwente, in collaboration with UMilano, a master student Anne Buijsrogge has started her graduation project on mathematical analysis and optimization of the data storage policies of the crawler, using the techniques from queueing theory and stochastic optimization.

Results are reported in [1].

Publications

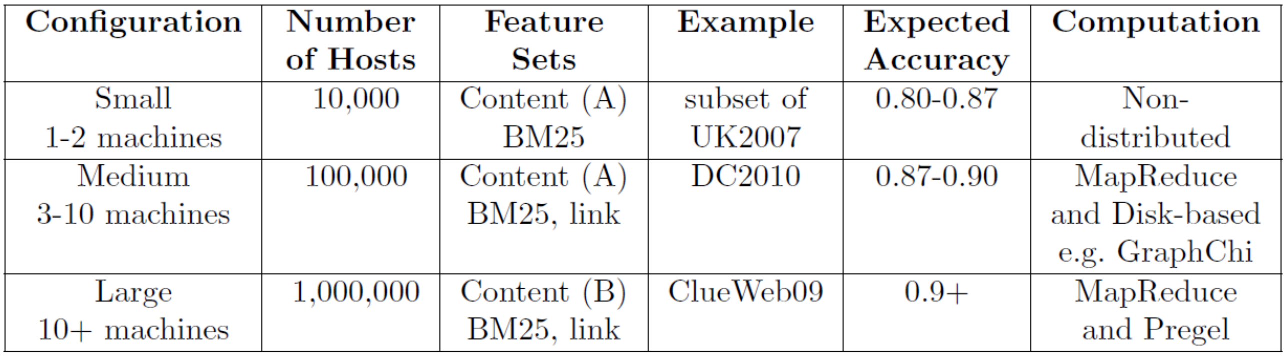
[1] P. Boldi, A. Marino, M.Santini, S. Vigna. "BUbiNG: Massive crawling for the masses"**,** submitted for publication, Oct 2013

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 [1] we address the computational efficiency of Web feature generation. The first expensive step involves parsing to create terms and links. The time requirement scales linearly with the number of pages. For a very large collection such as ClueWeb09, distributed processing was necessary: over 45 old machines running Hadoop 0.21, we parsed the uncompressed 9.5TB English part of ClueWeb09 in 36 hours. Additional tasks such as term counting, BM25 or content feature generation fits within the same time frame. Host level aggregation allows us to proceed with a much smaller size data. However for aggregation we need to store a large number of partial feature values for all hosts unless we sort the entire collection by host, again by external memory or Map-Reduce sort. The following features however remain expensive: Page level PageRank that is also required for all content features involving the maximum PageRank page of the host; as well as the page level features involving multi-step neighborhood such as neighbourhood size at distance k as well as graph similarity. In order to be able to process graphs of ClueWeb09 scale (4.7 billion nodes and 17 billion edges), we implemented message passing C++ codes.

As practical message, we may conclude that, as seen the Table, single machines may compute content and BM25 features for a few 10,000 hosts only. Link features need additional resources and either compressed, disk based or, in the largest configuration, Pregel-like distributed infrastructures.



Configuration	Number of Hosts	Feature Sets	Example	Expected Accuracy	Computation
Small 1-2 machines	10,000	Content (A) BM25	subset of UK2007	0.80-0.87	Non-distributed
Medium 3-10 machines	100,000	Content (A) BM25, link	DC2010	0.87-0.90	MapReduce and Disk-based e.g. GraphChi
Large 10+ machines	1,000,000	Content (B) BM25, link	ClueWeb09	0.9+	MapReduce and Pregel

Publications

[1] Miklos Erdelyi, Andras A. Benczur. Balint Daroczy, Andras Garzo, Tamas Kiss, David Siklosi. The classification power of Web features. Internet Mathematics, to appear

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)*

Research is in progress. We analyzed the time evolution of Wikipedia network in [1] in collaborative paper of P1 and P3. We study the time evolution of ranking and spectral properties of the Google matrix of English Wikipedia hyperlink network during years 2003 - 2011. The statistical properties of ranking of Wikipedia articles via PageRank and CheiRank probabilities, as well as the matrix spectrum, are shown to be stabilized for 2007 - 2011. A special emphasis is done on ranking of Wikipedia personalities and universities. We show that PageRank selection is dominated by politicians while 2DRank, which combines PageRank and CheiRank, gives more accent on personalities of arts. The Wikipedia PageRank of universities recovers 80 percents of top universities of Shanghai ranking during the considered time period.

And in [2], we develop methods to automatically discover temporal events along important connections. For our experiments we selected Wikipedia as a clean corpus where measurements are not biased for example by date identification, yet the methods can directly be applied for any hyperlinked collection. Wikipedia is certainly the most used and best-known online encyclopedia and knowledge-base of the past decade. Almost every action or event, be it tiny or slightly remarkable, immediately appears in blog posts, news articles or sometimes even in Wikipedia articles. For ranking we use both the link structure and the article content. The user can specify a query and should get a “temporally changing” subgraph of relevant articles. First we try to find the relevant articles respective to the query by utilizing a text search engine. As the next step, based on these articles, we try to find those nodes that have not only important changes according to the definition above but also their content is related to the original query. In a recursive definition reminiscent of PageRank and HITS, we will consider the change of a page relevant if relevant changes can be observed in the neighborhood of the page as well. We retrieve and present the top ranking articles and their linkage.

