Small or medium-scale focused research project (STREP)

Full proposal

ICT FET Open Call
FP7-ICT-2011-C

New tools and Algorithms for DIrected NEtwork analysis

(NADINE)

Date of preparation: 21.02.2012

Version number: 3 (includes Technical Annex, Deliverables updates)

Name of short proposal this full proposal refers to: NADINE (No 288956)

Type of funding scheme: Small or medium-scale focused research project (STREP)

Work programme topics addressed: ICT-2011.9.1 FET Open

Name of the coordinating person: Dima Shepelyansky

List of participants:

Participant no. *	Participant organisation name	Participant short name	Country
1 (Coordinator)	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	CNRS	France
2	UNIVERSITEIT TWENTE	UTWE	Netherlands
3	MAGYAR TUDOMANYOS AKADEMIA SZAMITASTECH-NIKAI ES AUTOMATIZALASI KUTATO INTEZET	MTA_SZTAKI	Hungary
4	UNIVERSITA DEGLI STUDI DI MILANO	UMIL	Italy

Proposal Abstract:

On the scale of the past ten years, modern societies have developed enormous communication and social networks. Their classification and information retrieval becomes a formidable task for the society. Various search engines have been developed by private companies which are actively used by Internet users. Due to the recent enormous development of World Wide Web and communication networks, new tools and algorithms should be invented to characterize the properties of these networks on a more detailed and precise level. It is also highly important to have new tools to classify and rank enormous amount of network information in a way adapted to internal network structures and characteristics. The project will 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 will be 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. New tools and algorithms produced by the project will create fundamental basis for developers of new types of search and social media services, which will put Europe on leading positions in this important area dominated at present by other countries. NADINE tools will find applications in modern networks, including mobile communication networks which will play more and more important role in future. New characterization of complex networks will allow stakeholders to manage information extraction for social networks, communication and other networks in an efficient and rapid way. The project will create efficient voting systems in social networks that will pave the way for new types of democracy solutions in societies at a high communication level.

Table of Contents:

Proposal

Section 1: Scientific and/or technical quality, relevant to the topics addressed by the call

1.1 Targeted breakthrough and long-term vision

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

To handle such enormous databases, fundamental mathematical tools and algorithms related to centrality measures and network matrix properties should be developed. Indeed, the PageRank algorithm, which was initially at the basis of the development of the Google search engine [1,2] and is still key ingredient in network analysis, is directly linked to the mathematical properties of Perron-Frobenius operators and Markov chains [3]. Due to its mathematical foundation, this algorithm determines a ranking order of nodes that can be applied to various types of directed networks. However, the recent enormous development of WWW and communication networks requires creation of new tools and algorithms to characterize the properties of these networks on a more detailed and precise level. For example, such networks contain weakly coupled or secret communities which can correspond to very small values of the PageRank and are hard to detect [2]. It is therefore highly important to have new tools to classify and rank enormous amount of network information in a way adapted to internal network structures and characteristics.

The present project will develop new tools and algorithms to facilitate classification and information retrieval from large networks recently created by human activity. It will develop new types of generalized centrality measures which will be sensitive to internal network structures and will lead to new efficient methods for node classification and information retrieval. Newly developed Monte Carlo methods will allow extracting these centrality measures at low computational cost for enormously large network sizes. The Google matrix formed by the links of the network will be analyzed by analytical tools of Random Matrix Theory [4] and quantum chaos and by efficient numerical methods for large matrix diagonalization including the Arnoldi method [5]. The combination of CheiRank with PageRank leads to two-dimensional ranking by communicativity and popularity of nodes, as proposed recently by partners. These new methods will be disseminated to create qualitatively new types of search and social media solutions and will be applied to various types of networks, including the WWW, social networks, citation networks, procedure call networks of software architecture, Wikipedia hyperlinks, phone call communication networks, workflows in economy and business process management. We will devise means to extract new information from networks including communities and other internal structures. A special emphasis will be put on social networks, for which specific tools will be developed, in particular related to the important problem of voting systems.

The new tools and algorithms created by the project will produce **broad scientific and technological impact** on **classification and information retrieval on modern directed networks**, and give n**ew recipes for their skilful design and development**. As concerns the

European dimension: New search ranking and social media analysis technologies will be proposed on the basis of new algorithms. This will help to create a European answer on the current situation when the market is dominated by search engines from USA, China, and Russia. The obtained research results will be made public, but the explanation of their importance and the know-how for their concrete application and efficiency will be provided in priority to European firms and Institutions. The project is based on interdisciplinary expertise of partners in mathematics, physics and computer science with the cross-fertilization of different fields of science bringing qualitatively new solutions.

Breakthrough: New tools and algorithms produced by the project will create fundamental basis for industrial development of new types of search and network analysis technologies that will put Europe on leading positions in this important area dominated at present by other countries. New characterization of directed complex networks will allow applications to manage information extraction for social networks, communication, and other networks in an efficient, rapid and dynamical way. These tools will be actively used for network applications and beyond, for instance, for mobile communication networks that will play more and more important role in future. The project will create efficient voting systems in social networks that will pave the way for new types of democracy solutions in a society at a high communication level.

1.2 Novelty and foundational character

All information about a directed network is contained in the Google matrix $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E}/N$, where **S** is constructed by normalizing to unity all columns of the adjacency matrix, and replacing columns with zero elements by 1/N, N being the network size, and **E** has unity as all entries [2]. The damping parameter α describes the probability to jump to any node for a random surfer. The matrix G belongs to the class of Perron-Frobenius operators appearing in dynamical systems and Markov chains [3]. It has a largest eigenvalue $\lambda=1$ with associated real nonnegative eigenvector, called the PageRank vector ρ. This vector plays an important role in the ranking of WWW nodes in commercial search engines. The statistical properties of the PageRank have been investigated by many groups and it was found that its distribution follows in general a universal algebraic decay with exponent $\nu = 2.1$ [6], also ρ is established to be on average proportional to a number of ingoing links [7]. Such type of algebraic dependence appears in scale-free type networks which have been actively investigated during the last decade [8, 9, 10]. A number of statistical models have been proposed for analysis of properties of such networks. One of the most known is the Albert-Barabasi model [8]. It naturally generates scale-free features but the recent studies [11] of the spectral properties of its corresponding Google matrix showed that it has a large gap in the spectrum of λ between unity and other eigenvalues. In contrast typical WWW networks, e.g. British university networks, have no such gap. Hence, new type of random matrix models should be developed to give correct description of eigenspectrum and eigenstates of real networks [11].

Among eigenvectors of G the PageRank is the most important one, giving a steady state probability distribution over the network. The PageRank gives at the top the most known and popular nodes. However, an example of the Linux procedure call network studied in [12] shows that in this case the PageRank puts at the top certain procedures which are not very important from the software view point (e.g. printk). As a result it was proposed to use an additional ranking vector. Thus, the CheiRank ρ^* has been proposed by partners. It is a PageRank vector of the Google matrix $\mathbf{G}^*$ built from the adjacency matrix with inverted link directions of the initial network [12, 13, 14]. Thus ρ^* is on average proportional to a number of outgoing links. While ρ determines node popularity, ρ^* highlights communicative property of nodes.

Since each node belongs both to ρ and ρ∗ we obtain a new two-

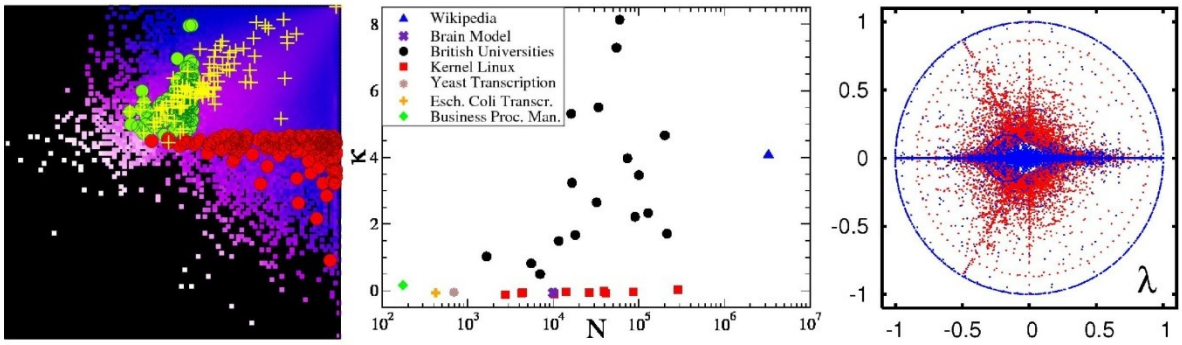


Fig.1: three illustrations of the proposed new network based ranking methods.

dimensional ranking on 2D plane of ranks K, K* corresponding to ρ, ρ∗ [13,14]. Density distribution of N=328257 Wikipedia English articles, which form a directed citation hyperlink network, is shown in log-log scale in 2D plane ($0 < \ln K, \ln K^* < \ln N$) taken from [13] (left Fig., $\alpha = 0.85$, white/blue for large/low density, black for 0). Here red/green points show top 100 personalities according to PageRank/CheiRank, pluses show Hart's book ranking. Among personalities PageRank selects broadly known ones with 58% in politics, 15% in science, while CheiRank gives here 15% and 52% finding composers, architects, botanists and astronomers who are not much known but who e.g. discovered a lot of Australian butterflies (George Lyell) or asteroids (Nikolai Chernykh). Thus 2DRanking retrieves information in a new more rich way. Our initial study shows that various networks have small (e.g. Linux kernel software and gene networks) or large (UK universities and Wikipedia networks) κ correlator values between CheiRank and PageRank (middle Fig. Taken from [14]) and hence these directed networks are qualitatively different.

The eigenvalues spectrum of G matrix of Cambridge University network (N = 200823, α=1, right Fig., taken from [15]), obtained by the Arnoldi method [5], has isolated communities with λ=1 (subset size about 40000, blue points) and remaining single component with λ<1 (red points). The spectral density of **G** for such networks shows well-pronounced patterns and eigenstates of **G** display prominent structures, in particular communities are well visible even at the PageRank tail. This matrix analysis will allow extracting important information hidden inside network. Even though the PageRank decay is rather universal, real networks show complex internal structure which should be extracted by new tools and methods. These complex and rich structures require adapted centrality measures in order to rank and classify network information in a most efficient and relevant way. Recent results obtained by the team demonstrate universal features of PageRank emergence at small damping factors related to isolated community structures [15].

Creation of new tools gains enormous importance with the recent development of new types of networks including social networks or phone call communication networks [2, 8, 9, 10]. These methods will also find applications in new fields of science where network analysis has been applied only recently, such as software architecture [2, 12], dynamical systems and Ulam networks [3], workflows in economy and world trade [16], gene networks [14, 17] or new problems emerging with the conception of voting systems on social networks [18]. The new tools will create a basis for industrial development of search engines of qualitatively new type.

In the following we list in detail the state of the art of each WP and the progress to be made by NADINE.

WP1: CheiRank versus PageRank, centrality measures and network structure

Numerous studies prove that the second-order network structural characteristics, such as assortativity, clustering and correlations, are crucial in network design and applications. However, current network models, e.g. the famous Preferential Attachment rule [9], are too generic to capture these important traits which are genuinely network-specific (see e.g. [11]). Centrality measures, due to their global nature, contain the needed network-specific information. We will develop and analyze new advanced probabilistic models that establish adequate mathematical relations between centrality measures and network structure. The consortium, being the first to mathematically explain the well-known power law behaviour of PageRank values [19], has strong expertise to meet this new challenge.

The new concept of CheiRank versus PageRank with 2DRanking [12-14] (left Fig.) will be analyzed in detail giving more rich characterization of nodes communicativity and popularity. We use also 2DRank which combines these two properties to generate a 1d-ranking. For Wikipedia (left Fig.) we have top articles: 1.United States, 2. United Kingdom, 3. France (PageRank); 1. India, 2. Singapore, 3. Pakistan (2DRank); 1. Portal: Contents/Outline-of knowledge/Geography-and-places, 2. List-of-state-leaders-by-year, 3. Portal: Contents/ Index/ Geography-and-places (CheiRank). Among Dow-Jones companies at Wikipedia we have at top Cisco (CheiRank) and Microsoft (PageRank) that shows again difference between communication and "popularity". For networks with small correlator κ CheiRank gives at top more significant nodes compared to PageRank (e.g. for procedure call Linux software network and business process management). We note that CheiRank and PageRank naturally appear for the world trade network of international trade, where they and linked with export and import flows for a given country respectively [16]. The Google matrix built from United Nations COMTRADE database allows to rank all countries in a democratic way independently of their richness [16]. 2DRanking has certain features of HITS with its hubs and authorities but it gives a global query independent ranking. The parallels between HITS and 2DRanking will be investigated. The fundamental properties of 2DRanking will be studied in WP1 in combination with various centrality measures.

Exactly as the mean value cannot be used to describe power law distributions, the common correlation measures miss important information on network dependencies due to averaging and accumulating the results in just one number. The consortium has recently developed novel dependency measures based on the state-of-the-art techniques in extreme value theory, that characterize the correlations for most important nodes by a probability-like measure on a finite interval [20, 21, 22]. This enables the exactly right level of mathematical accuracy, both insightful and analytically tractable, for studying the 2Dranking. In addition, Markov Decision Processes will be applied for optimal modelling of modern directed networks.

We also develop Monte Carlo methods for efficient estimation of PageRank and CheiRank. Preliminary results indicate that to detect few top nodes with large PageRank values, we can use less random walk runs than the number of the nodes [22]. We plan to investigate the convergence of the list of top nodes in terms of ranking rather than in terms of PageRank and CheiRank probability values, since ranking converges much quicker than the rank values. We expect that Monte Carlo approach can help us to develop very efficient methods for detecting top nodes with respect to various centrality measures, and 2DRanking opens qualitatively new opportunities for creation of search engines of new types.

WP2: Network analysis through Google matrix eigenspectrum and eigenstates

The spectrum of the Google matrix [11, 15] also contains characteristic patterns clearly seen in the right Fig. In addition, eigenstates other than the PageRank also display peaks and correlations visible even at the tail of the PageRank. The consortium has recently adapted the Arnoldi method in order to obtain numerically eigenvalues with largest $|\lambda|$ and their associated eigenvectors, reaching network sizes N of several millions [15]. The precise analysis of these eigenstates will allow extracting communities and network structures in a very efficient way. We will systematically relate the characteristics of the spectrum and eigenstates to the internal structure of the network, enabling generic characteristics of networks to be determined by this matrix analysis. Significant degeneracies are present in the spectrum, being related to specific properties of the networks (right Fig.). Our preliminary data also show that for certain networks, e.g. Linux kernel software network, the spectrum is characterized by a fractal Weyl law recently established for quantum chaotic scattering and the Perron-Frobenius operators of chaotic dynamical systems [23]. This law determines the number of modes with large $|\lambda|$ in the matrix **G**, depending on the fractal dimension of the network, and thus identifies the most important eigenstates. We find for Linux network the fractal dimension d=1.3. In addition, the studies of network models and Ulam networks has shown that under certain conditions the PageRank can become completely delocalized [24]. Such delocalization has certain similarities with the Anderson transition in disordered solids [4, 25, 26] and its appearance would produce a drastic impact on the performance of the PageRank algorithm, since the PageRank becomes flat being dominated by small random fluctuations. The whole world would go blind the day such delocalization occurs... At the moment the PageRanks of many networks including WWW are located in the localized phase but the conditions of localization and its properties should be investigated in detail in order to meet the danger of delocalization transition due to future network growth.

