Extraction and Analysis of Facebook Friendship Relations

Salvatore Catanese[1], Pasquale De Meo[1], Emilio Ferrara[2], Giacomo Fiumara[1] and Alessandro Provetti[1,3]

[1] Dept. of Physics, Informatics Section. University of Messina, Italy.
[2] Dept. of Mathematics, University of Messina, Italy.
[3] Oxford-Man Institute, University of Oxford, UK.

Summary. Online Social Networks (OSNs) are a unique Web and social phenomenon, affecting tastes and behaviors of their users and helping them to maintain/create friendships. It is interesting to analyze the growth and evolution of Online Social Networks both from the point of view of marketing and offer of new services and from a scientific viewpoint, since their structure and evolution may share similarities with real-life social networks. In social sciences, several techniques for analyzing (offline) social networks have been developed, to evaluate quantitative properties (e.g., defining metrics and measures of structural characteristics of the networks) or qualitative aspects (e.g., studying the attachment model for the network evolution, the binary trust relationships, and the link prediction problem). However, OSN analysis poses novel challenges both to Computer and Social scientists. We present our long-term research effort in analyzing Facebook, the largest and arguably most successful OSN today: it gathers more than 500 million users. Access to data about Facebook users and their friendship relations, is restricted; thus, we acquired the necessary information directly from the front-end of the Web site, in order to reconstruct a sub-graph representing anonymous interconnections among a significant subset of users. We describe our ad-hoc, privacy-compliant crawler for Facebook data extraction. To minimize bias, we adopt two different graph mining techniques: breadth-first search (BFS) and rejection sampling. To analyze the structural properties of samples consisting of millions of nodes, we developed a specific tool for analyzing quantitative and qualitative properties of social networks, adopting and improving existing Social Network Analysis (SNA) techniques and algorithms.

1 Introduction

The increasing popularity of Online Social Networks (OSNs) is witnessed by the huge number of users that MySpace, Facebook etc. acquired in a short amount of time. The growing accessibility of the Web, through several media, gives to most users a 24/7 online presence and encourages them to build a online mesh of relationships.

As OSNs become the tools of choice for connecting people, we expect that their structure will increasingly mirror real-life society and relationships. At the same time, with an estimated 13 millions transactions per seconds (at peak) Facebook is one of the most challenging computer science artifacts, posing several optimization, scalability and robustness challenges.

The essential feature of Facebook is the friendship relation between participants. It consists, mainly, in a permission to consult each others' friends list and posted content: news, photos, links, blog posts, etc; such permission is mutual. In this chapter we consider the Facebook friendship graph as the (non-directed) graph having FB users as vertices and edges represent their friendship relation.

The analysis of OSN connections is a fascinating topic on multiple levels. First, a complete study of the structure of large *real* (i.e., off line) communities were impossible or at least very expensive before, even at fractions of the scale considered in OSN analysis. Second, data is clearly defined by some structural constraints, usually provided by the OSN structure itself, w.r.t. real-life relations, often hardly identifiable. The interpretation of these data opens up new fascinating research issues, e.g., is it possible to study OSNs with the tools of traditional Social Network Analysis, as in Wasserman-Faust [89] and [69]? To what extent the behavior of OSN users is comparable to that of people in real-life social networks [39]? What are the topological characteristics of the relationships network (friendship, in the case of FB) of OSN [4]? And what about their structure and evolution [58]?

To address these questions, further Computer Science research is needed to design and develop the tools to acquire and analyze data from massive OSNs. First, proper social metrics need to be introduced, in order to identify and evaluate properties of the considered OSN. Second, scalability is an issue faced by anyone who wants to study a large OSN independently from the commercial organization that owns and operate it. For instance, last year Gjoka et al. [42] estimated the crawling overhead needed to collect the whole Facebook graph in 44 Terabytes of data. Moreover, even when such data could be acquired and stored locally (which however raises storage issues related to the social network compression [17, 16]), it is non-trivial to devise and implement effective functions that traverse and visit the graph or even evaluate simple metrics. In literature, extensive research has been conducted on sampling techniques for large graphs; only recently, however, studies have shed light on the bias that those methodologies may introduce. That is, depending on the method by which the graph has been explored, certain features may result over/under-represented w.r.t. the actual graph.