Finally, we extended link-based similarity algorithms by proposing metrics to capture the linkage change of Web pages over time [3]. We describe a method to calculate these metrics efficiently on the Web graph and then measure their performance when used as features in Web spam classification. We propose an extension of two link-based similarity measures: XJaccard and PSimRank.

Publications

[1] Y.-H.Eom, K.M.Frahm, A.Benczur and D.L.Shepelyansky, "Time evolution of Wikipedia network ranking", submitted EPJB, arXiv:1304.6601 [physics.soc-ph] (2013)

[2] Julianna Göbölös-Szabó, András Benczúr. Temporal Wikipedia search by edits and linkage. SIGIR 2013 Workshop on Time-aware Information Access, 28 July – 1 August 2013, Dublin, Ireland

[3] Miklos Erdelyi, Andras A. Benczur. Balint Daroczy, Andras Garzo, Tamas Kiss, David Siklosi. The classification power of Web features. Internet Mathematics, to appear

Deliverable D5.1: 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	Delive red
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	prog
7	Protocols for large-scale network processing	4,5	R	PU	36	prog
8	Characterization of multiproduct world trade network	4	R	PU	36	prog
9	Webcrawler development and database collection	5	R	PU	36	prog
10	Monte Carlo algorithms for centrality measures	1	R	PU	36	prog
11	Delocalization conditions for Google matrix eigenstates	2	R	PU	36	prog
12	New protocols for social voting and recommendation	3	R	PU	36	prog
13	Characterization of ranking of Wikipedia and other networks	4	R	PU	36	prog
14	Characterization of time evolving Web structures	5	R	PU	36	prog
	Project website	1-5	O	PU	6	yes
	1st report	1-5	R	RE	18	yes

	2nd report					
	Final report					
	Final plan for using and disseminating knowledge					

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 first 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 - 31 March 2015 (30 months); PhD student Vivek Kandiah also participates in the project being supported by CNRS-Region-Midi-Pyrenees.

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)

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. Planified and actual working person-months are reported in Annex person-month status table.

List of project meetings

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:

- kick off NADINE meeting was organized in the frame of Workshop Spectral Properties of Complex Networks, *European Center for Theoretical Studies in Nuclear Physics and Related Areas (ECT*)*, Trento, Italy (23.07. – 27.07.2012); web page http://www.quantware.ups-tlse.fr/complexnetworks2012/

- workshop Directed networks days was organized at Dipartimento di Informatica, Universita degli Studi di Milano, 13-14 June 2013; web page http://www.quantware.ups-tlse.fr/FETNADINE/dnd2013/

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 37 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 first reporting period

[1] **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)

This 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] **Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles**

Authors: Y.-H.Eom and D.L. Shepelyansky
Journal: PLoS ONE v.8(10), p.e74554 (2013) (arXiv:1306.6259 [cs.SI], 2013)

This paper determines top 30 persons for Wikipedia editions in 9 different languages. The network of links between languages or cultures is determined attributing persons to their native culture in these 9 editions. We show that the global hero for 9 editions is Napoleon for PageRank and Michael Jackson for 2DRank.

[3] **Degree-degree correlations in random graphs with heavy-tailed degrees**

Authors: N. Litvak, and R. van der Hofstad,
Journal: Accepted in Internet Mathematics. (arXiv:1202.3071 [math.PR], 2013)

We investigate the degree-degree dependencies in networks as described by the Pearson correlation coefficient, and show that it is non-negative in the large graph limit when the asymptotic degree distribution has an infinite third moment. Furthermore, we provide examples where the Pearson's correlation coefficient converges to zero in a network with strong negative degree-degree dependencies, and another example where this coefficient converges in distribution to a random variable. We suggest the alternative degree-degree dependency measure, based on Spearman's parameter, and prove that this statistical estimator converges to an appropriate limit under quite general conditions. These conditions are proved to hold in common network models, such as the configuration model and the preferential attachment model. We conclude that rank correlations provide a suitable and informative method for uncovering network mixing patterns.

[4] **Temporal influence over the Last.fm social network.**

Authors: Róbert Pálovics, András Benczúr.
Conference: The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM 2013 Niagara Falls, Canada, August 25-28, 2013

This paper is the first step of NADINE research towards exploiting the graph of social contacts in recommender systems. The work initiated in this paper is continued in two directions. First, we constructed a densification law model for the spread of information over the social network and at present we analyze its properties. Second we extend our research on recommender systems by embedding both the network and additional information such as geographic locations into the core factor model. The extended version of the paper is invited for submission to the “Social Network Analysis and Mining” journal, (SNAM) by Springer.