WP3: Applications to voting systems in social networks

We will elucidate the relation between different centrality measures for different types of social networks. We will study centrality measures of the first category (measuring how a node is well connected to the other nodes, instances are node degree, closeness centrality, PageRank) and second category (measuring to what extent a node helps communication flow inside the network, instances are betweenness centrality, random walk betweenness centrality, second order centrality). The study of these different measures applied to various types of social networks will make it possible specifying which centrality measure is better suited for which type of online social network analysis problem. We also hope to gain insights on the behaviour of the spectral indices studied by the other work packages by comparing them with indices which are more geometrically inspired, and on which more intuition has been built by the sociologists in the last 50 years.

We remark that such an evaluation is missing from the literature for several reasons. First of all, many of these indices are difficult or impossible to compute exactly on large datasets (e.g., betweenness). We intend to approach this problem using massive parallel computations and new (possibly approximate) algorithms, using also Monte-Carlo techniques developed by the other work packages. Second, while when discussing toy examples made by a few nodes geometrical and sociological intuitions can help in judging the correctness or the properties of the index under study, large networks provide no obvious clue, and an *external ground truth* is necessary to assess the index. Missing an external ground truth, another interesting approach is to compare the correlation between an index to be studied and existing indices whose properties are well known. In a sense, we leverage what we know about an index to infer

something about another. The high scalability of some correlation index (e.g., Kendall's tau) makes this approach feasible.

In parallel, we will focus on the important question of voting systems on such social networks [18, 27-29]. Voting systems are systems that allow one to collapse the preferences of a population into a single vote that, in a sense, should reflect the desires of all individuals, or the majority of them. Although largely studied in classical sociology textbooks, *direct democracy* voting in a large social network turns out not to be a good idea. *Spectral voting* is similar to what is sometimes called *liquid democracy* (or proxy voting), but with an attenuation factor introduced to reduce transitivity and to avoid paradoxes. Given a social network, every individual chooses one of its direct acquaintances and delegates its voting power to him/her. The delegation is weakened by the attenuation factor that may be seen as a sort of friction in the transitivity of the vote. In the end, one obtains a *voting graph* with constant out-degree (one), and computes Katz's path-based index on the voting graph. 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.

This fact suggests that, as it happens in other ranking systems (e.g. HITS), other eigenvectors (beyond the dominant one) may provide further information about the structure of the network, in particular helping to find emergent opinions. The expertise of the consortium will allow studying the entire spectrum and the associated eigenvector spaces to explore these issues. Second, when voting is not complete, the presence of censored data can be dealt with by modelling missing votes using random variables (e.g., a uniform choice of neighbours). In this case, probability theory can be used to provide results and even a *centrality index* based on a total random vote. We will investigate such indexes using the consortium expertise in probability theory.

Another idea we intend to explore is the usage of other centrality indices (*in lieu* of Katz's) to structure the delegation process. As long as a voting graph is defined, every centrality index provides a different way of spreading the voting power of delegating users. While Katz's index is endowed with a very simple interpretation (giving away one's vote with an attenuation factor), other indices provide different combination of the voting power. Closeness centrality, for instance, induces a voting scheme which in essence computes the arithmetic mean of the distances of the votes from the final recipient taking the decision.

We will thus elucidate the relation between different centrality measures for different types of social networks. We will study centrality measures of the first category (measuring how a node is well connected to the other nodes, instances are node degree, closeness centrality, PageRank) and second category (measuring to what extent a node helps communication flow inside the network, instances are betweenness centrality, random walk betweenness centrality, second order centrality). The study of these different measures applied to various types of social networks will make it possible specifying which centrality measure is better suited for which type of online social network analysis problem. We also hope to gain insights on the behaviour of spectral indices studied by the other work packages by comparing them with indices which are more geometrically inspired, and on which more intuition has been built by the sociologists in the last 50 years.

We remark that such an evaluation is missing from the literature for several reasons. First of all, many of these indices are difficult or impossible to compute exactly on large datasets (e.g., betweenness). We intend to approach this problem using massive parallel computations and new (possibly approximate) algorithms, using also Monte-Carlo techniques developed by the other work packages. Second, while when discussing toy examples made by a few nodes geometrical and sociological intuitions can help in judging the correctness or the properties of the index under study, large networks provide no obvious clue, and an *external ground truth* is necessary to assess the index. Missing an external ground truth, another interesting approach is to compare the correlation between an index to be studied and existing indices whose properties are well known. In a sense, we leverage what we know about an index to infer something about another. The high scalability of some correlation index (e.g.,

Kendall's tau) makes this approach feasible.

Finally, we intend to develop a Facebook application which will be used to perform experiments with voluntary participation from Facebook users. The idea is to describe a decision and let people either state their preference or choose some friend that is better suited to take the decision. More than in the decision itself, we are interested in gathering the delegation graphs that will be generated, to study their properties. If we are able to make the process sufficiently viral, this will provide us with interesting non-toy datasets. Partner 4 has a large data base (a few hudnred millions nodes) provided by Facebook which can be used for these studies (cf. Arxiv:1111.4570).

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

The new tools and algorithms developed in WP1-WP3 will provide new insights to a large variety of important networks, using the databases provided by WP5. These networks will include WWW, social networks, phone call communication networks, academic citation networks, procedure call networks of Open Source software (including Linux kernel), Wikipedia hyperlink network, and other networks appearing in various areas of science such as biological neuronal networks, protein networks, management networks, linguistic networks. Scalability is a key issue in handling real networks. We address graph partitioning for efficient distributed processing and consider how a variety of graph properties from WP 1-3 can be dynamically updated as new data arises. Of particular importance is fingerprinting for similarity (e.g. SimRank [30]) and ranking (e.g. personalized PageRank and CheiRank). Several graph algorithms can be parallelized by splitting the graph into pieces and distributing the pieces to different servers where they fit into internal memory. Such an architecture can serve shortest path queries, local network flows for community core queries ("give me the nearest community"), similarity search and more. We devise new ways to formulate and implement large-scale graph and matrix processing algorithms using improvements of Hadoop, Open Message Passing Interface, graph processing frameworks in the Bulk Synchronous Parallel model similar to Google's Pregel.

Due to the large and ever increasing financial gains resulting from high search engine ratings, it is no wonder that a significant amount of human and machine resources are devoted to artificially inflating the rankings of certain web pages. Amit Singhal, principal scientist of Google Inc. estimated that the search engine spam industry had a revenue potential of $4.5 billion in year 2004 if they had been able to completely fool all search engines on all commercially viable queries. As a result, web data collections suffer from ***quality deterioration*** in view of the fact that, under different measurement and estimates, roughly 10% of the Web sites and 20% of the individual HTML pages constitute spam. The baseline methods for Web spam filtering are summarized in [31] while the Web Spam Challenges [32] provide forum for researchers to compare their technologies. The Consortium includes pioneers of this area: [33] is early and highly cited result on Web spam detection. The Consortium includes an institution that made the first progress in the LiWA—Living Web Archives FP7 project to aid Internet archives in collaborate in Web spam filtering [34, 35, 36, 37]. Our present Web classifiers however do not scale up to Petabytes and even filtering a Terabyte scale archive may require a parallel environment. As a combination of distributed indexing and matrix processing procedures developed in the NADINE project, we will be able to classify and filter Web hosts at the Peta-scale.

WP5: Database development of real-world networks: WP5 will provide data collection for the many types of real networks which will be investigated in other Work Packages. This will be performed using the expertise of the consortium which already provided large datasets

for various networks. The collected data will be open for public access.

Datasets are essential in the evaluation of centrality measures, ranking algorithms, and for experimenting with large-scale data analysis. In the context of this project, a dataset is a *snapshot* of a part of the web graph, or of a social network. A snapshot is a copy that should exhibit a minimum of temporal coherence, and thus should be gathered in a very short time (say, no more than few weeks).

In some cases, it might be useful to gather several snapshots on a regular basis: this is very useful to explore the way in which measures evolve in time, and the way in which communities appear and dissolve. One of the few publicly available examples is the temporally labelled snapshot of the UK domain gathered by the LAW [38] for the DELIS EU FP6 project. Other organizations gathering temporal snapshots are the Stanford WebBase and the Internet Archive. However, the Internet Archive does not distribute publicly its data, and the way in which datasets are provided by the Stanford WebBase is not suited to large data transfers. Currently, the largest, publicly available snapshot of the web graph has around 100 million nodes. There are two notable exceptions, which however have serious problems.

One is the Altavista 2002 snapshot distributed by Yahoo! in its WebScope program. This dataset is unusable from scientific viewpoint as its features are completely different by those of web snapshots usually gathered. The so-called "giant component", which collects the largest set of mutually reachable pages, and it is experimentally known to contain at least 50% of the pages of a snapshot, is less than 4%. A large proportion of nodes have no outlinks, which is probably due to the inclusion of discovered but not yet visited nodes to the snapshot. No information is available on the crawling algorithm, or on the seed.

The other is the ClueWeb09 snapshot developed for the TREC conference. The dataset is excellent from the information-retrieval tasks it has been designed for, but from the perspective of studying networks it is problematic. Besides being quite expensive (as it has to be shipped on several large hard disks), the crawl has been arranged so to contain the kinds of web pages that commercial search engines such as Google, Yahoo and Live Search would contain and rank highly some query. This suggests some kind of content-based optimization that does not agree well with the intent of obtaining a representative sample of the network.

For social networks the situation is more complex and varied. Relatively small databases (such as scientific co-authorship, or co-starring in movies) are easily downloaded from the net and need just some massaging. Larger networks (e.g., Facebook) must be crawled in a way that is similar to web graphs, but developing ad-hoc algorithms, as the interface to access links between entities (e.g., friendship) is not uniform. Unfortunately, strictly limited terms of use make the goal of crawling large social networks very difficult to attain.

Internet scale data management is related to the objectives of NADINE in two respects. First, highly important data items are spread across multiple data collections that need to be identified and accumulated upon which network analysis tasks will be performed. Second, of particular interest are methods for efficient distributed data analysis. No infrastructure has yet been developed for distributed document information sharing with tasks including duplicate detection or cross-linkage analysis. Large scale duplicate detection methods are based on fingerprinting but distributed operation is not considered yet. Distributed location servers introduced for caching documents [39] but new methods are needed if the location of the document is defined by external constraints and mass copying to new locations is infeasible. Several methods for large graph approximations have been devised including quite a few of Web relevance originating from within the proposed Consortium [30, 40]. Distributed architectures are used for various tasks [41, 42].

Web content analytics is increasingly gaining attention due to its value in gathering business intelligence. State-of-the-Art research takes a multidimensional view on text databases, but remain at the level of terms [43]. On the phrase-level recent studies have investigated media sites and blogs based on relatively larger computations over time by a so-called meme-tracking approach [44].

Cross-data analytics is still in its infancy. Main efforts so far mostly concentrate on standards for (meta-) data interoperability (such as [45]) instead of cross-data analytics. In addition, the most extensive available Web research collections only provide sparse disconnected snapshots (temporal as well as content wise), which are far less comprehensive than collections of NADINE consortium members and supporting partners. While there has been sporadic papers on the temporal analysis of Web content for retrieval and classification [46] and change frequency estimation [47], there has been a recent burst of papers in this area considered hot topic in the WWW, SIGIR and WSDM community [34, 48].

The current best research collections such as [38] occupy huge space and (except for small portions such as the Web graph) they cannot be easily distributed. Indeed, it took two weeks to extract this data locally and then remote copy a 500GB portion for further research within the Consortium. When searching or classifying distributed content, we face the difficulty that the coverage, depth, breadth and several statistical properties may be very different for different portions of data gathered for different purposes by different institutions and strategies. Classifying heterogeneous Web data is a hard task in [35].

NADINE backed by the available huge data collections will allow unique studies of Web data to the whole scientific community, which have not been possible that comprehensibly before. Hence, NADINE will support discovery, modelling and (even) prediction of novel patterns of occurrence on the Web and social media.

REFERENCES:

[1] S.Brin and L.Page, *The anatomy of a large-scale hypertextual Web search engine,* Comp. Netw. ISDN Syst. 30, 107 (1998).

[2] A.M.Langville and C.D.Meyer, *Google's PageRank and Beyond: The Science of Search Engine Ranking* (Princeton University Press 2006).

[3] M.Brin and G. Stuck, *Introduction to dynamical systems,* (Cambridge University Press, 2002).

[4] Thomas Guhr, Axel Müller–Groeling and Hans A. Weidenmüller, *Random-matrix theories in quantum physics: common concepts* Phys. Rep. 299, 189 (1998).

[5] G. W. Stewart, *Matrix Algorithms Volume II: Eigensystems* (SIAM, 2001).

[6] D.Donato, L.Laura, S.Leonardi and S.Millozzi, *Large scale properties of the Webgraph,* Eur. Phys. J. B 38, 239 (2004); G.Pandurangan, P.Raghavan and E.Upfal, *Using PageRank to Characterize Web Structure,* Internet Math. 3, 1 (2005).

[7] N.Litvak, W.R.W.Scheinhardt, and Y.Volkovich, *In-Degree and PageRank of Web pages: Why do they follow similar power laws?* Internet Math. 4, 175 (2007).

[8] R.Albert and A.-L.Barabasi, *Statistical mechanics of complex networks,* Rev. Mod.

Phys. 74, 47 (2002).

[9] S.N.Dorogovtsev and J.F.F.Mendes, *Evolution of Networks*, (Oxford University Press, 2003).

[10] S.Fortunato, *Community detection in graphs*, Phys. Rep. 486, 75 (2010).

[11] O.Giraud, B.Georgeot and D.L.Shepelyansky, *Delocalization transition for the Google matrix*, Phys. Rev. E 80, 026107 (2009); ibid, *Spectral properties of the Google matrix of the World Wide Web and other directed networks* 81, 056109 (2010).

[12] A.D.Chepelianskii, *Towards physical laws for software architecture*, preprint arxiv:1003.5455 (2010)

[13] A.O.Zhirov, O.V.Zhirov and D.L.Shepelyansky, *Two-dimensional ranking of Wikipedia articles*, EPJB 77, 523 (2010).

[14] L.Ermann, A.D.Chepelianskii and D.L.Shepelyansky, *Towards two-dimensional search engines*, preprint arxiv:1106.6215[cs.IR]. (2011)

[15] K.M.Frahm, B.Georgeot and D.L.Shepelyansky, *Universal emergence of PageRank*, preprint arxiv:1105.1062[cs.IR] (2011)

[16] L.Ermann and D.L.Shepelyansky, *Towards Google matrix of the world trade network*, preprint arxiv:1103.5027 (2011).

[17] R.Milo, S.Shen-Orr, S.Itzkovitz, N.Kashtan, D.Chklovskii and U.Alon, Science 298, 824 (2002).

[18] M.Balinski and R.Lakari, *A theory of measuring, electing and ranking*, PNAS 104, 8720 (2007)

[19] N.Litvak, W.R.W. Scheinhardt and Y.Volkovic, *In-Degree and PageRank: Why Do They Follow Similar Power Laws?* Intenet Math. 4, 175 (2007)

[20] Y.Volkovich, N. Litvak and D.Donato, *Determining Factors Behind the PageRank Log-Log Plot*, Proceeding of WAW2007, LNCS 4863, 108 (2007)

[21] Y.Volkovich, N.Litvak. and B.Zwart, *Measuring extremal dependencies in Web graphs*, Proc. 17th Int. Conf. WWW 2008 1113 (2008).

[22] Y.Volkovich, N. Litvak and B.Zwart, *Extremal Dependencies and Rank Correlations in Power Law Networks*, Proc. First Int. Conf. Complex Sciences: Theory and Applications (Complex 2009), Shanghai, China, p.16 (2009) Springer.