Our long-term research on these topics is presented in this book chapter. We describe in detail the architecture and functioning modes of our ad hoc Facebook crawler, by which, even on modest computational resources, we can extract large samples containing several milions of nodes. Two recently-collected samples of about 8 millions of nodes each are described and analyzed in detail. To comply with the FB end-user licence, data is made anonymous

upon extraction, hence we never memorize users' sensible data. Next, we describe our newly-developed tool for graph analysis and visualization, called LogAnalysis. LogAnalysis may be used to compute the metrics that are most pertinent to OSN graph analysis, and can be adopted as an open-source, multiplatform alternative to the well-known NodeXL tool.

2 Background and Related Literature

The task of extracting and analyzing data from Online Social Networks has attracted the interest of many researchers e.g. in [7, 39, 93]. In this section we review some relevant literature directly related to our approach.

In particular, we first discuss techniques to crawl large social networks and collect data from them (see Section 2.1). Collected data are usually mapped onto graph data structures (and sometimes hypergraphs) with the goal of analyzing their structural properties.

The ultimate goal of these efforts is perhaps best laid out by Kleinberg [56]: topological properties of graphs may be *reliable indicators* of human behaviors. For instance, several studies show that node degree distribution follows a power-law, both in real and online social networks. That feature points to the fact that most social network participants are often *inactive,* while few key users generate a large portion of data/traffic. As a consequence, many researchers leverage on tools provided from *graph theory* to analyze the social network graph with the goal -among others- of better interpreting personal and collective behaviors on a large scale. The list of potential research questions arising from the analysis of OSN graphs is very long; in the following we shall focus on three themes which are directly relevant to our research:

i) *Node Similarity Detection*, i.e., the task of assessing the degree of similarity of two users in an OSN (see Section 2.2),
ii) *Community Detection*, i.e., the task of of finding groups of users (called *communities*) who frequently interact with each other but seldom with those outside their community (see Section 2.3)
iii) *Influential User Detection*, i.e., the task of identifying users capable of stimulating other users to join activities/discussions in their OSN (see Section 2.4).

2.1 Data Collection in OSN

The most works focusing on data collection adopt techniques of Web information extraction [34], to crawl the front-end of Websites; this because OSN datasets are usually not publicly accessible; data rests in back-end databases that are accessible only through the Web interface.

In [63] the discussed the problem of sampling from large graphs adopting several graph mining techniques, in order to establish whether it is possible to

avoid bias in acquiring a subset of the whole graph of a social network. The main outcome of the analysis in [63] is that a sample of size of 15% of the whole graph preserves the most of the properties.

In [69], the authors crawled data from large online social networks like Orkut, Flickr and LiveJournal. They carried out an in-depth analysis of OSN topological properties (e.g., link symmetry, power-law node degrees, groups formation) and discussed challenges arising from large-scale crawling of OSNs.

[93] considered the problem of crawling OSNs analyzing quantitative aspects like the efficiency of the adopted visiting algorithms, and bias of data produced by different crawling approaches was .

The work by Gjoka et al. [42] on OSN graphs is perhaps the most similar to our current research, e.g. in [22]. Gjoka et al. have sampled and analyzed the Facebook friendship graph with different visiting algorithms namely BFS, Random Walk and Metropolis-Hastings Random Walks. Our objectives differ from those of Gjoka et al. because their goal is to produce a [[[[[*consistent* sample of the Facebook graph. A sample is defined consistent when some of its key structural properties, i.e., node degree distribution, assortativity and clustering coefficient approximate fairly well the corresponding properties of the original Facebook graph. Vice versa, our work aims at crawling a portion of the Facebook graph and to analitically study the structural properties of the crawled data.

A further difference with [42] is in the strategy for selecting which nodes to visit: Gjoka's strategy requires to know in advance the degree of the considered nodes. Nodes with the highest degree are selected and visited at each stage of the sampling. In the Facebook context, a node's degree represents the number of friends a user has; such information is available in advance by querying the profile of the user. Such an assumption, however, is not applicable if we consider other online social networks. Hence, to know the degree of a node we should preliminarily perform a complete visit of the graph, which may not be feasible for large-scale OSN graphs.