[5] **In-core computation of geometric centralities with HyperBall: A hundred billion nodes and beyond.**

Authors: P.Boldi, S.Vigna
Journal: To appear in the Proceedings of 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW 2013). (arXiv:1308.2144, 2013)

Abstract: Given a social network, which of its nodes are more central? This question has been asked many times in sociology, psychology and computer science, and a whole plethora of centrality measures (a.k.a. centrality indices, or rankings) were proposed to account for the importance of the nodes of a network. In this paper, we approach the problem of computing geometric centralities, such as closeness and harmonic centrality, on very large graphs; traditionally this task requires an all-pairs shortest-path computation in the exact case, or a number of breadth-first traversals for approximated computations, but these techniques yield very weak statistical guarantees on highly disconnected graphs. We rather assume that the graph is accessed in a semi-streaming fashion, that is, that adjacency lists are scanned almost sequentially, and that a very small amount of memory (in the order of a dozen bytes) per node is available in core memory. We leverage the newly discovered algorithms based on HyperLogLog counters, making it possible to approximate a number of geometric centralities at a very high speed and with high accuracy. While the application of similar algorithms for the approximation of closeness was attempted in the MapReduce framework, our exploitation of HyperLogLog counters reduces exponentially the memory footprint, paving the way for in-core processing of networks with a hundred billion nodes using “just” 2TiB of RAM. Moreover, the computations we describe are inherently parallelizable, and scale linearly with the number of available cores.

Dissemination: List and abstracts of papers and preprints appeared during the 1st reporting period within the framework of NADINE:

[1] **Towards two-dimensional search engines**

Authors: L.Ermann, A.D.Chepelianskii and D.L.Shepelyansky
Journal: J. Phys. A: Math. Theor. v.45, p.275101 (2012) (arXiv:1106.6215[cs.IR])

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

[2] **Ecological analysis of world trade**

Authors: L.Ermann and D.L.Shepelyansky
Journal: Phys. Lett. A v.377, p.250 (2013) (arXiv:1201.3584[q-fin.GN], 2012)

Abstract: Ecological systems have a high level of complexity combined with stability and rich biodiversity. Recently, the analysis of their properties and evolution has been pushed forward on a basis of concept of mutualistic networks that provides a detailed understanding of their features being linked to a high nestedness of these networks. It was shown that the nestedness architecture of mutualistic networks of plants and their pollinators minimizes competition and increases biodiversity. Here, using the United Nations COMTRADE database for years 1962 - 2009, we show that a similar ecological analysis gives a valuable description of the world trade. In fact the countries and trade products are analogous to plants and pollinators, and the whole trade network is characterized by a low nestedness temperature which is typical for the ecological networks. This approach provides new mutualistic features of the world trade highlighting new significance of countries and trade products for the world trade.

[3] **PageRank model of opinion formation on social networks**

Authors: V.Kandiah and D.L.Shepelyansky
Journal: Physica A v.391, p.5779 (2012) (arXiv:1204.3806v1 [physics.soc-ph], 2012)

Abstract: We propose the PageRank model of opinion formation and investigate its rich properties on real directed networks of Universities of Cambridge and Oxford, LiveJournal and Twitter. In this model the opinion formation of linked electors is weighted with their PageRank probability. We find that the society elite, corresponding to the top PageRank nodes, can impose its opinion to a significant fraction of the society. However, for a homogeneous distribution of two opinions there exists a bistability range of opinions which depends on a conformist parameter characterizing the opinion formation. We find that LiveJournal and Twitter networks have a stronger tendency to a totalitar opinion formation. We also analyze the Sznajd model generalized for scale-free networks with the weighted PageRank vote of electors.

[4] **PageRank of integers**

Authors: K.M.Frahm, A.D.Chepelianskii and D.L.Shepelyansky
Journal: J. Phys. A: Math. Theor. v.45, p.405101 (2012) (arXiv:1205.6343[cs.IR], 2012)

Abstract: We build up a directed network tracing links from a given integer to its divisors and analyze the properties of the Google matrix of this network. The PageRank vector of this matrix is computed numerically and it is shown that its probability is inversely proportional to the PageRank index thus being similar to the Zipf law and the dependence established for the World Wide Web. The spectrum of the Google matrix of integers is characterized by a large gap and a relatively small number of nonzero eigenvalues. A simple semi-analytical expression for the PageRank of integers is derived that allows to find this vector for matrices of billion size. This network provides a new PageRank order ofintegers.

[5] **Google matrix of Twitter**

Authors: K.M.Frahm and D.L.Shepelyansky
Journal: Eur. Phys. J. B v.85, p.355 (2012) (arXiv:1207.3414[cs.SI], 2012)

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

[6] **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: We study 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. We also show that the correlator between PageRank and CheiRank vectors distinguishes different organizations of information flow on BBC and Le Monde web sites.

[7] **Google matrix analysis of DNA sequences**

Authors: L.Ermann, K.M.Frahm and D.L.Shepelyansky
Journal: PLOS One v.8(5), p. e61519 (2013) (arXiv:1301.1626[q-bio.GN], 2013)

Abstract: For DNA sequences of various species we construct the Google matrix G of Markov transitions between nearby words composed of several letters. The statistical distribution of matrix elements of this matrix is shown to be described by a power law with the exponent being close to those of outgoing links in such scale-free networks as the World Wide Web (WWW). At the same time the sum of ingoing matrix elements is characterized by the exponent being significantly larger than those typical for WWW networks. This results in a slow algebraic decay of the PageRank probability determined by the distribution of ingoing elements. The spectrum of G is characterized by a large gap leading to a rapid relaxation process on the DNA sequence networks. We introduce the PageRank proximity correlator between different species which determines their statistical similarity from the view point of Markov chains. The properties of other eigenstates of the Google matrix are also discussed. Our results establish scale-free features of DNA sequence networks showing their similarities and distinctions with the WWW and linguistic networks.