[23] L.Ermann, A.D.Chepelianskii and D.L.Shepelyansky, *Fractal Weyl law for Linux Kernel Architecture*, Eur. Phys. J. B 79, 115 (2011)

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

[25] F.Evers and A.Mirlin, *Anderson transitions*, Rev. Mod. Phys. 80, 1355 (2008)

[26] N.Perra, V.Zlatic, A.Chessa, C.Conti, D.Donato and G.Caldarelli, *PageRank equation and localization in the WWW*, Eur. Phys. Lett.88, 48002 (2009)

[27] P.Boldi, F.Bonchi, C.Castillo, and S.Vigna. *Voting in social networks*. In Proc. of ACM 18th Conference on Information and Knowledge Management (CIKM), Napa Valley,

CA, USA, 2009. ACM Press.

[28] P.Boldi, M.Rosa, M.Santini, and S.Vigna. *Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks*. In Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors, Proceedings of the 20th international conference on World Wide Web, pages 587–596. ACM, 2011.

[29] P.Boldi, M.Santini, and S.Vigna. *Permuting web and social graph*s. Internet Math. 6, 257 (2010).

[30] D. Fogaras and B. Racz. *Scaling link-based similarity search*, in Proc. 14th Int. Conf. World Wide Web, p. 641, NY, ACM 2005.

[31] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. *A reference collection for web spam*. SIGIR Forum, 40(2):11–24, December 2006.

[32] C. Castillo, K. Chellapilla, and L. Denoyer. *Web spam challenge* 2008. In Proc. 4th Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2008.

[33] A.A. Benczúr, K. Csalogány, T. Sarlós, M. Uher: *SpamRank -- Fully Automatic Link Spam Detection*. AIRWeb'05 in conjunction with WWW 2005

[34] M. Erdélyi, A. A. Benczúr, J. Masanés, and D. Siklósi. *Web spam filtering in internet archives*. In AIRWeb ’09: Proc. 5th Int. Workshop on Adversarial information retrieval on the web, 2009.

[35] A. A. Benczúr, M. Erdélyi, J. Masanés, and D. Siklósi. *Web spam challenge proposal for filtering in archives*. In AIRWeb ’09: Proc. of the 5th Int. Workshop on Adversarial Information Retrieval on the Web. ACM Press, 2009.

[36] M. Erdélyi and A.A. Benczúr. *Temporal analysis for web spam detection: An overview.* In Proc. TWAW in conjunction with WWW 2011. CEUR Workshop Proceedings.

[37] M. Erdélyi, A. Garzó, and A.A. Benczúr. *Web spam classification: a few features worth more*. In Proc. WebQuality 2011 in conjunction with WWW 2011.

[38] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. *A large time-aware graph.* SIGIR Forum, 42(2):33−38, 2008.

[39] D.Karger, E.Lehman, T.Leighton, M.Levine, D.Lewin, and R.Panigrahy, *Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the world wide web*, in ACM Symposium on Theory of Computing, May 1997, pp. 654-663

[40] T. Sarlós, A.A. Benczúr, K. Csalogány, D. Fogaras, B. Rácz: *To Randomize or Not To Randomize: Space Optimal Summaries for Hyperlink Analysis.* In Proc. WWW2006

[41] C. Sidló, A. Garzó, A. Molnár, A.A. Benczúr, *Infrastructures and Bound for Distributed Entity Resolution*, in Proc. QDB in conj. VLDB 2011.

[42] András Garzó, Dávid Nemeskey, Róbert Pethes, Dávid Siklósi, András A. Benczúr, SZTAKI @ TREC 2010, in TREC 2010

[43] R. Fagin, P. Kolaitis, R. Kumar, J. Novak, D. Sivakumar, and A. Tomkins. *Efficient Implementation of Large-scale Multi-structural Databases.* In VLDB, 2005.

[44] Leskovec, J., Backstrom, L., and Kleinberg, J. 2009. *Meme-tracking and the dynamics*

of the news cycle. In Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (Paris, France, June 28 - July 01, 2009). KDD 09. ACM, New York, NY, 497-506

[45] ISO, International Organisation for Standardisation: MPEG-21: Overview. http://www.-chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm, 2004.

[46] Z.Bar-Yossef, A. Z.Broder, R.Kumar, and A.Tomkins. *Sic transit gloria telae: Towards an understanding of the Web's decay.* In Proc. 13th WWW p.328. ACM Press (2004).

[47] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener, *A large-scale study of the evolution of web pages*, Software- Practice and Experience 34, 213 (2004).

[48] P.Bordino, P.Boldi, D.Donato, M.Santini, and S.Vigna. *Temporal evolution of the UK web.* In Workshop on Analysis of Dynamic Networks (ICDM-ADN'08), 2008.

[49] P.Chen, H.Xie, S.Maslov and S.Redner, *Finding scientific gems with Google's PageRank algorithm*, J.Informet. 1, 8 (2007)

[50] L.Li, D.L.Alderson, J.C.Doyle and W.Willinger, Internet Math.2, 431 (2005)

[51] J.G.Foster, D.V.Foster, P.Grassberger and M.Paczuski, *Edge direction and the structure of networks*, PNAS 107, 10815 (2010)

[52] Y.Volkovich and N.Litvak, *Asymptotic analysis for personalized Web search*, Adv. Appl. Prob. 42, 577 (2010)

[53] P.R.Jelenkovic and M.Olvera-Cravioto, I*nformation ranking and power laws on trees*, Adv. Appl. Prob. 42, 1057 (2010)

[54] R. van der Hofstad, G.Hooghiemstra and P. van Mieghem, *Distances in random graphs with finite variance degrees,* Random Structures Algorithms 27, 76 (2005)

[55] R. van der Hofstad, Random graphs and complex networks, http://www. win. tue. nl/rhofstad/NotesRGCN.pdf (2010)

[56] K.Avrachenkov, N.Litvak, D.Nemirovsky and N.Osipova, Monte Carlo methods in PageRank computation: when one iteration is sufficient, SIAM J. Numer. Anal. 45, 890 (2007)

[57] B.Bahmani, K.Chakrabarti and D.Xin, *Fast personalized PageRank on MapReduce*, Proc. 2011 international conference on Management of data, p.973 ACM (2011)

[58] K.Avrachenkov, N.Litvak, D.Nemirovsky, E.Smirnova and M.Sokol, *Quick Detection of Top-k Personalized PageRank Lists*, Algorithms and Models for the Web Graph, p.50 Springer (2011)

[59] L.Ermann and D.L.Shepelyansky, *Google matrix and Ulam networks of intermittency maps*, Phys. Rev. E 81, 036221 (2010); *Ulam methos and fractal Weyl law for Perron-Frobenius operators*, Eur. Phys. J. B 75, 299 (2010)

[60] S.Nonnenmacher, J.Sjoestrand and M.Zworski, *From open quantum systems to open quantum maps*, Comm. Math. Phys. 304, 1 (2011)

[61] P.Boldi, F.Bonchi, C.Castillo, and S.Vigna. *Viscous democracy for social networks.* Commun. ACM, 54:129−137, June 2011.

[62] P.Boldi, M.Rosa, and S.Vigna. *HyperANF: Approximating the neighbourhood function of very large graphs on a budget*. In Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, Eds., Proc. 20th Int. Conf. on WWW, p625 ACM (2011).

[63] M. Kurucz, A.A. Benczúr, A. Pereszlényi, Large-scale principal component analysis on LiveJournal friends network. Proceedings of SNAKDD 2008.

[64] B. Georgeot and O. Giraud, *The game of go as a complex network*, preprint arXiv:1105.2470

[65] Paolo Boldi, Bruno Codenotti, Massimo Santini, and Sebastiano Vigna. UbiCrawler: A scalable fully distributed web crawler. Software: Practice & Experience, 34(8):711−726, 2004.

[66] Paolo Boldi and Sebastiano Vigna. The WebGraph framework I: Compression techniques. In Proc. 13th International World Wide Web Conference (WWW 2004), pp.595−601, Manhattan, USA, 2004. ACM Press.

[67] J Göbölös-Szabó, Z. Fekete, et al., Outstanding Integration of Visual and Computational methods. in IEEE VAST 2011 Symposium, part of VisWeek 2011.

1.3 S/T methodology

1.3.1 Overall strategy and general description

To achieve the goals of the project, the Consortium is built up out of four research groups with leading and complementary expertise in mathematics, computer science, software engineering, complex networks and theoretical physics. The duration of the project is 36 months with an annual workshop which allows enhancing collaborations and exchange of ideas within the Consortium and including other leading groups in these scientific areas.

The project is composed of five different interrelated scientific workpackages (**WP**s). **WP1** will focus on various centrality measures and propose new ways of efficient ranking of information on directed networks. **WP2** will use the Google matrix eigenspectrum and eigenstates for network analysis, community detection and characterization of fractal properties of networks. The theoretical advances gained in **WP1** and **WP2** will be applied to the important problem of voting systems in social networks in **WP3**, and enable to create new tools and algorithms to obtain new features and characteristics of real-world network structures (**WP4**). **WP5** will provide data collection for various types of modern networks, which will be treated by the methods developed in the other WPs. Special emphasis will be put on reliable collection of data from actual networks of petasize including spam filtering.

1.3.2 Timing of work packages and their components

The NADINE project is intended to last 36 months. We envisage an increasing level of integration of all workpackages as a function of time. Due to the theoretical and open nature of this proposal, the detailed findings corresponding to a deliverable in one workpackage should not severely influence the success of the deliverables in other workpackages but only the specific type and direction of following research. A detailed list of all tasks and deliverables is visualized in the following chart; a list which specifies the deliverables is given in Table 1.3b.

		Year 1				Year 2				Year 3			
WP1	D1.1												
	D1.2												
WP2	D2.1												
	D2.2												
WP3	D3.2												
	D3.1												
WP4	D4.1												
	D4.2												
WP5	D5.1												
	D5.2												

The WPs and respective Leading Partners (LPs) are summarized in the following table 1.3a Work package list.

Table 1.3a: Work package list

Work package No	Work package title	Type of activity	Lead partic no.	Lead partic. short name	Person-months	Start month	End month
1	CheiRank versus PageRank, centrality measures and network structure	RTD	2	UTWE	35	1	36
2	Network analysis through Google matrix eigenspectrum and eigenstates	RTD	1	CNRS	34	1	36
3	Applications to voting systems in social networks	RTD	4	UMIL	31	1	36
4	Applications of new tools and algorithms to real-world network structures	RTD	3	MTA_SZTAKI	31	1	36
5	Database development of real-world networks	RTD	4	UMIL	34	1	36
6	Management	MGT	1	CNRS	3	1	36
7	Dissemination	OTHER	1	CNRS	8	1	36
		TOTAL			176		

Table 1.3b: List of Deliverables

Del. no.	Deliverable name	WP no.	Nature	Dissemination level	Delivery date (proj. month)
1.1	**Period 1 scientific report on WP1**	1	R	PU	18
1.2	**Period 2 scientific report on WP1**	1	R	PU	36
2.1	**Period 1 scientific report on WP2**	2	R	PU	18
2.2	**Period 2 scientific report on WP2**	2	R	PU	36
3.1	**Period 1 scientific report on WP3**	3	R	PU	18
3.2	**Period 2 scientific report on WP3**	3	R	PU	36
4.1	**Period 1 scientific report on WP4**	4	R	PU	18
4.2	**Period 2 scientific report on WP4**	4	R	PU	36
5.1	**Period 1 scientific report on WP5**	5	R	PU	18
5.2	**Period 2 scientific report on WP5**	5	R	PU	36
6.1	**Period 1 scientific report**	6	R	PU	18
6.2	**Period 1 periodic report**	6	R	PU	18
6.3	**Period 2 scientific report**	6	R	PU	36
6.4	**Period 2 periodic report**	6	R	PU	36
6.5	**Final Report including Report on awareness and wider societal implications, and Final plan for the use and dissemination of Foreground**	6	R	PU	36
7.1	**Project website**	7	R	PU	3
7.2	**Initial plan for the use and dissemination of foreground**	7	R	PU	4
7.3	**Contribution to portfolio and concertation activities at FET-Open level**	7	R	PU	36

Table 1.3c: List of milestones

Milestone number	Milestone name	Work package(s) involved	Expected date	Means of verification
1	**Correlation properties of directed networks**	1	12	Realization of task **WP1.1**
2	**Statistical characterization of 2DRanking**	1,2	12	Realization of tasks **WP1.2, WP2.1, WP4.3**
3	**Eigenstate community detection**	2,3	12	Realization of tasks **WP2.2** and **WP3.1**
4	**Spam filter protocols**	4	18	Realization of task **WP4.2**
5	**Network-specific centrality measures**	1,3	18	Realization of tasks **WP1.1, WP1.3** and **WP3.1, WP3.2**
6	**Fractal Weyl law properties of networks**	2	24	Realization of task **WP2.3**
7	**Protocols for large-scale network processing**	4,5	24	Realization of tasks **WP4.1** and **WP5.2**
8	**Characterization of multi-product world trade network**	4	24	Realization of task **WP4.4**
9	**Webcrawler development and database collection**	5	24	Realization of task **WP5.1**
10	**Monte Carlo algorithms for centrality measures**	1	36	Realization of task **WP1.4**
11	**Delocalization conditions for Google matrix eigenstates**	2	36	Realization of task **WP2.4**
12	**New protocols for social voting and recommendation**	3	36	Realization of tasks **WP3.3, WP3.4**

	systems			
13	**Characterization of ranking of Wikipedia and other networks**	4	36	Realization of task **WP4.3**
14	**Characterization of time-evolving Web structures**	5	36	Realization of task **WP5.3**

Table 1.3d: Work package description

Description of Work package 1

Work package number	**1**	**Start date or starting event:**		**Month 1**
Work package title	**CheiRank versus PageRank, centrality measures and network structure**			
Activity type	RTD			
Participant number	1	**2**	3	4
Participant short name	CNRS	**UTWE**	MTA_SZTAKI	UMIL
Person-months per participant	9	**16**	5	5

Objectives 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 [1] and citation graph [49] 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), see [12,13,14]. 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. Effects of dynamical link variations in time and their influence on 2DRanking will be analyzed. 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. The results of WP1 will be used for W2.1, WP2.2, WP2.4; for WP3.2, WP3.3; for WP4.3, WP4.4. The real network data from WP5 will be used for analytical and numerical investigations in this WP1.

Description of work (broken down into tasks) and role of partners

WP 1.1 Measuring and modelling network-specific dependencies between node parameters (**UTWE**, CNRS, UMIL). Correlations between the components are crucial in 2DRanking, and vary from one network to another [13]. Current dependency measures such as the widely accepted assortativity measure are often inappropriate or even ill-defined in power law networks. Alternative measures have been suggested that partially remedy this deficiency, see e.g. [50]. However, most of suggested measures miss important information on network correlations due to averaging and summarising the results in just one number. The consortium has recently developed novel dependency measures for in-degree and PageRank of a node [21,22], based on the state-of-the-art techniques in extreme value theory, that characterize the correlations for most important nodes by a probability-like measure on a finite interval. This enables the exactly right level of mathematical accuracy, both insightful and analytically tractable, for studying the 2Dranking. We will further develop these methods for evaluating a wide range of network correlations including the degree-degree correlations. The measurements on directed networks confirmed large (in- and out-) degree correlations between neighbouring nodes in social networks, World Wide Web, food webs and word adjacency networks [51]. We aim to 1) evaluate network correlations, including CheiRank and PageRank correlations and degree-degree correlations in directed graphs; 2) formally interpret and provide techniques for reproducing the observed correlations in random networks. In addition, Markov Decision Processes will be applied for optimal modelling of modern directed networks.