2.2 Similarity Detection

Finding similar users of a given OSN is a key issue in several research fields like Recommender Systems, especially Collaborative Filtering (CF) Recommender Systems [3]. In the context of social networks, the simplest way to compute user similarities is by means of accepted similarity metrics such as the Well-known Jaccard coefficient [50]. However, the usage of the Jaccard coefficient is often not satisfactory because it considers only the acquaintances of a user in a social network (and, therefore, local information) and does not take global information into account. A further drawback consists of the fact that users with a large number of acquaintances have a higher probability of sharing some of them w.r.t. users with a small number of acquaintances; therefore, they are likely to be regarded as similar even if no real similarity exists between

them. [1] proposed that the similarity of two users increases if they share acquaintances who, in turn, have a low number of acquaintances themselves.

In order to consider global network properties, many approaches rely on the idea of *regular equivalence*, i.e., on the idea that two users are similar if their acquaintances are similar too. In [13] the problem of computing user similarities is formalized as an optimization problem. Other approaches compute similarities by exploiting matrix based methods. For instance, the approaches of [27, 61] use a modified version of the Katz coefficient, SimRank [53], provides an iterative fixpoint method. The approach of [14] operates on directed graphs and uses an iterative approach relying on their spectral properties.

Some authors studied computational complexity of social network analysis with an emphasis on the problem of discovering links between social network users [85, 86]. To this purpose, tools like Formal Concept Analysis and Matrix Factorization are described and employed in this Chapter.

To describe these approaches, assume to consider a social network and let $G = \langle V, E \rangle$ be the graph representing it; each node in V represents a user whereas an edge specifies a tie between a pair of users (in particular, the fact that a user *knows* another user).

In the first stage, *Formal Concept Analysis* is applied to map G onto a graph G'. The graph G' is more compact than G (i.e., it contains less nodes and edges of G) but, however, it is *sparse*, i.e., a node in G' still has few connections with other nodes. As a consequence, the task of predicting if two nodes are similar is quite hard and comparing the number of friends/acquaintances they share is not effective because, in most cases, two users do not share any common friend and, therefore, the similarity degree of an arbitrary pair of users will be close to 0. To alleviate sparsity, *Singular Value Decomposition* [46] (SVD) is applied. Experiments provided in [85] show that the usage of SVD is effective in producing a more detailed and refined analysis of social network data.

The SVD is a technique from Linear Algebra which has been successfully employed in many fields of Computer Science like Information Retrieval; in particular, given a matrix $\mathbf{A}$, the SVD allows the matrix $\mathbf{A}$ to be decomposed as follows:

$$\mathbf{A} = \mathbf{U\Sigma V}$$

being $\mathbf{U}$ and $\mathbf{V}$ two *orthogonal matrices* (i.e., the columns of U and V are pairwise orthogonal); the matrix $\mathbf{\Sigma}$ is a diagonal matrix whose elements coincide with the square roots of the eigenvalues of the matrix $\mathbf{AA}^T$; as usual, the symbol $\mathbf{A}^T$ denotes the transpose of the matrix $\mathbf{A}$.

The SVD allows to decompose a matrix $\mathbf{A}$ into the product of three matrices and if we would multiply these three matrices, we would reconstruct the original matrix $\mathbf{A}$. As a consequence, if $\mathbf{A}$ is the adjacency matrix of a social network, any operation carried out on $\mathbf{A}$ can be equivalently performed on the three matrices $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ in which $\mathbf{A}$ has been decomposed.

The main advantage of such a transformation is that matrices $\mathbf{U}$ and $\mathbf{V}$ are dense and, then, we can compute the similarity degree of two users even if the number of friends they share is 0.

2.3 Community Detection

The problem of detecting groups of related nodes in a single social network has been largely analyzed in the Physics, Bioinformatics and Computer Science literature and is often known as *community detection* [71, 72] and studied, among others, by Borgatti et al. [18].