[8] **Time evolution of Wikipedia network ranking**

Authors: Y.-H.Eom, K.M.Frahm, A.Benczur and D.L.Shepelyansky
Journal: submitted Eur. Phys. J. B (2013) (arXiv:1304.6601 [physics.soc-ph], 2013)

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

[9] **PageRank model of opinion formation on Ulam networks**

Authors: L.Chakhmakhchyan and D.L. Shepelyansky
Journal: to appear in Phys. Lett. A (2013) (arXiv:1305.7395 [nlin.CD], 2013)

Abstract: We consider a PageRank model of opinion formation on Ulam networks, generated by the intermittency map and the typical Chirikov map. The Ulam networks generated by these maps have certain similarities with such scale-free networks as the World Wide Web (WWW), showing an algebraic decay of the PageRank probability. We find that the opinion formation process on Ulam networks have certain similarities but also distinct features comparing to the WWW. We attribute these distinctions to internal differences in network structure of the Ulam and WWW networks. We also analyze the process of opinion formation in the frame of generalized Sznajd model which protects opinion of small communities.

[10] **Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles**

Authors: Y.-H.Eom and D.L. Shepelyansky
Journal: PLoS ONE v.8(10), p.e74554 (2013) (arXiv:1306.6259 [cs.SI], 2013)

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

[11] **Poincare recurrences and Ulam method for the Chirikov standard map**

Authors: K.M.Frahm and D.L. Shepelyansky
Journal: Eur. Phys. J. B v.86, p.322(2013) (arXiv:1302.2761 [nlin.CD] , 2013)

Abstract: We study numerically the statistics of Poincare recurrences for the Chirikov standard map and the separatrix map at parameters with a critical golden invariant curve. The properties of recurrences are analyzed with the help of a generalized Ulam method. This method allows to construct the corresponding Ulam matrix whose spectrum and eigenstates are analyzed by the powerful Arnoldi method. We also develop a new survival Monte Carlo method which allows us to study recurrences on times changing by ten orders of magnitude. We show that the recurrences at long times are determined by trajectory sticking in a vicinity of the critical golden curve and secondary resonance structures. The values of Poincare exponents of recurrences are determined for the two maps studied. We also discuss the localization properties of eigenstates of the Ulam matrix and their relation with the Poincare recurrences.

[12] **Google matrix of the citation network of Physical Review**

Authors: K.M.Frahm, Y.-H.Eom, D.L.Shepelyansky

Journal submitted to Phys Rev E, arXiv:1310.5624[physics.soc-ph] (2013)

Abstract: We study the statistical properties of spectrum and eigenstates of the Google matrix of the citation network of Physical Review for the period 1893 - 2009. The main fraction of complex eigenvalues with largest modulus is determined numerically using the computational precision with up to 16384 binary digits that allows to resolve hard numerical problems for small eigenvalues. The nearly nilpotent matrix structure allows to obtain semi-analytical computation of eigenvalues. We find that the spectrum is characterized by the fractal Weyl law with a fractal dimension approximately 1. It is found that the majority of eigenvectors are located in a localized phase. The statistical distribution of articles in the PageRank-CheiRank plane is established providing a better understanding of information flows on the network. The concept of ImpactRank is proposed to determine an influence domain of a given article. We also discuss the properties of random matrix models of Perron-Frobenius operators.

[13] **Uncovering disassortativity in large scale-free networks**

Authors: N. Litvak and R. van der Hofstad
Journal: Physical Review E, v. 87, p. 022801 (2013) (arXiv:1204.0266 [physics.soc-ph])

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

[14] **Degree-degree correlations in random graphs with heavy-tailed degrees.**

Authors: N. Litvak and R. van der Hofstad
Journal: Internet Mathematics, accepted. (arXiv:1202.3071 [math.PR]) (2013)

Abstract: Mixing patterns in large self-organizing networks, such as the Internet, the World Wide Web, social and biological networks are often characterized by degree-degree dependencies between neighbouring nodes. In *assortative* networks, the degree-degree dependencies are positive (nodes with similar degrees tend to connect to each other), while in *disassortative* networks, these dependencies are negative. One of the problems with the commonly used Pearson correlation coefficient, also known as the *assortativity coefficient* is that its magnitude decreases with the network size in disassortative networks. This makes it impossible to compare mixing patterns, for example, in two web crawls of different sizes. As an alternative, we have recently suggested to use rank correlation measures, such as Spearman's rho. Numerical experiments have confirmed that Spearman's rho produces consistent values in graphs of different sizes but similar structure, and it is able to reveal strong (positive or negative) dependencies in large graphs. In this paper we analytically investigate degree-degree dependencies for scale-free graph sequences. In order to demonstrate the ill behaviour of the Pearson's correlation coefficient, we first study a simple model of two heavy-tailed highly correlated random variables X and Y, and show that the sample correlation coefficient converges in distribution either to a proper random variable on [-1,1], or to zero, and the limit is non-negative a.s. if X,Y≥ 0. We next adapt these results to the degree-degree dependencies in networks as described by the Pearson correlation coefficient, and show that it is non-negative in the large graph limit when the asymptotic degree distribution has an infinite third moment. Furthermore, we provide examples where the Pearson's correlation coefficient converges to zero in a network with strong negative degree-degree dependencies, and another example where this coefficient converges in distribution to a random variable. We suggest the alternative degree-degree dependency measure, based on Spearman's rho, and prove that this statistical estimator converges to an appropriate limit under quite general conditions. These conditions are proved to hold in common network models, such as the configuration model and the preferential attachment model. We conclude that rank correlations provide a suitable and informative method for uncovering network mixing patterns.