WP1.2 Analytical tools for the 2DRanking distribution in directed graphs (**CNRS**, UTWE, MTA_SZTAKI, UMIL). In contrast to the wealth of empirical studies of centrality measures, their analytical properties are largely unknown, even for basic random networks the results are lacking bar rare exceptions. The strikingly insufficient mathematical understanding makes any justifiable improvement virtually impossible. The consortium pioneered probabilistic approach for the analysis of the PageRank distribution [19,20,52]. This approach, based on a branching process approximation of the neighbourhood of a node, allows to obtain the power law behaviour of PageRank distribution analytically, in a closed-form expression that depends on the model parameters. The aim of WP1.2 is to provide a probabilistic analysis and derive the probability distribution for the 2DRanking. To this end, we assume an arbitrary joint distributions of in-and out-degrees and use analytical approach [52] and the large deviation sample path approach [53]. After measuring and formalising graph correlations (WP 1.1) we will include the network-specific correlation patterns in the approximate model for 2DRanking to analytically obtain the behaviour of centrality measures in networks of different nature. At this step, a neighbourhood of a node is formally modelled as a branching tree with correlated degrees of neighbouring nodes. The tree approximation of the node's neighbourhood is often adequate and is widely used in the network literature, e.g., in the analysis of distances in random networks; see, e.g. [54] for analysis of the tree approximation in the configuration model, and [55] for a comprehensive overview. For practical application of centrality measures, however, we aim to assess this approximation and to extend the analysis from a tree-shaped neighbourhood approximation to realistic network graphs. This fundamental step has not yet been done in the literature. The developed analytical tools for studies of branching-process-based approximations and as well as for the analysis of structural properties of networks, have only recently achieved sufficient maturity to address this highly challenging task.

Modern networks have dynamical link variations in time. The related changes and sensitivity of PageRank and Cheirank, induced by such dynamical modifications of network links will be analyzed using developed models and real network data. We expect that highly communicative nodes at the top of CheiRank should play an important role in the dynamical variations of 2DRanking.

WP1.3 Design and analysis of new model-based centrality measures (**UTWE**, CNRS, MTA_SZTAKI, UMIL). The obtained analytical results will reveal the influence of the network parameters and network correlations on the centrality measures. This will provide sufficient grounds for designing new, model-based and network-specific centrality measures that take advantage of network structure in an optimal way. For instance, in highly centralised networks, such as Preferential Attachment networks [8] PageRank gives the same result as in-degree while in Wikipedia PageRank and in-degree are almost uncorrelated [21]. Likewise, the same centrality measure is not equally informative in

different networks. We aim for new measures that are: (1) informative in a specified network environment, and (2) computationally tractable. The properties of 2DRanking in standard models of scale-free networks will be investigated (e.g. for the Albert-Barabasi model). New methods of information retrieval based on 2DRanking will be developed. As a test bed for investigations of properties of 2DRanking we will use university networks, Wikipedia hyperlink networks of articles in English, French and Spanish, world trade networks. The analytical and numerical results obtained for the properties of 2DRanking will allow to put the fundamental grounds for development of search engines of new type. Such studies are crucial for translating our results in prototype software that can be used for World Wide Web, social networks and other networks.

WP1.4 Design and analysis of Monte Carlo algorithms for computation of importance measures (**UTWE**, CNRS, MTA_SZTAKI, UMIL). Traditional computational methods for centrality measures require the knowledge and storage of a complete network graph. However, obtaining network graph is often an intricate task for social reasons (e.g. personal friendships or sexual contacts) or due to technical restrictions, e.g. enormous computer capacity and skills are required to retrieve a complete World Wide Web or Internet graphs. Besides, processing a complete network data is technically demanding and time consuming process. This motivates the development of the Monte Carlo methods [30,56,57] allow fast randomized estimations of centrality measures, which can be computed on-line and without knowing the complete network. We develop Monte Carlo methods for efficient estimation of two-dimensional centrality measures. Our recent results demonstrate that in order to detect few top nodes with large PageRank values, we can use less random walk runs than the number of the nodes [58]. We will investigate the convergence of the list of top nodes in terms of ranking rather than in terms of values of centrality scores, since ranking converges much quicker than the rank values. We expect that Monte Carlo approach will help us to develop very efficient methods for detecting top nodes with respect to a rich class of centrality measures. Endowed with fast computational procedures, 2DRanking opens qualitatively new opportunities for creation of search engines of new types.

Deliverables (brief description) and month of delivery

D1.1 Period 1 scientific report on WP1 (M18)

D1.2 Period 2 scientific report on WP1 (M36)

Description of Work package 2

Work package number	**2**	**Start date or starting event: Month 1**		
Work package title	**Network analysis through Google matrix eigenspectrum and eigenstates**			
Activity type	RTD			
Participant number	**1**	2	3	4
Participant short name	**CNRS**	UTWE	MTA_SZTAKI	UMIL
Person-months per participant	**17**	9	4	4

Objectives WP2 will investigate 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 in [12,23] established the validity of fractal Weyl law, found recently in systems of quantum chaotic scattering and Perron-Frobenius operators of dynamical systems. WP2 will investigate the spectrum of Wikipedia network analyzed recently in [13]. The eigenmodes with eigenvalue modulus being close to α correspond to slow relaxation modes in networks. Such modes should be linked with specific communities hidden inside network. The Arnoldi method will allow detecting 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 with decomposition of degenerate subspaces with $\lambda=1$ are shown in Fig. right panel. The size of degenerate subspaces can be rather large (around 40000 for the case shown in Fig.). This Arnoldi approach will be also applied for networks of Wikipedia articles, open source software networks, university networks considered in [11,15,24]. 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 results presented in [13]. 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 [24] will also be determined for various networks. The results of WP2 are used for WP3.1, WP3.2, WP4.2, WP4.3, WP4.4. This WP2 uses real network data from WP5.

Description of work (broken down into tasks) and role of partners

WP 2.1 Random matrix models of Google type matrices (**CNRS**, UTWE, UMIL). WP2.1 develops random matrix models of Google type matrices and investigates their properties. From the physics of electron transport in disordered systems it is known that probability spreading can be suppressed by interference processes leading to exponential localizations in disordered solids known as the Anderson localization [25,26]. WP2.1 will develop various random models for Google matrices of various complex networks and will study conditions under which PageRank and other eigenmodes are extended or localized. These models should reproduce the important property of real networks which have no gap in the spectrum of Google matrix (at $\alpha=1$). The usual models used in the field of complex networks, e.g. the Albert-Barabasi model, have very large gap (see e.g. results presented in [11]), and hence new models should be developed to have new random matrix models of G without spectral gap. The sensitivity of PageRank to the Google damping factor α will be analyzed by analytical methods and extensive numerical simulations. Recent results obtained in [24] on the basis of Ulam networks generated by chaotic nonlinear maps show that PageRank can become delocalized for a certain range of α. Such delocalization would produce a drastic effect on efficiency of search engines since a delocalized PageRank becomes very sensitive to various fluctuations that make such ranking vulnerable. The results in progress show that band random matrix models of Google matrices have exponentially localized eigenstates including the PageRank. This is rather similar to the Ander-

son localization in disordered solids. However, an addition of new links of small-world type can generate a delocalization transition. Preliminary obtained results show close similarities between band random matrices studied for disordered electrons with Anderson localization and Markov chains with band structure and disorder. WP2.1 will clarify how the characteristics of network are linked with the localization properties of eigenstates of the Google matrix G. The experience gained by our group in the fields of quantum chaos, many-body quantum systems and multi-qubit systems will allow developing new approaches to the analysis of matrix properties of complex networks.

WP2.2. **Eigenspectrum and eigenfunctions of Google matrix of directed networks** (**CNRS**, UTWE, MTA_SZTAKI, UMIL). The Arnoldi method allows to determine the dominant part of the spectrum and eigenstates of Google matrices of sizes of a few millions [15]. With this approach isolated subspaces which have $\lambda=1$ at a damping factor unity are well-identified. In addition, other eigenstates with eigenvalue modulus being close to one are also efficiently computed by this approach. The eigenstates with such eigenvalues correspond to slow relaxation modes in such directed networks being typical for real scale-free network [11,15]. The physical intuition tells that such eigenmodes should correspond to isolated and weakly coupled communities which are of great interest for network characterization. To verify this conjecture, we will study the eigenstates of the Wikipedia Google matrix [13,14,15], where the names of the articles allow to identify communities in a relatively easy way. Other tests will be performed with other types of directed networks including the world trade networks obtained on the basis of UN COMTRADE database [16]. The Google matrix for these networks will be generalized to the multiproduct trade between all UN countries which corresponds to matrix size of about 100 000.

WP2.3 Fractal dimensions and fractal Weyl law for directed networks (**CNRS**, UTWE, MTA_SZTAKI, UMIL) . Recent results obtained by the consortium show that the spectrum of Ulam networks of dynamical systems [24,59] and of real scale-free networks [23] is characterized by a non-integer scaling in N with an exponent $\nu=d/2$ smaller than unity. Only a small fraction of eigenvalues has finite values of λ while almost all eigenvalues drop to zero. This fact can be related to the fractal Weyl law which appears in quantum systems with chaotic scattering [60]. The spectral properties of the Linux kernel procedure call network are connected to the network topological features, which in this particular case follow programming convenience rules. The confirmation of the fractal Weyl law ($\nu=d/2$) with $\nu \approx 0.65$ and $d \approx 1.3$ has been found for different Kernel versions [23] with growing number of procedures, where the relevant number of resonances grow as $N_\lambda \propto N^\nu$ (number of states with $0.1<|\lambda|\leq 1$). The fractal dimension of this Linux network was computed using the cluster growing method and is shown to be approximately 1.4. The conditions of appearance of the fractal Weyl law in the complex directed networks will be firmly established during this project. Fractal Weyl law implies that the number of important states with finite values of relaxation rate $\gamma=-2\ln|\lambda|$ is relatively small and thus their characterization will allow to extract important information about network properties. In this WP2.3 we will investigate fractal Weyl properties for other scale-free networks including Wikipedia hyperlink network, open source softwares, universities and other Web networks taking into account the variation of their size N with time.

WP2.4 Localization and delocalization properties of Google matrix eigenstates (**CNRS**, UTWE, MTA_SZTAKI) .The studies of network models and Ulam networks [11,24,26,59] have shown that under certain conditions the PageRank can become completely delocalized. Such delocalization has certain similarities with the Anderson transition in disordered solids [25], and its appearance would produce a drastic impact on the performance of the PageRank algorithm. In this case the PageRank becomes flat being dominated by small random fluctuations. The whole world would go blind the day such delocalization occurs. At the moment the PageRanks of many networks including WWW are located in the localized phase but the conditions of localization and its properties should be investigated in detail in order to meet the danger of delocalization transition due to future network growth. Indeed, new types of citation via pdf-files and other formats generate links in automatic way that can lead to drastic changes of PageRank and CheiRank properties [14]. We will study evolution of localization properties of eigenstates (e.g. participation ratio and eigenstate entropy used in solid

state systems [25]) with the increase of network size with time for various types of networks including open software, Wikipedia network, universities networks, Ulam networks and random models of Markov chains. The progress of WP2.4 will use the results obtained in WP1.2, WP2.1. Conditions of Anderson type delocalization transition in scale-free networks will be obtained. Effects of dynamical link variations on localization will be inverstigated.

Deliverables (brief description) and month of delivery

D2.1 Period 1 scientific report on WP2 (M18)

D2.2 Period 2 scientific report on WP2 (M36)

Description of Work package 3

Work package number	**3**	**Start date or starting event:**		**Month 1**
Work package title	**Applications to voting systems in social networks**			
Activity type	RTD			
Participant number	1	2	3	**4**
Participant short name	CNRS	UTWE	MTA_SZTAKI	**UMIL**
Person-months per participant	9	5	4	**13**

Objectives *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) to 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, this data base is available for UMIL, cf. Arxiv:1111.4570); 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 [27,61] 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.

WP3 will analyze the voting graphs and opinion formation on real social networks (e.g. Facebook, Twitter) using the algorithmic tools developed in WP1, WP2. The effects of PageRank of neighbours on their vote and opinion formation on the whole network will be investigated in detail. Voting on social networks is getting an enormous importance for commercial applications (e.g. like/unlike opinion for a given hotel or restaurant, cf www.nomao.com). WP3 will give real tools which will allow to understand the process of opinion formation on social networks. The efficiency analysis of opinion propagation/formation on such networks will have important applications for existing and future social networks. WP3 uses real network data from WP5.

Description of work (broken down into tasks) and role of partners

WP 3.1 Eigenvectors for spectral voting (**UMIL**, CNRS, MTA_SZTAKI).
Spectral voting just scratches the surface of the information provided by the voting graph. The eigenstates of this voting graph will be analyzed by the analytical and numerical tools developed in

WP2. The first question is (as it happens in other ranking systems, such as HITS), is which knowledge can be extracted by other eigenvectors (beyond the dominant one). For instance, we expect emergent but latent opinion to be detectable by the study of the entire spectrum. The expertise of the consortium will allow studying the entire spectrum and the associated eigenvector spaces to explore these issues. We will use Facebook data base available to UMIL (cf. Arxiv:1111.4570).

WP3.2. **Social voting analysis through centrality measures** (**UMIL**, CNRS, UTWE, MTA_SZTAKI). When voting is not complete, the presence of censored data can be dealt with by modelling missing votes using random variables (e.g., a uniform choice of neighbours). In this case, probability theory can be used to provide results and even a *centrality index* based on a total random vote. We will investigate such indexes using the consortium expertise in probability theory. In particular, we would like to understand their position among centrality measures of the first category (measuring how a node is well connected to the other nodes—instances are node degree, closeness centrality, PageRank) and second category (measuring to what extent a node helps communication flow inside the network, instances are betweenness centrality, random walk betweenness centrality, second order centrality). The study of these different measures applied to various types of social networks will allow specifying which centrality measure is better suited for which type of online social network analysis problem. The analytical methods and numerical tools of WP1, WP2 will find here an important range of applications. For example, CheiRank vector selects mostly communicative nodes which are very efficient for the information propagation and hence such nodes will play an important role in opinion formation and its propagation through the whole network. PageRank probability of neighbours also produce an important influence on the choice of opinion of a given node (e.g. + or -, as in spin systems of Ising model). Such voting spin like systems are investigated in the frame of physical research (e.g. Volovik, Redner arxiv:1111.3883), however, the scale-free networks and the pageRank of nodes have not been analyzed till present. We intend to develop a Facebook application which will be used to perform experiments with voluntary participation from Facebook users. The idea is to describe a decision and let people either state their preference or choose some friend that is better suited to take the decision. More than in the decision itself, we are interested in gathering the delegation graphs that will be generated, to study their properties. If we are able to make the process sufficiently viral, this will provide us with interesting non-toy datasets.

WP3.3 Social network analysis through graph neighbourhood function (**MTA_SZTAKI**, UMIL, UTWE). A recent breakthrough made by member of the consortium [62] that makes such analysis possible is the design and the very carefully engineered implementation of an algorithm for computing with arbitrary precision the *neighbourhood function* of a graph, that is, given the function associating with each *t* the number of pairs of nodes reachable in less than *t* steps. From the neighbourhood function it is easy to compute the cumulative distribution of distances, which, besides providing basic measures such as the average distance and the harmonic diameter, makes a fine perturbation analysis possible: how does the distribution change when nodes with a high ranking are removed. This type of analysis, well-known in the study of complex networks, has been carried on for very small networks, and mostly on synthetic model, due to the unavailability of feasible algorithms and datasets: both the datasets created by WP5 and the expertise of the consortium on computing the neighbourhood function will be used to analyze the fine perturbations of the distance distribution when nodes of high rank are removed, highlighting connections between first and second category centrality measures. The properties of the *neighbourhood function* of a graph will be analyzed by the tools developed in WP1 including CheiRank, PageRank and 2DRanking properties of nodes: it is important to know not only those nodes which are in a close vicinity to a given one, but also to know its global importance in the network. We will test the efficiency of opinion propagation (e.g. like/unlike) through Facebook application developed in WP3.2.