A number of community detection algorithms are based on the concept of *network modularity*. In particular, if we assume that a OSN S (represented by means of a graph G) has been partitioned into m communities, the corresponding network modularity Q is defined as follows:

$$Q = \sum_{s=1}^{m} \left[\frac{l_s^2}{L} - \frac{d_s^2}{2L} \right] \tag{1}$$

where L is the number of edges in G, l_s is the number of edges between nodes belonging to the s-th community and d_s is the sum of the degrees of the nodes in the s-th community.

High values of Q reflect a high value of l_s for each identified community; in turn, this implies that detected communities are highly cohesive and weakly coupled. Therefore, it is not surprising that the idea inspiring many community detection algorithms is to maximize the function Q. Unfortunately, maximizing Q is an NP-hard [20], thus viable heuristics must be considered.

A first heuristics is the Girvan-Newman algorithm (hereafter, GN) [41]. It relies on the concept of *edge betweenness*; in particular, given and edge e of S, its edge betweenness $B(e)$ is defined as

$$B(e) = \sum_{n_i \in S} \sum_{n_l \in S} \frac{np_e(n_i, n_l)}{np(n_i, n_l)} \tag{2}$$

where n_i and n_l are nodes of S, $np(n_i, n_l)$ is the number of distinct shortest paths between n_i and n_l and $np_e(n_i, n_l)$ is the number of distinct shortest paths between n_i and n_l containing e itself.

Intuitively, edges with a high betweenness connect nodes belonging to different communities. As a consequence, their deletion would lead to an increase of Q. Algorithm GN relies on this intuition. It first ranks edges on the basis of their betweenness; then it removes the edge with the highest betweenness (Step 1). After this, it recomputes the betweenness of the remaining edges and the value of Q (Step 2). It repeats Steps 1 and 2 until it does not observe any significant increase of Q. At each iteration, each connected component of S identifies a community. The computational complexity of GN is $O(N^3)$, N being the number of nodes of S. The cubic complexity algorithm may not

be scalable enough for the size of online social networks but a more efficient $-O(Nlog^2N)-$ implementation of GN can be found in [25].

[15] proposes to apply GN on the neighborhood of each node rather than on the whole network. Once communities have been identified, the group of nodes associated with a community is replaced by a supernode, thus producing a smaller graph. This process is iterated and, at each iteration, the function Q is re-computed. The algorithm ends when Q does not significantly increase anymore.

[80] illustrates an algorithm which strongly resembles GN. In particular, for each edge e of S, it computes the so-called *edge clustering coefficient* of e, defined as the ratio of the number of cycles containing e to the maximum number of cycles which could potentially contain it. Next, GN is applied with the edge clustering coefficient (rather than edge betweenness) as the parameter of reference . The most important advantage of this approach is that the computational cost of the edge clustering coefficient is significantly smaller than that of edge betweenness.

All approaches described above use the greedy technique to maximize Q. In [48], the authors propose a different approach which maximizes Q by means of the simulated annealing technique. That approach achieves a higher accuracy but can be computationally very expensive.

[75] describes *CFinder*, which, to the best of our knowledge, is the first attempt to find overlapping communities, i.e., communities which may share some nodes. In CFinder communities are detected by finding cliques of size k, where k is a parameter provided by the user. Such a problem is computationally expensive but experiments showed that it scales well on real networks and it achieves a great accuracy.

The approach of [73] uses a Bayesian probabilistic model to represent an online social network. An interesting feature of [73] is the capability of finding *group structures*, i.e., relationships among the users of a social network which go beyond those characterizing conventional communities. For instance, this approach is capable of detecting groups of users who show forms of aversion with each other rather than just users who are willing to interact. Experimental comparisons of various approaches to finding communities in OSNs are reported in [35, 65].

In [55] the authors propose CHRONICLE, an algorithm to find time-evolving communities in a social network. CHRONICLE operates in two stages: in the first one it considers T "snapshots" of the social network in correspondence of T different timestamps. For each timestamp it applies a density-based clustering algorithm on each snapshot to find communities in the social network. After this, it builds a T-partite graph G_T which consists of T layers each containing the communities of nodes detected in the corresponding timestamp. It adds also some edges linking adjacent layers: they indicate that two communities, detected in correspondence of consecutive timestamps, share some similarities. As a consequence, the edges and the paths in G_T identify similarities among communities over time.