[15] **Quck detection of nodes with large degrees**
Authors: K.Avrachenkov,N.Litvak, M.Sokol, D.Towsley
Journal: Conference proceedings (extended abstract): 9th International Workshop on Algorithms and Models for the Web Graph, WAW 2012, 22-23 June 2012, Halifax, NS, Canada. pp. 54-65. Lecture Notes in Computer Science 7323. Springer Verlag. (arXiv:1202.3261v1 [cs.DS], 2012) Journal (full version): Internet Mathematics, accepted.

Abstract: Our goal is to quickly find top k lists of nodes with the largest degrees in large complex networks. If the adjacency list of the network is known (not often the case in complex networks), a deterministic algorithm to find the top k list of nodes with the largest degrees requires an average complexity of O(n), where n is the number of nodes in the network. Even this modest complexity can be very high for large complex networks. We propose to use the random walk based method. We show theoretically and by numerical experiments that for large networks the random walk method finds good quality top lists of nodes with high probability and with computational savings of orders of magnitude. We also propose stopping criteria for the random walk method which requires very little knowledge about the structure of the network.

[16] **Alpha current flow betweenness centrality**

Authors:K.Avrachenkov, N. Litvak, V. Medyanikov, and M. Sokol
Journal: Conference proceedings: 10th Workshop on Algorithms and Models for the Web Graph, WAW2013, 15-16 December, 2013, Harvard University (arXiv:1308.2591v1 [cs.SI], 2013)

Abstract: A class of centrality measures called betweenness centralities reflects degree of participation of edges or nodes in communication between different parts of the network. The original shortest-path betweenness centrality is based on counting shortest paths which go through a node or an edge. One of shortcomings of the shortest-path betweenness centrality is that it ignores the paths that might be one or two hops longer than the shortest paths, while the edges on such paths can be important for communication processes in the network. To rectify this shortcoming a current flow betweenness centrality has been proposed. Similarly to the shortest-path betweenness, it has prohibitive complexity for large size networks. In the present work we propose two regularizations of the current flow betweenness centrality, alpha-current flow betweenness and truncated alpha-current flow betweenness, which can be computed fast and correlate well with the original current flow betweenness.

[17] **Quick detection of popular entities in large directed networks.** Submitted. The paper has not been not published on arXiv for the purpose of the blind review.

Authors: L. Ostroumova, K. Avrachenkov and N. Litvak
Journal/Conference: Submitted to a top Computer Science conference.

Abstract: In this paper, we address a problem of quick detection of popular entities in large online social networks. Practical importance of the problem is attested by a large number of companies that continuously collect and update statistics about popular entities.
We suggest an efficient two-stage algorithm for solving this problem. For instance, our algorithm needs only one thousand API requests in order to find the top-50 most popular users in Twitter, a

network with more than a billion of registered users. Our algorithm is easy to implement, it outperforms existing methods, and serves many different purposes, such as finding most popular users or most popular interest groups in social networks. An important contribution of this work is he analysis of the proposed algorithm using the Extreme Value Theory~-- a branch of probability that studies extreme events and properties of largest order statistics in random samples. Using this theory, we derive accurate predictions for the algorithm's performance and show that the number of API requests for finding top-k most popular entities is sublinear in the number of entities. Moreover, we formally show that the high variability among the entities, expressed through heavy-tailed distributions, is the reason for the algorithm's efficiency. We quantify this phenomenon in a rigorous mathematical way.

[18] **Degree-degree correlations in directed networks with heavy-tailed degrees.**

Authors: P. van der Hoorn and N. Litvak
Journal: submitted October, arXiv:1310.6528[math.PR] (2013)
Abstract: The network theory of Pearson's correlation coefficients are commonly used to measure the degree assortativity of a network. We investigate the behavior of these coefficients in the setting of directled networks with heavy-tailed degree disequences. We prove that for graphs where the in- and out-degree sequences satisfy a power law, Pearson's correlation coefficients converge to a non-negative number in the infinite network size limit. We propose alternative measures for degree-degree correlations in directed networks based on Spearman's rho and Kemdall's tau. Using examples and calculations on the Wikipedia graphs for nine different languages, we show why these rank correlation measures are more suited for measuring degree assortativity in directed graphs with heavy-tailed degrees.

[19] **Temporal influence over the Last.fm social network.**

Authors: Róbert Pálovics, András Benczúr.