WP3.4 Recommendation systems in social networks (**MTA_SZTAKI**, UMIL, CNRS).
In WP3.4, in addition to voting systems, the new technologies are also applied in recommendation systems. Recommendation systems involve a high level of feedback from the user about the quality of the results. Measuring the results in recommendation applications and analyzing implicit and ex-

plicit user feedback will measure the quality of the innovative network analysis methods. The recommendation solutions use specific local information from the system to which the recommendation engine is applied: an Internet site, an appliance TV solution, etc. The recommendation builds personalized preference models based on the historical usage data. The historical data represents user interaction events with the system. For example, interactions with an item or item independent user activities like logging in or out.

Recommendation systems are significantly affected by the so-called cold start problem. Specifically, the automated preference model building approach used in the recommendation algorithms cannot draw inference for users or items because of the lack of sufficient data relating to the new user. Consequently, the quality of recommendations for new users is poor, and also new items in the system are not recommended efficiently until sufficient information is gathered. The efficient recommendation of new items is particularly important in certain applications, such as news, product, and TV program recommendations. News articles have dominantly very short lifetime, therefore efficient and instant characterization of fresh articles in within the article space is crucial. Similarly, upcoming TV shows that haven't been broadcast are difficult to target; here event information gathered during broadcasting can only be used in catch-up service and in show repetition.

To partly overcome the item specific cold start problem, item metadata can be exploited. In this regard new items are characterized by means of their metadata similarity to already known items. This solution is unfortunately limited by the amount and quality of metadata available locally when new items are inserted into the item catalogue. The enriched data gathered from the web and social media provided by WP5 can help to tackle this problem. Identifying and correlating information from the this larger scope with items within the realm of the recommendation engine allows recommendations to be made with much broader information than would be available to it normally leading to much higher quality results and helping to avoid the cold start problem. Reciprocally the information gathered by the recommendation engine can be used to help refine the focus of the harvesting of information from the web and social media. The results obtained in WP3.1, 3.2, 3.3 will useful for the performance and realization of this WP3.4 .

Deliverables (brief description) and month of delivery

D3.1 Period 1 scientific report on WP3 (M18)

D3.2 Period 2 scientific report on WP3 (M36)

Description of Work package 4

Work package number	**4**	**Start date or starting event:**		**Month 1**
Work package title	**Applications of new tools and algorithms to real-world network structures**			
Activity type	RTD			
Participant number	1	2	**3**	4
Participant short name	CNRS	UTWE	**MTA_SZTAKI**	UMIL
Person-months per participant	3	6	**16**	6

Objectives

Methods of WP1-WP3 are implemented in large scale applications based on real data collected in WP5. Achievements in this WP4 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. Effects of dynamical variations of links in rapidly changing mass media networks, like e.g. BBC and LeMonde, will be investigated. 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. This WP4 uses results of WP1-WP3 and real network data from WP5.

Description of work (broken down into tasks) and role of partners

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

We address the problems of graph partitioning for efficient distributed processing. We consider a variety of graph algorithms from WP 1-3 that we plan to dynamically update as new data arises. Of particular importance is fingerprinting for similarity and ranking. This task contributes to open source frameworks (Hadoop, HAMA, etc.). One means of scaling a variety of graph algorithms from WP1-3 involves graph partitioning for efficient distributed processing. We also address the task of dynamic update as new data arises. Of particular importance is fingerprinting for similarity and ranking, well-known technologies to scale SimRank and personalized PageRank that may be applicable for the novel network analysis technologies as well. Several graph algorithms can be parallelized by splitting the graph into pieces and distributing the pieces to different servers where they fit into internal memory. Such an architecture can serve shortest path queries, local network flows for community core queries ("give me the nearest community"), similarity search and more. In these applications it is crucial to partition into pieces so that most queries can be solved locally or with a few hops. So this is basically a graph partitioning problem where the quality of the partitioning is defined by the running time or number of communication hops for hopefully some realistic input for one of the above tasks.

Due to the size of the graphs in question, distributed partitioning is an additional challenge. We are planning to build on our spectral partitioning results, a method that, on real networks is thought to perform bad prior to our results that resolve this problem [63]. Given a good graph partitioning, or a predefined distributed graph arising as the result of a distributed Web crawl, our next goal is to update query structures over this data. We consider path fingerprint update [30] as an exemplary application. If a batch of new vertices and edges are added with possible deletions, then each server can update its portion of the path fingerprints but changes may propagate to other nodes. We would like

to measure the amount and rounds of communication in various algorithms.

Hadoop, an open source implementation of the map-reduce framework, is considered SoA in distributed Web processing. However large graph and matrix processing goes beyond Hadoop and map-reduce capabilities. First of all we devise new ways to formulate and implement large-scale graph and matrix processing algorithms. We use iterative matrix-vector multiplication with generalizations and optimizations on Hadoop: an open source effort is http://www.cs.cmu.edu/~pegasus/. We also plan to rely on direct node messaging to minimize communication costs that can be based on a Message Passing Interface implementation such as OpenMPI, a framework typically supported on clouds. In addition, incubatory implementations of graph processing frameworks in the Bulk Synchronous Parallel model similar to Google's Pregel also exist. Finally, we can reuse Hadoop codes for inter-process communication and file I/O.

WP4.2. **Network quality and trust classification and spam filtering** (**MTA_SZTAKI**, UMIL, UTWE)

Networked entities express trust and recommendation along their connections that can be exploited for quality prediction. Likewise, malicious behaviour such as spamming, vote or opinion manipulation, Google bombing appear in networks of connected true or virtual, generated entities. Selecting the trustworthy, central and opinion forming nodes while avoiding noise and selfish or malicious actors is a main goal of NADINE.

We apply NADINE methodology to machine learning assisted trust and quality assessment by building on both the new technologies developed in WPs1-2 on link analysis as well as the continuously improving machine learning technologies of the consortium. As a speciality of the NADINE environment, we devise methods that are robust towards the heterogeneity and multilingualism of the data sources and leverage on language independent link analysis techniques to the greatest possible extent. We will develop technologies to normalize features and models across statistically very different subcollections as well as deploy cross-lingual techniques as existing in the Consortium.

As a combination of distributed procedures developed in the NADINE project, we will be able to

- Classify and filter networks at the Peta-scale;
- Automatically learn parameters to support use of the technologies beyond the Consortium.

WP4.3 2DRanking and centrality measures of Wikipedia, open software and other networks(CNRS, UMIL, UTWE, MTA_SZTAKI)

Using the algorithms and tools developed in WP1-WP3 as well as in WP4.1 and WP4.2, we will analyze databases generated by WP5. Special attention will be devoted to Wikipedia hyperlink network [13,14], including comparisons between English, German, French, Italian and Spanish editions. The evolution of the obtain characteristics with time will be analyzed in detail. We will also study open software procedure call networks [12,14,23] for Linux, OpenGL, and Python. In parallel we will apply the above algorithms to the study of gene regulation networks [14,17]. 2DRanking and other centrality measures will be applied to large scale parts of the World Wide Web, including University networks, mass media web sites like BBC, LeMonde, CNN and others. These sites rapidly evolve in time, and our algorithms will allow capturing most rapid and important variations of information flow in these mass media networks. Dynamical variations of ranking will be investigated and the sensitivity of 2DRanking to dynamical changes will be determined. The obtained results will allow developing efficient methods to understand and analyze time-dependent networks. Recent developments for networks appearing in games (e.g. go) allow extracting promising strategies with the Google matrix methods [64]. These studies will be further extended to various types of games.

WP 4.4 Analysis of Google matrix of multiproduct world trade network (**CNRS**, UTWE, MTA_SZTAKI) Using the United Nations Commodity Trade Statistics Database we recently constructed the Google matrix of the world trade network and analyzed its properties for various trade commodities for all countries and all available years from 1962 to 2009 [16]. The trade flows on this network are classified with the help of PageRank and CheiRank algorithms developed for the World Wide Web and other large scale directed networks. For the world trade this ranking treats all countries on equal democratic grounds independent of country richness. Still this method puts at the top a

group of industrially developed countries for trade in all commodities. Our study establishes the existence of two solid states like domains of rich and poor countries which remain stable in time, while the majority of countries are shown to be in a gas like phase with strong rank fluctuations. In WP4.4, we will generalize this approach to the most important and actual situation of multiproduct trade which corresponds to a new case of colored directed networks with nontrivial interactions between products and countries participating in the trade. The centrality measures and 2DRanking tools developed in WP1-WP3 will enable to shed new light on the important field of international trade. We note that recent Google+ Circles have a certain analogy with multiproduct trade networks that will allow having cross-fertilization between trade flows and social networks.

Deliverables (brief description) and month of delivery

D4.1 Period 1 scientific report on WP4 (M18)

D4.2 Period 2 scientific report on WP4 (M36)

Description of Work package 5

Work package number	**5**	**Start date or starting event:**		**Month 1**
Work package title	**Database development of real-world networks**			
Activity type	RTD			
Participant number	1	2	3	**4**
Participant short name	CNRS	UTWE	MTA_SZTAKI	**UMIL**
Person-months per participant	5	4	11	**14**

Objectives 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. WP5 provides real network data which are extensively used in WP1,WP2,WP3,WP4.

Description of work (broken down into tasks) and role of partners

WP 5.1 Crawler development and database collection (**UMIL**, MTA_SZTAKI, CNRS)
Existing open source crawlers do not meet the requirements of the NADINE project, including linear scalability on multiple machines, easily pluggable procedure to extract link from visited pages (to be able to crawl social networks), constant-memory crawling (data structure must be partially offline, and the crawler must be able to complete the crawl, no matter the size, in a fixed or logarithmically growing amount of memory), speed of processing, very flexible filtering system for all crawling phases, antispam capabilities.

In this task we implement a fast, distributed crawler, building on the know-how developed with UbiCrawler [65]. We aim to develop a fully functional and configurable crawler that can be used by researchers interested in gather both general and *focused* snapshots (i.e., snapshots in a single language, snapshots related to a set of keywords, etc.). The crawler will be distributed as open-source software at the end of the project. Due previous positive experience of members of the project we plan to implement the crawler in Java. Java has a number of advantages in terms of speed of development, large number of open-source libraries available, and programming tools.

In the last phase, the crawler will be used to gather large snapshots (500 million to one billion pages) from which large web graphs will be extracted. Crawls of this size are known to be extremely complex activities, due not only to design, implementation and efficiency issues, but also to practical, legal and administrative difficulties. The crawls will be made available initially to the project members and, after a grace period of one year, to the general public.

In parallel, we will investigate the possibility of gathering datasets based on social networks. Several such datasets are available to the public in several non-standard formats: we aim at creating a single-source database in a single, highly compressible, yet quickly accessible, format, based on our previous work on WebGraph [66], a Java framework for graph compression.

We remark that making access to datasets easy and uniform is essential to push researchers to perform experiments on a wider range of type of networks. Indeed, as it has happened in the past, the mere availability of a dataset in a very easily accessible format can push researchers of a community to use it widely, even if it is, actually, a very bad dataset.

WP5.2. **Internet Scale Data Management** (**MTA_SZTAKI**, UMIL, UTWE)
The NADINE approach will enhance existing approaches in at least two dimensions. First, NADINE will develop strategies to facilitate the processing of ultra-heavy computations by efficiently identifying relevant data and distributing only the required data among its nodes. Second, unlike existing approaches, in the case of social networks, data contains rapidly emerging and evolving digital content

where it not clear up-front, which are the decisive parameters for analysis. Hence, NADINE will investigate novel methods to discover and perform analytic tasks on the fly.

We plan to develop distributed versions of duplicate detection and linkage analysis including neighbourhood analysis. Even computing the simplest graph based features such as in-degrees requires nontrivial methods in a distributed settings while applications such as Web classification and trust or quality ranking require complex features based on the multi-step neighbourhood of Web hosts.

We plan to develop methods to aggregate large matrix processing results computed individually for different subsets of a large unified graph. For example we may compute graph ranking or SVD for large collections at different times and at a later time on demand fuse them by reusing the partial results.

WP5.3. Cross-data and temporal Web analytics (**MTA_SZTAKI**, UMIL, CNRS)

The comprehensive cross-data temporal Web analytics which will be provided by the NADINE tools will open a completely novel research area to the whole scientific community. While there is speculation that intelligence agencies and cross-platform Web companies analyze log data, there has not been a study of similar kind on the "wild" Web that is open to the scientific community. Hence, NADINE backed by the available huge data collections will allow unique studies of Web data to the whole scientific community, which have not been possible that comprehensibly before. Hence, NADINE will support discovery, modelling and (even) prediction of novel patterns of occurrence on the Web and social media.

We develop distributed active learning where the human assessor may have access only to portion of the data. Restricted access may result for technological reasons: data may be local as for example all embedded elements of a Web page will not be transmitted. Human factors may however also restrict access: the quality assurance team of a French archive may not be able to assess Web pages in Hungarian, or may lack the domain knowledge for topical assessment on specialized fields. By *visualizing connections between entities* we will turn the available large volumes of disparate data into actionable intelligence that reveals hidden patterns and relationship based on our evolving visualization solution [67].

Deliverables (brief description) and month of delivery

D5.1 Period 1 scientific report on WP5 (M18)

D5.2 Period 2 scientific report on WP5 (M36)

Description of Work package 6

Work package number	**6**	**Start date or starting event:**		**Month 1**
Work package title	**Management**			
Activity type	MGT			
Participant number	**1**	2	3	4
Participant short name	**CNRS**	UTWE	MTA_SZTAKI	UMIL
Person-months per participant	**3**	0	0	0

Objectives To ensure the successful delivery of the aims of the project in an efficient and cost-effective manner, and to establish successful channels of communication between project partners, the European Commission, other interrelated European and national projects and the scientific community.

Description of work (broken down into tasks) and role of partners
The project will be coordinated by the French scientific partner (CNRS). The project coordinator will assume overall responsibility for the efficient administrative, financial and scientific management of the proposed project, together with the implementation, integration, reporting and dissemination fall components of the project. They will be responsible for ensuring clear, effective and rapid channels of communication within the group and for establishing and maintaining contact with the European Commission and their interrelated projects and with other research.

Deliverables D6.1 and **6.3** will contain the ready-to-publish Executive Summary plus an Introduction that explains the inter-relation of the WPs, that will then constitute the chapters, from the overall project perspective. Annexes with publications and dissemination activities will follow.

Deliverables D6.2 and **D6.4:** (in agreement with the instructions for project reporting (ftp://ftp.cordis.europa.eu/pub/fp7/docs/project_reporting_en.pdf), periodic reports will discuss financial/management issues as well as technical details; they will be structured according to the following sections:

- A publishable summary
- Project objectives
- Work progress and achievements
- Project management
- Explanation of the use of the resources

Deliverable D6.5 will represent the **Final report** including **Report on awareness and wider societal implications** and **Final plan for the use and dissemination of Foreground**

Deliverables (brief description) and month of delivery
D6.1: Period 1 Scientific Report (M18)

D6.2: Period 1 Periodic Report (M18)

D6.3: Period 2 Scientific Report (M36)

D6.4: Period 2 Periodic Report (M36)

D6.5: Final Report including Report on awareness and wider societal implications and Final plan for the use and dissemination of Foreground(M36)

Description of Work package 7

Work package number	**7**	**Start date or starting event:**		**Month 1**
Work package title	**Dissemination**			
Activity type	Other			
Participant number	**1**	2	3	4
Participant short name	**CNRS**	UTWE	MTA_SZTAKI	UMIL
Person-months per participant	**2**	2	2	2

Objectives
The aim of this workpackage is to disseminate the project results, to make them accessible for a broader public, and thereby to raise the partners' public profile.

Description of work (broken down into tasks) and role of partners

Task **WP7.1:**
In order to support scientific cooperation at the FET-Open level and broad public awareness of project achievements, consortium members will ensure within the areas of interest of the project:

• Project results shall be published throughout the duration of the project in widely accessible science and technology journals, as well as through conferences and through other channels, including the Web, reaching audiences beyond the academic community.