2.4 Influential User Detection

A recent trend in OSN analysis is the identification of *influential users* [40, 74]. Influential users are those capable of stimulating others to join OSN activities and/or to actively contribute to them.

In Weblog (blog) analysis, there is a special emphasis on so-called *leader identification*. In particular, [87] suggested to model the blogosphere as a graph (whose nodes represent bloggers whereas edges model the fact that a blogger cites another one). In [66], the authors introduces the concept of *starter*, i.e., a users who first generate information that catchs the interest of fellow users/readers. Among other, the approach of [66] has deployed the *Random Walk* technique to find starters. Researchers from HP Labs analyzed user behaviors on Twitter [82]; they found that influential users should not only catch attention from other users but they should also overcome the passivity of other users and spur them to get involved in OSN activities. To this purpose, they developed an algorithm (based on the HITS algorithm of [57]) to assess the degree of influence of a user. Experimental results show that high levels of popularity of a user do not necessarily imply high values in the degree of influence.

3 Sampling the Facebook Social Graph

Our work on OSN analysis began with the goal to understand the organization of popular OSN, and as of 2010 Facebook was by far the largest and most studied. Facebook gathers more than 500 millions active users, and its growth rate has been proved to be the highest among all the other competitors in the last few years. More than 50% of users log on to the platform in any given day. Coarse statistics about the usage of the social network are provided by the company itself[4]. Our study is interested in analyzing the characteristics and the properties of this network on a large-scale. In order to reach this goal, first of all we had to acquire data from this platform, and later we proceeded to their analysis.

3.1 The Structure of the Social Network

The structure of the Facebook social network is simple. Each subscribed user can be connected to others via friendship relationships. The concept of friendship is *bilateral*, this means that users must confirm the relationships among them. Connections among users do not follow any particular hierarchy, thus we define the social network as *unimodal*.

This network can be represented as a graph $G = (V, E)$ whose nodes V represent users and edges E represent friendship connections among them.

---

[4] Please refer to `http://www.facebook.com/press/info.php?statistics`

Because of the assumption on the bilateralness of relationships, the graph is *undirected*. Moreover, the graph we consider is *unweighted*, because all the friendship connections have the same value within the network. However, it could be possible to assign a weight to each friendship relation, for instance, considering the frequency of interaction between each pair of users, or different criteria. Considering the assumption that loops are not allowed, we conclude that in our case is possible to use a simple unweighted undirected graph to represent the Facebook social network. The adoption of this model has been proved to be optimal for several social networks (see [45]).

Although choosing the model for representing a network could appear to be simple, this phase is important and could be not trivial. Compared to Facebook, the structure of other social networks requires a more complex representative model. For example, Twitter should be represented using a *multiplex* network; this because it introduces different types of connections among users ("following", "reply to" and "mention") [83]. Similar considerations hold for other OSNs, such as aNobii [5], Flickr and YouTube [69], etc.

How to get information about the structure of the network

One important aspect to be considered for representing the model of a social network is the amount of information about its structure we have access to. The ideal condition would be to have access to the whole network data, for example acquiring them directly from the company which manage the social networking service. For several reasons (see further), the most of the time this solution is not viable.

Another option is to obtain data required to reconstruct the model of the network, acquiring them directly from the platform itself, exploiting its public interface. In other words, a viable solution is to collect a representative sample of the network to correctly represent its structure. To this purpose, it is possible to exploit Web data mining techniques [34] to extract data from the front-end of the social network Websites. This implies that, for very large OSNs, such as Facebook, Twitter, etc., it is hard or even impossible to collect a complete sample of the network. The first limitation is related to the computational overhead of a large-scale Web mining task. In the case of Facebook, for example, to crawl the friend-list Web page (dimension $\simeq 200KB$) for the half billion users, it would approximately require to download more than $200KB \cdot 500M = 100$ Terabytes of HTML data. Even if possible, the acquired sample would be a snapshot of the structure of the graph at the time of the mining process. Moreover, during the sampling process the structure of the network slightly changes. This because, even if short, the data mining process requires a not negligible time, during which the connections among users evolve, thus the social network, and its structure changes accordingly. For example, the growth rate of Facebook has been estimated in the order of 0.2% per day [42]. In other words, neither all these efforts could ensure to acquire a perfect sample. For these reasons, a widely adopted approach