Conference: The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM 2013 Niagara Falls, Canada, August 25-28, 2013

https://dms.sztaki.hu/sites/dms.sztaki.hu/files/file/2013/lastfm-asonam.pdf

Abstract: Several recent results show the influence of social contacts to spread certain properties over the network, but others question the methodology of these experiments by proposing that the measured effects may be due to homophily or a shared environment. In this paper we justify the existence of the social influence by considering the temporal behavior of Last.fm users. In order to clearly distinguish between friends sharing the same interest, especially since Last.fm recommends friends based on similarity of taste, we separated the timeless effect of similar taste from the temporal impulses of immediately listening to the same artist after a friend. We measured strong increase of listening to a completely new artist in a few hours period after a friend compared to non-friends representing a simple trend or external influence. In our experiment to eliminate network independent elements of taste, we improved collaborative filtering and trend based methods by blending with simple time aware recommendations based on the influence of friends. Our experiments are carried over the two-year "scrobble" history of 70,000 Last.fm users.

[20] **Cross-lingual web spam classification.**

Authors: A. Garzó, B. Daróczy, T. Kiss, D. Siklósi, A.A. Benczúr.

Journal/conference: The 3rd Joint WICOW/AIRWeb Workshop on Web Quality Rio de Janeiro, Brasil. May 13, 2013. Proceedigs of the 22nd international conference on World Wide Web companion

https://dms.sztaki.hu/sites/dms.sztaki.hu/files/file/2013/crosslingual-short.pdf

Abstract: While English language training data exists for several Web classification tasks, most notably for Web spam, we face an expensive human labeling procedure if we want to classify a Web domain in a language different from English. In this paper we overview how existing content and link based classification techniques work, how models can be ``translated'' from English into another language, and how language-dependent and independent methods combine. In particular we show that simple bag-of-words translation works very well and in this procedure we may also rely on mixed language Web hosts, i.e. those that contain an English translation of part of the local language text. Our experiments are conducted on the ClueWeb09 corpus as the training English collection and a large Portuguese crawl of the Portuguese Web Archive. To foster further research, we provide labels and precomputed values of term frequencies, content and link based features for both ClueWeb09 and the Portuguese data.

[21] **The classification power of Web features.**

Authors: Miklos Erdelyi, Andras A. Benczur. Balint Daroczy, Andras Garzo, Tamas Kiss, David Siklosi.

Journal: Internet Mathematics, to appear

Abstract: In this paper we give a comprehensive overview of features devised for Web spam detection and investigate how much various classes, some requiring very high computational effort, add to the classification accuracy. We collect and handle a large number of features based on recent advances in Web spam filtering, including temporal ones, in particular we analyze the strength and sensitivity of linkage change. We propose new temporal link similarity based features and show how to compute them efficiently on large graphs. We show that machine learning techniques including ensemble selection, LogitBoost and Random Forest significantly improve accuracy. We conclude that, with appropriate learning techniques, a simple and computationally inexpensive feature subset outperforms all previous results published so far on our data set and can only slightly be further improved by computationally expensive features. We test our method on three major publicly available data sets, the Web Spam Challenge 2008 data set WEBSPAM-UK2007, the ECML/PKDD Discovery Challenge data set DC2010 and the Waterloo Spam Rankings for ClueWeb09.

We make several feature sets and source codes public, including the temporal features of eight .uk crawl snapshots that include WEBSPAM-UK2007 as well as the Web Spam Challenge features for the labeled part of ClueWeb09.

[22] **Temporal Wikipedia search by edits and linkage.**

Authors: J. Göbölös-Szabó, A.A. Benczúr.

Conference: SIGIR 2013 Workshop on Time-aware Information Access, 28 July - 1 August 2013, Dublin, Ireland

https://dms.sztaki.hu/sites/dms.sztaki.hu/files/file/2013/temp-wimmut.pdf

Abstract: In this paper we exploit the connectivity structure of edits in Wikipedia to identify recent events that happened at a given time via identifying bursty changes in linked articles around a specied date. Our key results include algorithms for node relevance ranking in temporal subgraph and neighborhood selection based on measurements for structural changes in time over the Wikipedia link graph. We measure our algorithms over manually annotated queries with relevant events in September and October 2011; we make the assessment publicly available. While our methods were tested over clean Wikipedia metadata, we believe the methods are applicable to general temporal Web collections as well.

[23] **Real-time streaming mobility analytics**

Authors: Andras Garzo, Andras A. Benczur, Csaba Istvan Sidlo, Daniel Tahara, Erik Francis Wyatt

Conference: IEEE Big Data 2013

https://dms.sztaki.hu/sites/dms.sztaki.hu/files/file/2013/pid2922315.pdf

Abstract: Location prediction over mobility traces may find applications in navigation, traffic optimization, city planning and smart cities. Due to the scale of the mobility in a metropolis, real time processing is one of the major Big Data challenges. In this paper we deploy distributed streaming algorithms and infrastructures to process large scale mobility data for fast reaction time prediction. We evaluate our methods on a data set derived from the Orange D4D Challenge data representing sample traces of Ivory Coast mobile phone users. Our results open the possibility for efficient real time mobility predictions of even large metropolitan areas.

[24] **SZTAKI @ ImageCLEF 2012 Photo Annotation.**

Authors: B. Daróczy, D. Siklósi and A.A. Benczúr.