• Beneficiaries shall deposit an electronic copy of the published version or the final manuscript accepted for publication of a scientific publication relating to foreground published before or after the final report in an institutional or subject-based repository at the moment of publication.

• Beneficiaries are required to make their best efforts to ensure that this electronic copy becomes freely and electronically available to anyone through this repository: immediately if the scientific publication is published "open access", i.e. if an electronic version is also available free of charge via the publisher, or within 6 months of publication.

• Periodic press releases shall be issued, and other means of disseminating project progress to a wider audience e.g. via video.

• Participation in FET-organised events, for example conferences, dedicated workshops & working groups, consultation meetings, summer schools, online fora, etc.

• International Co-operation - contribution to relevant national and international activities (e.g. Joint workshops, calls, etc...).

• The collected databases will be made available to public via the project website

The above activities will be reported in the project's Dissemination Plan and in periodic progress reports. In addition, the consortium agrees to include the following reference in all project-related publications, activities and events:

"The project NADINE acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 288956"

The scientific results obtained by the consortium are made open to public via publications,

website and other means. However, the explanation of the importance of the results for specific network applications and the know-how of their concrete applications and efficient means of their incorporation in the software codes and tools will be given to European public institutions and private companies. For example, the fundamental paper of Brin and Page [1] is publicly available but the know-how and specific implemetation tools are jealously kept by Google. For finding European applications we will leverage on our active connections with European large companies (Vodafone, T-Mobile), SMEs (Hanzo Archives - social media analysis; Gravity RD - recommender solutions, NOMAO.com), social network services (Last.fm, Xing) and institutions (Internet Memory Foundation) etc. We will seek collaboration with European initiatives and networks (Pascal2, TrebleCLEF, Chorus+, Theseus, FOC) as well.

Task WP7.2 is linked to Deliverable D7.2. Deliverable D7.2 will provide the initial dissemination plan that extends over the whole duration of the project. The plan will identify targeted international journals and conferences for research paper submissions and other opportunities to disseminate the project results. In addition, it will provide a tentative plan of special sessions and special issues, compatible with the current status of open and foreseen calls. The kick-off NADINE meeting is planed for the period 22-25 July 2012 during the workshop "Spectral properties of complex networks" at ECT Trento, Italy during 23-27 July 2012 (see www.quantware.ups-tlse.fr/complexnetworks2012/). The reviews of the project are planed at Toulouse on 19th month and at Brussels at 37th month.

Task **WP7.3: Contribution to portfolio and concertation activities at FET-Open level.**

Dissimination process will be under control, development and updating during the whole grant period.

In order to support scientific cooperation at the FET-Open level and broad public awareness of project achievements, consortium members will ensure within the areas of interest of the project: . Project results shall be published throughout the duration of the project in widely accessible science and technology journals, as well as through conferences and through other channels, including the Web, reaching audiences beyond the academic community. Beneficiaries shall deposit an electronic copy of the published version or the final manuscript accepted for publication of a scientific publication relating to foreground published before or after the final report in an institutional or subject-based repository at the moment of publication. Beneficiaries are required to make their best efforts to ensure that this electronic copy becomes freely and electronically available to anyone through this repository: - immediately if the scientific publication is published within open access, i.e. if an electronic version is also available free of charge via the publisher, or - within 6 months of publication. · Periodic press releases shall be issued, and other means of disseminating project progress to a wider audience e.g. via video. Participation in FET-organised events, for example conferences, dedicated workshops & working groups, consultation meetings, summer schools, online fora, etc. International Co-operation - contribution to relevant national and international activities (ex. Joint workshops, calls, etc for example with NSF). The above activities will be reported in the project's Dissemination Plan and in periodic progress reports. In addition, the consortium agrees to include the following reference in all project-related publications, activities and events: "The project NADINE acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number 288956".

Deliverables (brief description) and month of delivery

D7.1: Project Website (M3)

D7.2: Initial plan for the use and dissemination of foreground (M4)

D7.3: Contribution to portfolio and concertation activities at FET-Open level (M36)

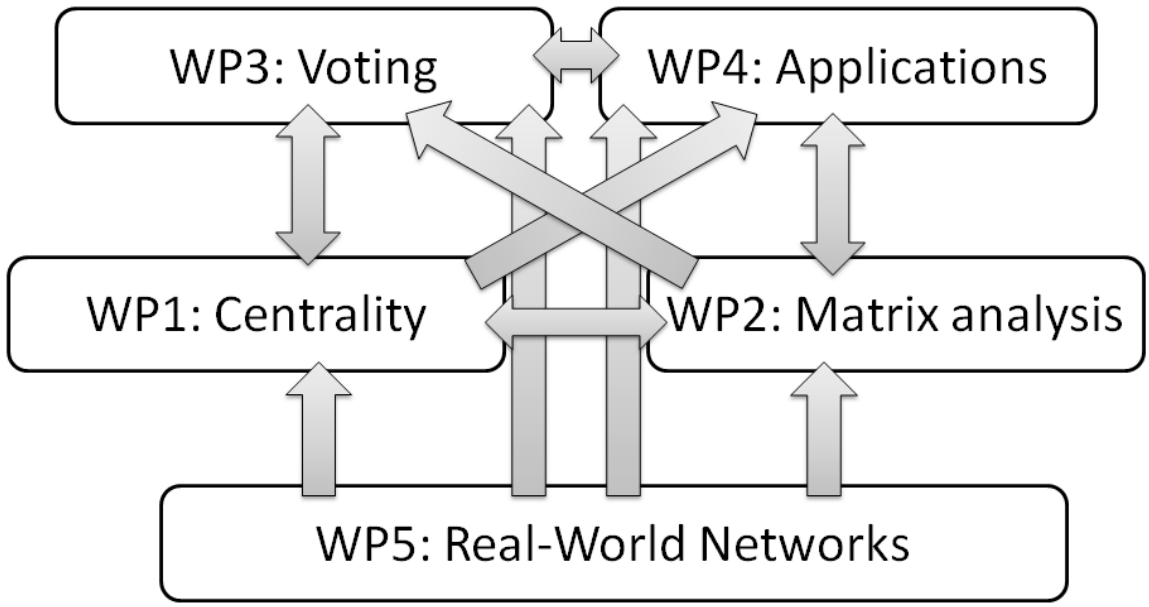
Table 1.3e: Summary of effort

Summary effort

The NADINE project is intended to last 36 months. We envisage an increasing integration of all workpackages as a function of time. The table 1.3e gives distribution of person-months per partner and per project WP (see also table 1.3a and time-Wp allocation Fig.). Each WP is leaded by a designated partner (marked in bold) who has a largest share of month-allocation. However, each partner contributes to all scientific WPs that ensures deep integration and collaboration inside the consortium. All scientific WPs have similar amount of person-months with a slightly higher accent to WP1,WP2 which will be used by all other WPs (see graph 1.3.3).

Partic. no.	**Partic. short name**	**WP1**	**WP2**	**WP3**	**WP4**	**WP5**	**WP6**	**WP7**	**Total person months**
1	CNRS	9	**17**	9	3	5	**3**	**2**	48
2	UTWE	**16**	9	5	6	4	0	2	42
3	MTA_SZTAKI	5	4	4	**16**	11	0	2	42
4	UMIL	5	4	**13**	6	**14**	0	2	44
Total		35	34	31	31	34	3	8	176

1.3.3. Graphical presentation of the components showing their interdependencies



1.3.4. Risks and Associated Contingency Plans

The project management team will perform continuous evaluation throughout the project, identifying any possible problems/risks at an early stage so that solutions can be elaborated in time. A systematic approach will be adopted for monitoring resource spending against project budget and achievements against schedule and critical success factors. In addition, a Risk & Contingencies Plan (RCP) will be applied to assess potential risks and take appropriate, realistic and effective actions within the project context. The RCP comprises four phases:

- Risk identification determines which risks might affect the project and documents their characteristics;
- Risk analysis elaborates a „Risk Matrix" in order to assess, prioritise and manage identified risks to the Project;
- Risk response actions include the design and implementation of answers to mitigate potential occurrence of the identified risks and to reduce threats to the project's objectives;
- Risk monitoring and control, finally, covers on-going follow-up of residual risks, identifying new risks, executing risk reduction actions and evaluating their effectiveness throughout the project life cycle.

The RCP will be planned in more detail in the beginning of the project and it will involve all partners. At the governing board meeting, identified risks will be revisited and the situation in the project will be analysed to identify newly upcoming risks. Furthermore, a channel will be established for WP leader to report on more detailed technology risks that might be identified in the works of the individual WPs. Identified risks are documented and integrated into Risk Matrixes in order to provide a clearer vision of the overall project situation from a risk management point of view.

Some preliminary risks have already been identified, and a possible contingency solution has been formulated for each of them. These are reported in the tables below.

Risk description	**Risk Assessment**	**Contingency solution**
Unforeseen technical problems may not be resolved with the assigned resources	**Impact**: Medium **Probability**: Medium Since the NADINE work plan contains various demanding research challenges this risk has to be considered. The risk is not too high, since the consortium members bring the required experience and expertise to judge the viability of the research topics within the planned project resources.	In case this risk occurs the partners are committed to invest a certain amount of additional own resources, since most of the addressed topics are also of high personal interest for them as researchers. In case this is not sufficient the governing board of the project in collaboration will assess the situation with the involved WP leader to decide about adequate replanning actions that ensure

		to overall project result.
Technology planned in NADINE becomes available from a third party	**Impact**: High **Probability**: Low-Medium This is a general risk for a three years research project; the probability is not too high, since directed network analysis technology is a very innovative area, where not too many parties do development and research in.	The consortium will perform regular technology watch activities in all relevant areas to ensure that the NADINE teams are aware, when this risk shows up. If competing technology becomes available, this will be evaluated. Where appropriate such technology will be incorporated and built upon in the project.
Lack of consensus within consortium	**Impact**: High **Probability**: Low The good collaboration climate and the mutual understanding of the partners make this very improbable.	Within the implementation plan management procedures have been established for enabling effective decision making; the project coordinator and the members of the governing board have the necessary skill to resolve such conflicts by adequate negotiation as well as the means required to avoid a blocking of the project by a management decision.
Project Partner leaves the consortium	**Impact**: High **Probability**: Low All consortium partners are very interested and committed to the project results (although from different perspectives due to the various roles in the project). This makes the probability of one partner leaving the consortium very low.	In case a partner will leave the consortium, the missing contributions from this partner are assessed. Further, steps depend on the result of this assessment. Typically, some of the missing contributions can be assigned to other partners and/or a new partner with adequate competences has to be identifies; the consortium members have a sufficient professional network to identify an adequate new partner.

Section 2: Implementation

2.1 Management structure and procedures

Management structure

Coordinator Dima Shepelyansky

Leader WP1 Nelly Litvak

Leader WP2 Bertrand Georgeot and Dima Shepelyansky

Leader WP3 Sebastiano Vigna

Leader WP4 Andras Benczur

Leader WP5 Sebastiano Vigna

Leader WP6 Dima Shepelyansky

Leader WP7 Dima Shepelyansky

Role of coordinators and work package leaders

The coordinator will be responsible for the day-to-day management of the project, reporting, liaison with the European Commission, checking the quality of the deliverables, dissemination activities, and project evaluation. Furthermore, he will be responsible for organizing and chairing an annual Project Coordination meeting.

The workpackage leaders will be responsible for the management and coordination within the workpackage, and will be members of the project coordination committee.

Management procedures:

Project Coordination Meeting:

Given the small scale of the consortium, its management will be kept light. Every year, we will organize a small workshop in which outside researchers interested in the topics of this project are invited. The project coordination meeting will be part of it. The research progress will be analysed in detail, and future directions of investigations will be defined.

The integration of all partners will be started at the first NADINE meeting planified to take place during the international Workshop “Spectral properties of complex networks” at European Center for Theoretical Studies in Nuclear Physics and Related Areas (ECT*), Trento, Italy during 23-27 July 2012. Coordinator and Partner 2 are among organizers of this workshop (see www.quantware.ups-tlse.fr/complexnetworks2012/), The event will attract about 45 leading researchers in the network sciences from all over the world. The interdisciplinary structure of this workshop is directly visible from the main scopes of ECT*. This event will launch the FET Open NADINE project and will make it well-known among leading European experts opening new collaborative possibilities of NADINE with network research on EU scale. The kick-off meeting of NADINE project will take place at 22-25 July during the workshop at ECT, Trento.

Decision making:

Decisions will be made by seeking a consensus. In the very unlikely case where this is out of reach, the coordinator has the final word.

Reporting:

Each workpackage leader will submit input for the progress reports at the end of the reporting period. The project coordinator will include them into the progress report. All the reports will be made public, or in case of need their restricted status will be negotiated with EC FET Open.

Review meetings will take place in Toulouse at month 19 and in Brussels at month 37.

Management of intellectual property
IPR Management during the project
For the success of the project it is essential that all project partners agree on explicit rules concerning IP ownership, access rights to any Background and Foreground IP for the execution of the project and the protection of intellectual property rights (IPRs) and confidential information before the project starts. Therefore, such issues will be addressed in detail within the Consortium Agreement between all project partners. The main purpose of the Consortium Agreement is to establish a legal framework for the project in order to provide clear regulations for issues within the consortium related to the work, IP-Ownership, Access Rights to Background and Foreground IP for the duration of the project and any other matters of the consortium's interest.

Access Rights to Background and Foreground IP during the project
In order to ensure a smooth execution of the project, the project partners agree to grant each other royalty-free access rights to their Background and Foreground IP for the execution of the project. Any details concerning the access rights to Background and Foreground IP for the duration of the project will be defined in the Consortium Agreement.

IP Ownership
Foreground IP shall be owned by the project partner carrying out the work leading to such Foreground IP. If any Foreground IP is created jointly by at least two project partners and it is not possible to distinguish between the contributions of each of the project partners, such work will be jointly owned by the contributing project partners. The same shall apply if, in the course of carrying out work on the project, an invention is made having two or more contributing parties contributing to it, and it is not possible to separate the individual contributions. Any such joint inventions and all related patent applications and patents shall be jointly owned by the contributing parties. Any details concerning the exposure to jointly owned Foreground IP, joint inventions and joint patent applications will be addressed in the Consortium Agreement.

Consortium Agreement
The purpose of the Consortium Agreement is to establish a legal framework for the project in order to provide clear regulations for issues within the consortium related to the work, IP-Ownership, Confidential Information, Access Rights to Background and Foreground IP for the duration of the project and any other matters of the consortium's interest.

2.2 Individual participants

Participant 1 – CNRS – Toulouse, France

The CNRS research unit involved in the project is the 'Laboratoire de Physique Theorique' (LPT), UMR 5152 of CNRS. This research unit is not a legal entity. It is a Joint Research Unit

(JRU) managed both by CNRS and Universite Toulouse III Paul Sabatier (UPST). According to an agreement between CNRS and UPST, the contracts involving LPT are managed by CNRS. Please see the 'Third parties" section under part B2.4 for further details. Universite Paul Sabatier is the leading university in the south of France with about 27 thousands students. Laboratoire de Physique Theorique (LPT) du CNRS at Universite Paul Sabatier, Toulouse has 20 permanent members and about 7 post-docs and 10 PhD students. It is a part of the federal institute IRSAMC which includes in total 4 Labs: Collisions Agregats Reactivite, Chimie et Physique Quantiques, Physique et Chimie des Nano-Objets, LPT.