is to collect small samples of a network, trying to preserve characteristics about its structure. There are several different sampling techniques that can be exploited; each algorithm ensures different performances, bias of data, etc.

For our experimentation we collected two significant samples of the structure of the social network, of a size comparable to other similar studies [24, 92, 42]. In particular, we adopted two different sampling algorithms, namely "breadth-first-search" and "Uniform". The first is proved to introduce bias in certain conditions (e.g., in incomplete visits) towards high degree nodes [59]. The latter is proved to be unbiased by construction [42].

Once collected, data are compared and analyzed in order to establish their quality, study their properties and characteristics. We consider two quality criteria to evaluate the samples: i) coherency with statistical data provided by the social network itself; ii) congruency with results reported by similar studies. Considerations about the characteristics of both the "breadth-first-search" and the "Uniform" samples follow in Section 5.

How to extract data from Facebook

Companies providing online social networking services, such as Facebook, Twitter, etc., do not have economic interests in sharing their data about users, because their business model mostly relies on advertising. For example, exploiting this information, Facebook provides unique and efficient services to advertising companies. Moreover, questions about the protection of these data have been advanced, for privacy reasons, in particular for Facebook [47, 67].

In this social network, for example, information about users and the interconnections among them, their activities, etc. can only be accessed through the interface of the platform. To preserve this condition some constraints have been implemented. Among others, a limit is imposed to the amount of information accessible from profiles of users not in friendship relations among them. There are also some technical limitations, e.g the friend-list is dispatched through an asynchronous script, so as preventing naive techniques of crawling. Some Web services, such as the "Graph API"[5], etc., have been provided during the last few months of 2010, by the Facebook developers team, but they do not bypass these limitations (and they eventually add even more restrictions). As of 2010, the structure of this social network can be accessed only exploiting techniques typical of Web data mining.

3.2 The Sampling Architecture

In order to collect data from the Facebook platform, we designed a Web data mining architecture, which is composed of the following elements (see Figure 1): i) a server running the mining agent(s); ii) a cross-platform Java application, which implements the logic of the agent; and, iii) an Apache interface,

---

[5] Available from `http://developers.facebook.com/docs/api`

which manages the information transfer through the Web. While running, the agent(s) query the Facebook server(s) obtaining the friend-list Web pages of specific requested users (this aspect depends on the implemented sampling algorithm), reconstructing the structure of relationships among them. Collected data are stored on the server and, after a post-processing step (see Section 3.5), they are delivered (eventually represented using an XML standard format [21]) for further experimentation.
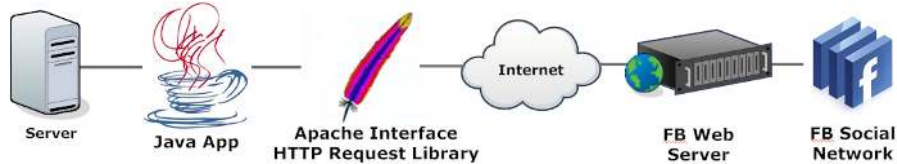

Fig. 1. Architecture of the data mining platform

The Facebook crawler

The cross-platform Java agent which crawl the Facebook front-end is the core of our mining platform. The logic of the developed agent, regardless the sampling algorithm implemented, is depicted in Figure 2. The first preparative step for the agent execution includes choosing the sampling methodology and configuring some technical parameters, such as termination criterion/a, maximum running time, etc. Thus, the crawling task can start or be resumed from a previous point. During its execution the crawler visits the friend-list page of a user, following the chosen sampling algorithm directives, for traversing the social graph. Data about new discovered nodes and connections among them are stored in a compact format, in order to save I/O operations. The process of crawling concludes when the termination criterion/a is/are met.