Conference: Working Notes of the ImageCLEF 2011 Workshop at CLEF 2012 Conference, Rome, Italy

https://dms.sztaki.hu/sites/dms.sztaki.hu/files/file/2012/clef2012labs_submission_135.pdf

Abstract: Our team made second place with tight margin at ImageCLEF 2012 Photo Flickr. We used our open-source GMM/Fisher vector toolkit based on our research for Gaussian Mixture Modeling and Fisher Kernel based learning methods implemented on graphics coprocessors. The machine learning methods used in this paper are also applicable in Web classification as shown by our ongoing work.

[25] **Four degrees of separation**

Authors: L.Backstrom, P.Boldi, M.Rosa, J.Ugander, and S.Vigna.

Conference: ACM Web Science 2012: Conference Proceedings, June 2012, pages 45-54, ACM Press (2012); best paper award, highlighted by New York Times; (arXiv:1111.4570, 2012)

Abstract: Frigyes Karinthy, in his 1929 short story "Láancszemek" ("Chains") suggested that any two persons are distanced by at most six friendship links. (The exact wording of the story is slightly ambiguous: "He bet us that, using no more than five individuals, one of whom is a personal acquaintance, he could contact the selected individual […]". It is not completely clear whether the selected individual is part of the five, so this could actually allude to distance five or six in the language of graph theory, but the "six degrees of separation" phrase stuck after John Guare's 1990 eponymous play. Following Milgram's definition and Guare's interpretation, we will assume that “degrees of separation” is the same as “distance minus one”, where “distance” is the usual path length—the number of arcs in the path.) Stanley Milgram in his famous experiment challenged people to route postcards to a fixed recipient by passing them only through direct acquaintances. The average number of intermediaries on the path of the postcards lay between 4.4 and 5.7, depending on the sample of people chosen.

We report the results of the first world-scale social-network graph-distance computations, using the entire Facebook network of active users (≈721 million users, ≈69 billion friendship links). The average distance we observe is 4.74, corresponding to 3.74 intermediaries or “degrees of separation”, showing that the world is even smaller than we expected, and prompting the title of this paper. More generally, we study the distance distribution of Facebook and of some interesting geographic subgraphs, looking also at their evolution over time.

The networks we are able to explore are almost two orders of magnitude larger than those analysed in the previous literature. We report detailed statistical metadata showing that our measurements (which rely on probabilistic algorithms) are very accurate.

[26] **Four degrees of separation, really**

Authors: P.Boldi and S.Vigna.

Conference: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE, 2012: 1222-1227 (arXiv:1205.5509, 2012)

Abstract: We recently measured the average distance of users in the Facebook graph, spurring comments in the scientific community as well as in the general press (“Four Degrees of Separation”). A number of interesting criticisms have been made about the meaningfulness, methods and consequences of the experiment we performed. In this paper we want to discuss some methodological aspects that we deem important to underline in the form of answers to the questions we have read in newspapers, magazines, blogs, or heard from colleagues. We indulge in some reflections on the actual meaning of “average distance” and make a number of side observations showing that, yes, 3.74 “degrees of separation” are really few.

[27] **Arc-Community Detection via Triangular Random Walks**

Authors: P.Boldi, M.Rosa,.

Conference: LA-WEB 2012: 48-56 (2012)

Abstract: Community detection in social networks is a topic of central importance in modern graph mining, and the existence of overlapping communities has recently given rise to new interest in arc clustering. In this paper, we propose the notion of triangular random walk as a way to unveil arc-community structure in social graphs: a triangular walk is a random process that insists differently on arcs that close a triangle. We prove that triangular walks can be used effectively, by translating them into a standard weighted random walk on the line graph; our experiments show that the weights so defined are in fact very helpful in determining the similarity between arcs and yield high-quality clustering. Even if our technique gives a weighting scheme on the line graph and can be combined with any node-clustering method in the final phase, to make our approach more scalable we also propose an algorithm (ALP) that produces the clustering directly without the need to build the weighted line graph explicitly. Our experiments show that ALP, besides providing the largest accuracy, it is also the fastest and most scalable among all arc-clustering algorithms we are aware of.

[28] **Robustness of social and web graphs to node removal**

Authors: P.Boldi, M.Rosa, S.Vigna,

Journal: Social Network Analysis and Mining, Springer: 1-14 (2012)

Abstract: Given a social network, which of its nodes have a stronger impact in determining its structure? More precisely: which node-removal order has the greatest impact on the network structure? We approach this well-known problem for the first time in a setting that combines both web graphs and social networks. Our experiments are performed on datasets that are orders of magnitude larger than those appearing in the previous literature: this is possible thanks to some recently developed algorithms and software tools that approximate accurately the number of reachable pairs and the distribution of distances in large graphs. Our experiments highlight deep differences in the structure of social networks and web graphs, show significant limitations of previous experimental results; at the same time, they reveal clustering by label propagation as a new and very effective way of locating nodes that are important from a structural viewpoint.