Coordinator and node leader is Dima Shepelyansky : directeur de recherche du CNRS (DR1), leader of quantware group, 55 years old, born 1956 Novosibirsk, USSR; male; citizenship: Russia, France; web page: **www.quantware.ups-tlse.fr/dima**; graduated from Novosibirsk State Univ. with distinction(1978); PhD supevised by B.V.Chirikov (1982), Russian habilitation (1989). His research directions include: **1) classical and quantum chaos**, chaos in nonlinear chains and classical Yang-Mills fields, microwave ionization of hydrogen and Rydberg atoms, nonlinear chains, classical and quantum synchronization, synchronization and chaos in planetary systems and rings. quantum fractals and fractal Weyl law; **2)** effects of **interactions** on Anderson localization, destruction of Anderson localization by weak nonlinearity, mesoscopic transport **and** quantum chaos, microwave control of electron **transport in mesoscopic systems,** deterministic ratchets in asymmetric nanostructures, dynamics of cold atoms and Bose-Einstein condensates (BEC) in **optical lattices**, Frenkel-Kontorova model and Wigner crystal in a perodic potential; **3) quantum computing** algorithms for classical and quantum chaos, imperfections effects for quantum computing; **4)** information retrieval, spectral properties of **Google matrix**, 2DRanking **of** Wikipedia articles and **directed networks**. **He has 202 scientific publications** (189 are visible by ISI for all spellings "Shepel*nsk*" with 4681 citations and H-index 36) including 44 Phys. Rev. Lett., 1 Phys. Rep., 1 lecture at Nobel Symposium; 11 popular articles including contributions to 2 books. He is co-organizer of 8 International conferences/workshops including E.Fermi Summer School (2005) and Programme at the Inst. H.Poincare on quantum computers memorized in video (2006). He was node PI of 2 EU grants, coordinator of EU FET project EDIQIP, PI of ARO/NSA/ARDA USA grant, node PI of 2 ANR PNANO . He was elected outstanding APS referee (2008), member of Editorial Board of Nonlinearity (1993/96), Phys. Rev. E (2011/13), Scholarpedia Editor for Quantum Chaos. **He supervised or co-supervised** 2 habilitations (K.Frahm and B.Georgeot, team members); 11 PhDs, 8 post-doc => 10 of them are permanent researchers/professors in physics. Klaus Frahm (Prof. UPS) and Bertrand Georgeot (DR2 at CNRS) are key persons of NADINE project. The group includes also Gabriel Lemarie (CR2) and Vivek Kandiah (PhD student CNRS/Region Midi-Pyrenees), it has close collaboration with A.Chepelianskii (U. Cambridge), L.Ermann (TANDAR CNEA Buenos Aires), O.Giraud (Orsay), O.Zhirov (Budker Inst of Nuclear Physics, Novosibirsk). It collaborates with members of SME company www.nomao.com at Toulouse, who may participate providing network data sets without any additional cost for the project (NOMAO contact: L.Candillier, S.Phan). The group results directly related to NADINE project are published in [11-16,23,24,59,64].

Participant 2 – UTWE – University of Twente, Enschede, Netherlands

The university's motto is `High tech, human touch'. Some 3,300 scientists and other professionals working together on cutting-edge research, innovations with real-world relevance and inspiring education for more than 9,000 students. The enterprising university encourages students to develop an entrepreneurial spirit, organizes numerous activities for secondary schools and is a partner of Kennispark Twente. The Stochastic Operations Research (SOR) group at

the Faculty of Electrical Engineering, Mathematics and Computer Science focuses its research and education on stochastic processes and their applications in the analysis of stochastic networks in telecommunications and logistics. With six staff members and eight PhD students and post-docs, the groups provides a motivating environment for innovative research. Dr. Nelly Litvak, the node project leader, is working on complex stochastic networks within the SOR group. We will also involve Dr. Werner Scheinhardt, expert in Markov processes and importance sampling. UTWE has regular collaboration with Dr. Konstantin Avrachenkov (INRIA Sophia Antipolis), Dr. Mariana Olvera-Cravioto (Columbia University), Prof.Dr. Remco van der Hofstad (Eindhoven University of Technology) and Dr. Yana Volkovich (Barcelona Media). The project will be embedded in the Centre for Telematics and Information Technology (CTIT) at the UT, the research school Business Engineering and Technological Applications (BETA), and the Institute for Applied mathematics in the 3TU Federation of the three Universities of Technology in the Netherlands (3TU.AMI).

Node leader: Dr. Nelly Litvak obtained her PhD in 2002 from Eindhoven University of Technology; web page **http://wwwhome.math.utwente.nl/~litvakn/**. She received Stieltjes Prize for the best PhD thesis in mathematics. From 2002 she is an Assistant Professor at the University of Twente. She has worked on a large variety of topics including performance analysis of warehousing systems, queuing, and healthcare logistics. From 2004 she has been working on the analysis of complex networks, in particular, on ranking algorithms and Web search. She is a laureate of a Meervoud grant `Ranking of nodes in complex stochastic networks', 2005-2009. She supervised the PhD research of Dr. Yana Volkovich, who defended her thesis in April 2009. Nelly Litvak has more than 50 publications in top journals and conferences. She is member of programme committees and invited speaker at many conferences, most recent are Workshop on Algorithms and Models for the Web Graph (WAW2011, PC); COST meeting 2011 and INFORMS Applied Probability Society conference 2011 (invited speaker). She is a managing editor of the *Internet Mathematics* journal. Nelly speaks English, Russian, Dutch and basic Bengali. She is also author of the book `Our sweet teenagers' (Alpina-non-fiction, Moscow, 2010, in Russian), of which more than 2000 copies are sold in Russia. Her publications related to NADINE project are [7,19-22,52,56,58].

Participant 3 – MTA SZTAKI – Magyar Tudomanyos Akademia, Szamitastechnikai es Automatizalasi Kutatointezet Budapest, Hungary

Organization description: As an EU Centre of Excellence, member of ERCIM and W3C, the leading national institute for computer science, software engineering and applied mathematics, the institute has a staff of over 70 PhD's and another more than 100 BSc's. Project tasks are accomplished by the Research Group for Data Mining and Web Search (http://datamining.sztaki.hu), a team of 6 senior researchers, 3 post-docs, 10 doctorate students and 5 developers. The Group was founded in 2000 for R&D in breaking technologies. We specialize in data mining for community and link analysis, custom solutions for extreme large systems (large Intranets, high traffic portals) as well as for languages with particularly complex syntax in collaboration with computational linguistic groups. In 2007 the Group achieved First Prize on the prestigious KDD Cup, a competition involving the best data mining groups around the world. Several of our former PhD students work now at the research centres of the top internet search companies (Google, Yahoo).
Previous experience related to NADINE tasks: Our major software products include a customer relation management software capable of visualizing the connection between entities (persons, objects, contracts) as well as mining hidden relations, a set of data mining software components (Clustering, classification, similarity search; clickstream analysis, frequent patterns, association rules, Hidden Markov models; storage framework with compression for 70-

90% size reduction over general purpose methods) that can be flexibly interconnected via a graphical interface, as well as is a search engine with integrated linguistic tools (http://search.sztaki.hu) for the Hungarian language that serves the Intranet of major national companies. Selected results of relevance to the proposal: KDD Cup 2007 Taks 1 winner solution, a recommender algorithm; KDD Cup 2009 result among the top results presented at KDD 2009, a machine learning method to classify users by marketing aspects; Web spam filtering results, participation in Web Spam Challenges; Participation at ImageCLEF 2007-2009, TRECVID 2009; Several publications on mining the Web graph, telecommunications and friendship networks.
Node leader: András Benczúr (web page http://datamining.sztaki.hu/, http://www.sztaki.hu/) is the head of Informatics Laboratory hosting the Research Group for Data Mining and Web Search with near 30 doctoral students, post-docs and developers. Benczúr received his Ph.D. at the Massachusetts Institute of Technology in 1997, since then his interest turned to Information Retrieval and Web Search. He was representing SZTAKI as principal investigator in several EU and national R&D projects. His research on spam filtering and low space approximations for very large scale Web analysis was awarded by a Yahoo Faculty Research Grant in 2006. Group publications related to NADINE project are [30,33-37,40-42,63].

Participant 4 – UMIL – University of Milano, Italy

Organization description: the Dipartimento di Scienze dell'Informazione (Department of Information Sciences) of the Università degli Studi di Milano, Italy, is one of the first departments created in Italy for studies related to computer science. It has several areas of expertise, in particular in the fields of web algorithmics, network analysis, formal-language theory, automata theory, image processing, and constructive logic. The DSI hosts the LAW (Laboratory for Web Algorithmics), which investigates a large spectrum of phenomena related to the web and social networks. Beyond scientific papers, the LAW gathers datasets that can be publicly distributed, and creates open-source software for large-scale analysis. Several papers published on the major journals and conference of the area have used software or datasets provided by the LAW.
Node leader: Sebastiano Vigna, born 1967 in Milano, Italy; 1996: Ph.D. in Computer Science, Università degli Studi di Milano, Italy; 1996–2003: Assistant professor, DSI; 2003–present: Associate professor, DSI, Università degli Studi di Milano (head of the LAW); web page **http://vigna.dsi.unimi.it/** .Expertise: web graphs, social networks, web algorithmics, large-scale analysis, web-graph compression, social-network compression, efficient data structures and algorithms for large datasets, high-performance web crawling. Past work related to the proposal includes the development of a high-performance crawler that has been used to gather large (100 million pages) datasets, in particular for the DELIS (Dynamically Evolving Large-scale Information Systems) European FP6 project; development and distribution of algorithms and Java libraries for the compressed in-memory representation of web graphs and social networks, which makes it possible to study very large networks in main memory; development of highly scalable algorithms for computing the neighbourhood function (and thus the distance distribution) of large graphs; the first accurate analysis of the behaviour of PageRank when the damping factor changes [6]; development of a complete, high-performance, open-source indexing engine, MG4J. Group publications related to NADINE project are [27-29, 38, 48, 61, 62, 65, 66].

2.3 Consortium as a whole

The consortium is composed of four groups. One of our main assets is that all of us have com-

plementary skills. In particular, the leader of WP1 has a very strong mathematical and computer science background. The leaders of WP2, WP6 and WP7 are experts from theoretical physics and specialists in the study of complex classical and quantum systems and directed networks analysis. The leader of WP3 and WP5 has expertise in computer science and large database crawling. The leader of WP4 is expert in large scale distributed data analytics. The interdisciplinary expertise of partners in mathematics, physics and computer science will create new solutions for important tasks of the NADINE project with European and worldwide applications.

The need for a European (rather than a national or local) approach
Given the international scope of the media market, the complexity and variety of the technologies concerned, the dominance of American companies in web-based information searching, and the pre-eminence of the US computer industry, it is clear that no single European country or manufacturer can compete alone in the field of IT productivity tools or just understanding the complex processes hidden in social networks.
On the other hand, this proposal has the ambitious goal of identifying novel methods for network analysis by leveraging on the complementary expertise of the partners.
Moreover, we truly believe that this would not be possible at a national level because the expertise breadth needed is large and no country has research groups at the state of the art for all of it. For this reason, each group is involved in very specific tasks that make use of their expertise.

i)Sub-contracting: no specific subcontracting

ii) Other countries: n. a.

2.4 Resources to be committed

The total budget is summarized in the forms A. The bulk of the budget will be spent on personnel. In total, we estimated the total manpower needed for a successful completion of the program at 176 person months. The personnel will consist of senior researchers, postdocs and PhD students. In addition a decent amount of travel costs has to be covered, enabling strong interactions within the consortium but also those with the outside theoretical and experimental scientific community. Both types of co-operations and collaborations will be strengthened by a small annual workshop.

The groups have already at their disposal the equipment and administrative support available for a successful achievement of the project, and this will complement the requested contribution. This includes the necessary office space, computer facilities and access to scientific journals and databases. Thus MTA_SZTAKI posesses a high performance cluster consisting of 200 CPU cores, 400GB RAM and 100TB storage that SZTAKI provides available for the NADINE consortium to perform large-scale distributed experiments. CNRS has access to high-performance supercomputers CALMIP (Univ. P.Sabatier, Toulouse) and IDRIS (Orsay). No specific additional equipment is requested for this project. The management part of the budget is kept to its absolute minimum.

Personal to be hired include: 1 PhD student for three years (UTWE), 1 postdoc for two and half years (CNRS), 1 postdoc for two years (MTA_SZTAKI) and 1 postdoc for two years (UMIL).

The Consortium will use as **background** the following assets in the project.

A significant set of high-quality software tools have been developed for the analysis of very large graphs by members of the consortium. In particular this includes:

- Fastutil, a collection of high-performance classes for managing large collections in

Java (UMIL).
- The WebGraph framework, a toolkit for storing and accessing efficiently graphs in a compressed representation that builds on fastutil. The framework will be used, for example, to perform iterative computations on very large graphs (UMIL).
- MG4J, a full-text, large-scale search engine that has been used in several research projects because of its efficiency and its extremely open design (UMIL).
- SZTAKI has home developed recommender components used for data mining contests that will be used as baseline for NADINE developments.

The consortium has at its disposal a significant set of ready-to-use datasets including web graphs of various sizes, social networks graphs (e.g., Twitter, DBLP, etc.), road graphs, telephone call graphs from Hungarian Telekom, insurance customer data from AEGON Hungary, and so on. In particular, UMIL gathered a series of monthly snapshots of the .uk domain, comprising more than 1 billion pages that have been combined into a single time-aware labelled graph.

OTHER DIRECT COSTS:

CNRS: Travel costs during 3 years: kick-off meeting (Trento July 2012), consortium and review meetings, and conferences for 4 permanent members of the group (K.Frahm, B.Georgeot, G.Lemarie, D.Shepelyansky), and post-doc to be hired with about 800euro per researcher per year. Total: 12000euro

UTWENTE: Laptop for a PhD student: 1500 euro. Travel costs: 2,5 kEuro per person per year for attending Trento workshop (kick-off meeting July 2012), consortium and review meetings, and conferences (15000euro for 2 persons, N.Litvak and PhD student, for 3 years). Total: 16500 euro

MTA_SZTAKI: travel costs: 1611,2 Euros per person per year for attending Trento workshop, consortium and review meetings, and other conferences (two persons involved). Total for two person for three years: 9667 EUR. Total: 9667euro

UMIL: The direct costs incurred by the UMIL partner (€105000) are divided mainly intravel costs (20000euro) and hardware costs (85000euro). Travel costs include Trento kick-off meeting, consortium and review meetings, and other conferences (three permanent members of the group and post-doc involved).

The hardware will be used for two important goals of the project:

- Developing a new distributed crawler;

- Performing analysis of social networks and their metrics at an unprecedented scale.

The hardware is thus divided in two parts: a large server (>=1/2T RAM, >20 cores) for multicore, shared-memory algorithms and 8 smaller units mainly aimed at distributed, message-passing algorithms. Overall, we aim also at >20T of disk capacity for storing the data resulting from crawling activity. Total: 105000euro

Third parties:

A) the valid UMR5152 CNRS-UPST agreement in place means that work carried out by the Laboratoire Physique Theorique du CNRS/UPST is under direct supervision, management and control of CNRS (in agreement with Université Paul Sabatier, Toulouse (UPST)).

B) cost declarations to be made by CNRS and the costs will be recorded in the accounts of CNRS and in addition no costs incurred by UPST.

C) therefore no special clause 10 in this case, only a third party making resources available to a beneficiary (CNRS).

Section 3: Impact

3.1 Transformational impact on science, technology and/or society

The new tools and algorithms created by the project will produce **broad scientific and technological impact** on analyzing **modern directed networks**, and give n**ew recipes for their skilful design and development**. As concerns the **European dimension**, new types of network analysis technologies will be proposed that are applicable for ranking in search engines, analyzing human behaviour in social networks, identify the opinion leaders as well as malicious or selfish actors.