[29] **Injecting Uncertainty in Graphs for Identity Obfuscation**

Authors: P.Boldi, F.Bonchi, A.Gionis, T.Tassa,

Journal: PVLDB 5(11): 1376-1387 (2012)

Abstract: Data collected nowadays by social-networking applications create fascinating opportunities for building novel services, as well as expanding our understanding about social structures and their dynamics. Unfortunately, publishing social-network graphs is considered an ill-advised practice due to privacy concerns. To alleviate this problem, several anonymization methods have been proposed, aiming at reducing the risk of a privacy breach on the published data, while still allowing to analyze them and draw relevant conclusions. In this paper we introduce a new anonymization approach that is based on injecting uncertainty in social graphs and publishing the resulting uncertain graphs. While existing approaches obfuscate graph data by adding or removing edges entirely, we propose using a finer-grained perturbation that adds or removes edges partially: this way we can achieve the same desired level of obfuscation with smaller changes in the data, thus maintaining higher utility. Our experiments on real-world networks confirm that at the same level of identity obfuscation our method provides higher usefulness than existing randomized methods that publish standard graphs.

[30] **Axioms for centrality**

Authors: P. Boldi, S. Vigna.

Journal: accepted for publication on Internet Mathematics. (arXiv:1308.2140, 2013)

Given a social network, which of its nodes are more central? This question has been asked many times in sociology, psychology and computer science, and a whole plethora of centrality measures (a.k.a. centrality indices, or rankings) were proposed to account for the importance of the nodes of a network. In this paper, we try to provide a mathematically sound survey of the most important classic centrality measures known from the literature and propose an axiomatic approach to establish whether they are actually doing what they have been designed for. Our axioms suggest some simple, basic properties that a centrality measure should exhibit.

Surprisingly, only a new simple measure based on distances, harmonic centrality, turns out to satisfy all axioms; essentially, harmonic centrality is a correction to Bavelas's classic closeness centrality designed to take unreachable nodes into account in a natural way.

As a sanity check, we examine in turn each measure under the lens of information retrieval, leveraging state-of-the-art knowledge in the discipline to measure the effectiveness of the various indices in locating web pages that are relevant to a query. While there are some examples of this comparisons in the literature, here for the first time we take into consideration centrality measures based on distances, such as closeness, in an information-retrieval setting. The results match closely the data we gathered using our axiomatic approach.

Our results suggest that centrality measures based on distances, which have been neglected in information retrieval in favour of spectral centrality measures in the last years, are actually of very high quality; moreover, harmonic centrality pops up as an excellent general-purpose centrality index for arbitrary directed graphs.

[31] **In-core computation of geometric centralities with HyperBall: A hundred billion nodes and beyond**

Authors: P. Boldi, S. Vigna.

Conference: To appear in the Proceedings of 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW 2013). (arXiv:1308.2144, 2013)

Abstract: Given a social network, which of its nodes are more central? This question has been asked many times in sociology, psychology and computer science, and a whole plethora of centrality measures (a.k.a. centrality indices, or rankings) were proposed to account for the importance of the nodes of a network. In this paper, we approach the problem of computing geometric centralities, such as closeness and harmonic centrality, on very large graphs; traditionally this task requires an all-pairs shortest-path computation in the exact case, or a number of breadth-first traversals for approximated computations, but these techniques yield very weak statistical guarantees on highly disconnected graphs. We rather assume that the graph is accessed in a semi-streaming fashion, that is, that adjacency lists are scanned almost sequentially, and that a very small amount of memory (in the order of a dozen bytes) per node is available in core memory. We leverage the newly discovered algorithms based on HyperLogLog counters, making it possible to approximate a number of geometric centralities at a very high speed and with high accuracy. While the application of similar algorithms for the approximation of closeness was attempted in the MapReduce framework, our exploitation of HyperLogLog counters reduces exponentially the memory footprint, paving the way for in-core processing of networks with a hundred billion nodes using “just” 2TiB of RAM. Moreover, the computations we describe are inherently parallelizable, and scale linearly with the number of available cores.

[32] **BUbiNG: Massive crawling for the masses**

Authors: P. Boldi, A. Marino, M.Santini, S. Vigna.

Confrerence: submitted for publication.

Abstract: Although web crawlers have been around for twenty years by now, there is virtually no freely available, open-source crawling software that guarantees high throughput, overcomes the limits of single-machine tools and at the same time scales linearly with the amount of resources available. This paper aims at filling this gap.

We describe BUbiNG, our next-generation web crawler built upon the authors' experience with UbiCrawler and on the last ten years of research on the topic. BUbiNG is an open-source Java fully distributed crawler (no central coordination), and single BUbiNG agents using sizeable hardware can crawl several thousands pages (per agent) per second respecting strict politeness constraints, both host- and IP-based. Unlike existing open-source distributed crawlers that rely on batch techniques (like MapReduce), BUbiNG job distribution is based on modern high-speed protocols so to achieve very high throughput.

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	4	4	8	8	3	3	2	2	1	1	1 / 1	1 / 1	20	20
Beneficiary P2 UTWE	6	8	4	5	2	2	2	3	1	2	0 / 0	1 / 1	16	21
Beneficiary P3 MTA_SZTAKI	2	2	2	2	2	2	10	10	7	7	0 / 0	1 / 1	24	24
Beneficiary P4 UMIL	4	4	0	0	6	6	0	0	9	9	0 / 0	1 / 1	20	20

[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