NADINE research aims at increasing the **scale** and **scope** of network analysis to a global level of understanding the Web, the social media, and in a final turn, the society. Understanding and improving mechanisms behind Web ranking and spam detection is highly urgent in society where current clever algorithms still can be cheated, and, in times of global economic crisis, anonymous spammers gain considerable incomes from web link trading (see e.g. recent *New York Times* articles: D. Segal. The dirty little secrets of search. The New York Times, February 13, 2011; D. Segal. A bully finds a pulpit on the web. The New York Times, November 28, 2010). With increasing ability of users to share, to create and to consume content it is extremely important to understand the dependencies between the users in social networks and social media, how the users organize themselves and whose opinion they trust. The recent example of the Arab spring that brought Twitter from on-line world into off-line world could demonstrate the importance of this study. By novel means of analyzing networks, in communication, social networks, and collaboration, we reach closer to the goal of exploring social life on Earth.

Our results will help to create European answer on the current situation when the market is dominated by search engines, social media and networking services from USA, China, and Russia. Our results will be applicable to Web scale classification and information retrieval.

The project is based on interdisciplinary expertise of partners in mathematics, physics and computer science with the cross-fertilization of different fields of science bringing qualitatively new solutions.

A project of this kind requires competence beyond the national scale – only a strong international (European) collaboration can bring in a sufficiently broad spectrum of expertise. The participating groups, located in the top research institutes of four European countries, are in all respects ideally positioned for this enterprise.

The scale of the Internet, and of web-based content, has become almost unimaginable. Between one and two-and-a-half billion people are connected, which is expected to grow to five billion by 2015. Five hundred million are expected to have mobile broadband in 2010, and Internet traffic has increased twenty-fold times in the past five years. Today there are more than 230 million Web servers, with something over twelve billion static Web pages. The number of dynamic pages is uncountable and literally unbounded. In this context of fast pace of change, the main driver today is *social media* and social networks.

Social media is the new buzzword of the Internet and the main effect behind the activity in social networks. Social media is a comparatively new term, which covers a wide range of online applications, platforms and media that support on-line interaction, collaboration and

the sharing of content. It includes all what constitutes "human interaction in a virtual world". Social media has grown explosively in a very short time, to become a new communication channel and a medium that will have a profound impact on advertising, publicity, marketing and opinion forming, as well as entertainment. NADINE wants to analyze and extend that impact to the world of the knowledge industry.

Social media is transforming both the scale and the nature of the Web. The vast majority of users are producing content and there is an accelerating transition from consumption towards participation and production. The scale and speed of growth of social media since 2006 can be illustrated in a few figures[1]. More than 80% of active Internet users worldwide watch video-clips on-line, and over 70% have read a blog. An extraordinarily large, and growing, number of people are producing, uploading and sharing media content. In 2008, 55% of the user sample uploaded photos; 23% uploaded videos, almost three-quarters of whom uploaded video at least once a week. There has been a real shift from passive media consumption to active choice and control over the media experience, moving in the direction of personal content creation.

Facebook and MySpace are just two of the largest and best-known services that use social media to establish explicit social networks. They are the most visible proof of the development of the Internet as a space for inter-personal interaction: in the study already mentioned, people maintained almost as many friendships on-line through social sites (thirty on average) as face-to-face (thirty-five).

Other websites establish a rather different kind of implicit social network between users who exchange data and opinions on the basis of shared interests. E-mail address books, instant messenger logs and wikis all point to deeper, wider and less obvious implicit social networks that offer immense information resources[2]. Even if the most popular explicit social networks, such as Facebook, are still dominated by the under thirty-fives, social networks as a whole engage the whole age spectrum. Many of the internet's early adopters are now in their sixties, and social networks of professionals including Linked In, Xing and Viadeo, has a more mature demographic.

Companies use the media to communicate messages about their brands and services to consumers: with social media, every one of these consumers not only gets the message from the company, but opinions from their peers on various social networks. Social media are highly dynamic and can react almost instantaneously. Social media gives power to the crowd with a multimodal, electronic, 'word of mouth', with a very high level of inter-personal interaction. The gossip and exchange of opinion that takes place at the bar, the coffee machine at work or over the garden fence now can happen around the world. Almost every commercial website encourages consumer input and interaction, with opinion ratings, feedback and voting opportunities, and the 'blogosphere' rivals all the other mass media in terms of reach, cultural, social and political impact. Within the Babel of social networks, though, how do we distinguish between the 'wisdom of the masses', expert opinion, covert advertising, deliberate misinformation or bias? Which voices can be trusted – and which have influence? Age is much less significant than might be imagined, and the evidence is that we trust online strangers almost as much as face-to-face recommendations from people we know. Paid-for

1 Taken from a study by Universal McCann *When did we start trusting strangers"?* September 2008. The study was based on a sample of 17,000 'active internet users' aged from sixteen to fifty-four in twenty-nine countries between September 2006 and March 2008. http://www.universalmccann.com/Assets/strangers_report_LR_20080924101433.pdf.

2 Leskovec J. and Horvitz E. *Planetary-scale Views on a Large Instant-Messaging Network*, Proceedings of WWW2008, Beijing, present a study of data from a month of communication across the whole of the Microsoft Messenger system. It examines characteristics and patterns emerging from 30 billion conversations among 240 million people at 180 million computers.

communications, advertising, and celebrity endorsements count for comparatively little. This is the world that NADINE will uncover.

The new algorithms developed by NADINE project will allow to perform qualitatively new ranking in various types of directed networks. The possible applications includes not only the WWW but also various other types on directed networks. The example of the World Trade Network [16] shows that 2DRanking is very natural for monetary and trade flows with CheiRank and PageRank being analogous to Export and Import flows. Thus 2DRanking will allow to construct a network map with important nodes linked to a given node. The 2D map representation of information flows in WWW, Wikipedia, World Trade, software networks like Linux, gene networks, neural brain networks and other networks will provide useful and efficient 2D graphical image of nodes with their neighbours. The understanding of actual roots of information flow on direct networks will put the foundation for further progress in e-voting and e-government of modern democratic society. This will also provide new efficient tools for commercial use of web and social networks including viral marketing and advertising. New ranking algorithms will give foundations for new search engine development in Europe making it competitive with USA, China and Russia. The project progress on this Roadmap will be regularly assessed and discussed in the reviewing reports.

The real impact of NADINE is beyond search, social media and social networks per se. NADINE wants to do a qualitative step forward in the capacity of European industry to exploit the new interaction paradigms that are behind digital social media and networks, through new insights stemming from the complementary expertise of NADINE partners.

3.2 Contribution at the European level towards the expected impacts listed in the work programme

The broader vision behind NADINE arises from the steadily increasing ability to design and manipulate information flow on the World Wide Web and other modern directed networks, which produce enormous influence on human society. This has and will have considerable impact not only within the scientific community.

In the foreseeable future society on the whole will profit more and more from the increasing capability of actively exploring network communication and information retrieval. NADINE complements the mainstream projects undertaken within EU FET ICT. The present proposal aims to enhance the European position in the development of Web content and social media services which play a strategic role in the modern highly communicative society.

Our applications for voting, leader identification and noise removal catch up with the speed at which new problems and opportunities are arising in our changing world as consequences of globalization, technological, demographic and environmental change, and make a contribution to strengthening our societies' adaptivity, resilience, and sustainability.

Our goal aligns with the goals of the FET flagship project FuturICT in understanding and managing complex, global, socially interactive systems. The NADINE consortium builds on our expertise in complexity science, theoretical physics, mathematics, computer science and ICT.

In NADINE we devise tools to handle big data. Our algorithms will be tested on Web scale data of Terabytes with special emphasis on parallel and cloud computing environments.

3.3 Dissemination and/or use of project results

The results obtained by the participants will be made accessible to a wide scientific community mainly by their publication in top international peer-reviewed scientific journals. A website of the project will be established and maintained. It shall include up-to date information about the project goals and a database of all publications that will be thus made freely available to the public. The members of the NADINE consortium will also present their results at the major international scientific conferences and workshops in the fields of computer science, physics and mathematics.

A major dissemination measure within NADINE will be the organization of a **workshops about spectral properties** of directed networks. In addition, members of the Consortium are active members of **program committees** for the major conferences in network analysis, Web information retrieval and data mining where they have the opportunity to initiate workshops; special sessions in NADINE related areas. Some selected previous activities include: Andras Benczur serving as WWW Workshop Chair (2011), WSDM senior PC (2012) and ECML/PKDD 2010 Discovery Challenge organizer; Sebastiano Vigna serving as WWW Workshop Co-chair (2011), WWW Search Track PC member (2006-present); Paolo Boldi serving as WSDM Co-chair (2010) and WWW "Social Systems and Graph Analysis" Track Co-chair (2011); Nelly Litvak serving as WAW workshop Co-chair (2009), WAW PC member (2011,2012), PC member of Complex'2009, managing editor of the Internet Mathematics journal.

The **data** gathered by the project and the tools developed for network analysis will be made available to the public under open-source licenses. All privacy of these data sets will be preserved (e.g. no private node names will be made public, only mathematical structure of the network and its links will be make publicly available, see also direct statement below). Distributing significant, large-scale datasets is a fundamental task that makes replication of results possible. Moreover, research groups lacking the infrastructure that is necessary to gather large-scale datasets will benefit from their availability. **Open source software libraries** will provide basic tools for social network data mining and for building large-scale applications. Researchers in the field currently miss such tools that need to be built case by case with in-house solutions.

We specially stress the following points:

Following the best practice in ethics, all data used in the project will be automatically anonymized and used for experimentation and prototype testing. No private information shall be disclosed to third parties. The data may include, but is not limited to, personal information about the user such as name, date of birth, location, images, or relations to other users. Should personalised data be available, these will be anonymized and the data protection act will be followed.

In order to provide products and services, online companies collect and store information from user account registration and site usage. A general, widely followed policy is to de-identify user log data within 18 months of collection, with limited exceptions to meet legal obligations. Certain types of log data such as ad views, ad clicks, page views and page clicks are usually retained for a longer period in order to power innovative product development, provide personalized and customized services, and better enable security systems to detect and defend against fraudulent activity.

We are not planning to collect new types of data during the project and will not disclose existing sensitive datasets to third parties. For members of the consortium, the data will be made available under restrictions typical in collaborations with academia, i.e., research with private data will have to be conducted on the infrastructure of the data owner.

Research organization will conduct experimentation and prototype testing only over anonymized data sets, following the standard procedures used in these organizations. Data sets and access to partner testbeds will be regulated by bilateral NDAs signed separately in addition to general rules in the CA. Finally, NADINE will observe European legal regulations concerning privacy. This is at a policy level, and will be monitored and reinforced by the Coordinator of NADINE.

We also plan to use NADINE data in organizing international **data mining contests** based on our experience with the Web Spam Challenges 2007 and 2008 in conjunction with the World Wide Web Conference where UMIL provided Web data and the ECML/PKDD 2010 Discovery Challenge where SZTAKI was main organizer in collaboration with Internet Memory as data provider and Yahoo! Research and Google as co-organizers.

The consortium will also leverage on SZTAKI expertise on implementing research results in industrial prototypes. Our good connections to the search industry also enable industrial exploitation of the results.

Partners will disseminate the results of NADINE by promoting them to projects under FP7 related to this proposal such as FuturICT, LivingKnowledge, WeKnowIt, MODAP, and others. We also envisage the use of our own tools for attracting users to the project and to disseminate the results of the project through the use of existing social networks.

Networked information has an atypical set of business models within the IT sector insofar as users expect the basic technology to be provided for free. The consumer is prepared to accept a certain amount of non-intrusive advertising in return for an improved service: in the best of cases, targeted advertising may become a positive benefit and more direct route to better services. In the service domain, we distinguish between four different approaches that we will consider:

- Services based in targeted advertising, which both reduce the annoyance of unwanted advertising and increase the return on advertising expenditure;
- Business to customer services, that use social media search technology to deliver new and better information or content to customers, either on a paid subscription or advertising model;
- Business to business services, based on media and social mining, that enable service providers to offer more and better services; and
- Consulting services for enterprises as well as advertisers and marketers, to help them understand different groups and customers, and communicate more effectively.

The financial stakes behind social media and social networks alone are very high. Advertisers, though, are still uncertain about how to take advantage of user-generated content and social media to promote their brands, and wary about taking risks. The potential economic impact is therefore very high, if the social media and marketing industries can begin to understand the

medium, and how to monetise its use. The impact will be felt not only on existing companies in the group but through the opportunities to create new businesses. The ability precisely to identify a precise interest group, target audience or skills group, to communicate with it directly, and understand how its members respond, will open the way to creating new on-line services and service-based companies.
One of the first direct impacts of NADINE may be in viral marketing and advertising, through the better understanding of social networks as a medium. In fact, spreading of influence and opinions through social networks is an important contribution to the success of initiatives whose success partly relies in the dissemination to a social network community. We believe that our research can contribute to create new ways to market and advertise products. In fact we have already contacted two companies that are interested in the results of this proposal to measure the impact of social networks on marketing and advertising.

Section 4: Consideration of gender aspects

The pursuit of scientific knowledge and its use in service to society requires the talent, perspectives and insight that can only be assured by increasing diversity. Therefore, an equal representation of women and men at all levels of STRIDES project is encouraged.

One of the priority identified in the European Commission adopted "Roadmap for Equality between Men and Women 2006-10" is to promote equal participation of men and women in decision making. Gender aspects in research have a particular relevance to the project's topic as risk factors, behaviour, causes, consequences, management, social responsibility may differ in men and women. Furthermore, roles and responsibilities, the relationship to the resources such as energy and the perception of risk and benefits may have a gender dimension.

The European Commission recognises a relationship between women and research:

o Women's participation in research must be encouraged both as scientists/technologists and within the evaluation, consultation and implementation processes; thus female researchers should be invited to participate in all project's activities,

o Equal opportunities will be promoted in recruitment at all levels

o Research must address women's needs, as much as men's needs,

o Research must be carried out to contribute to an enhanced understanding of gender issues.

The NADINE Project Coordinator will make special efforts to be as gender-balanced as possible, and, in addition, will monitor the impact of the activities on gender balance and gender-related issues. It will be ensured that women will be equally represented in events organised by or related to the project. Throughout the project the Project Coordinator with the help of the Project Management support team will collect gender statistics on the workforce and will monitor the progress made in terms of gender balance.

The NADINE consortium feels that it is important to have a female influence in the project management structure, in order to decide research direction and other policy decisions and make sure the project addresses both women's and men's needs. This helps to ensure the NADINE activities and project outputs are less/not gender biased. The consortium will try to ensure around 25% female members within the project management structure. It is noted that NADINE has 1 woman WP leader, Nelly Litvak.

Section 5: Ethical Issues ETHICAL ISSUES TABLE

	YES	PAGE
Informed Consent		
●Does the proposal involve children?		
●Does the proposal involve patients or persons not able to give consent?		
●Does the proposal involve adult healthy volunteers?		
●Does the proposal involve Human Genetic Material?		
●Does the proposal involve Human biological samples?		
●Does the proposal involve Human data collection?		
Research on Human embryo/foetus		
●Does the proposal involve Human Embryos?		
●Does the proposal involve Human Foetal Tissue / Cells?		
●Does the proposal involve Human Embryonic Stem Cells?		
Privacy		
●Does the proposal involve processing of genetic information or personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)		
●Does the proposal involve tracking the location or observation of people?		
Research on Animals		
●Does the proposal involve research on animals?		
●Are those animals transgenic small laboratory animals?		
●Are those animals transgenic farm animals?		
●Are those animals cloned farm animals?		
●Are those animals non-human primates?		
Research Involving Developing Countries		
●Use of local resources (genetic, animal, plant etc)		
●Impact on local community		
Dual Use		
●Research having direct military application		
●Research having the potential for terrorist abuse		
ICT Implants		
●Does the proposal involve clinical trials of ICT implants?		
I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	X	