Fig. 2. State diagram of the data mining process

During the data mining step, the platform exploits the Apache HTTP Request Library[6] to communicate with the Facebook server(s). After an authentication phase which uses a secure connection and "cookies" for logging into the Facebook platform, HTML friend-list Web pages are obtained via HTTP requests. This process is described in Table 1.

---

[6] http://httpd.apache.org/apreq

N.	Action Protocol Method URI	KB
1	open the Facebook page	
	HTTP GET www.facebook.com/	242
2	login providing credentials	
	HTTPS POST login.facebook.com/login.php	234
	HTTP GET /home.php	87
3	visit the friend-list page of a specific users	
	HTTP GET /friends/ajax/friends.php?id=#&filter=afp	224

Table 1. HTTP requests flow of the crawler: authentication and mining steps

Limitations

During the data mining task we noticed a technical limitation imposed by Facebook on the dimension of the dispatched friend-list Web pages, via HTTP requests. To reduce the traffic through its network, Facebook provides shortened friend-lists not exceeding 400 friends. During a normal experience of navigation on the Website, if the dimension of the friend-list Web page exceeds 400 friends, an asynchronous script fills the page with the remaining. This results is not reproducible using an agent based on HTTP requests. This problem can be avoided using a different mining approach, for example adopting visual data extraction techniques [34]. Data can be retrieved directly from the Web page using specific scripts designed for a Web browser, or alternatively by developing an agent which integrates a Web browser for rendering the pages. This approach is not viable for large-scale mining tasks, but we already dealt with this approach in [22] for a smaller experimentation. In Section 5.2 we investigated the impact of this limitation on the samples.

3.3 Breadth-first-search Sampling

The breadth-first-search (BFS) is a uninformed traversal algorithm which aims to visit a graph. Starting from a "seed node", it explores its neighborhood; then, for each neighbor, it visits its unexplored neighbors, and so on, until the whole graph is visited (or, alternatively, if a termination criterion is met). This sampling technique shows several advantages: i) ease of implementation; ii) optimal solution for unweighted graphs; iii) efficiency. For these reasons it has been adopted in a variety of OSN mining studies, including [69, 24, 92, 42, 93, 22]. In the last year, the hypothesis that the BFS algorithm produces biased data, toward high degree nodes, if adopted for partial graph traversal, has been advanced by [59]. This because, in the same (partial) graph, obtained adopting a BFS visiting algorithm, are both represented nodes which have been visited (high degree nodes) and nodes which have just been discovered, as neighbors of visited ones (low degree nodes). One important aspect of our experimentation has been to verify this hypothesis, in order to highlight which properties of a partial graph obtained using the BFS sampling are preserved,

and which are biased. To do so, we had to acquire a comparable sample which is certainly unbiased by construction (see further).

Description of the breadth-first-search crawler

The BFS sampling methodology is implemented as one of the possible visiting algorithms in our Facebook crawler, described before. While using this algorithm, the crawler, for first, extracts the friend-list of the "seed node", which is represented by the user actually logged on the Facebook platform. The user-IDs of contacts in its friend-list are stored in a FIFO queue. Then, the friend-lists of these users are visited, and so on. In our experimentation, the process continued until two termination criteria have been met: i) at least the third sub-level of friendship was completely covered; ii) the mining process exceeded 240 hours of running time. As discussed before, the time constraint is adopted in order to observe a short mining interval, thus the temporal evolution of the network is minimal (in the order of 2%) and can be ignored. The obtained graph is a partial reconstruction of the Facebook network structure, and its dimension is used as a yardstick for configuring the "Uniform" sampling (see further).

Characteristics of the breadth-first-search dataset

This crawler has been executed during the first part of August 2010. The acquired sample covers about 12 millions friendship connections among about 8 millions users. Among these users, we performed the complete visit of about 63.4 thousands of them, thus resulting in an average degree $\overline{d} = \frac{2 \cdot |E|}{|V_v|} \simeq 396.4$, considering V as the number of *visited users*.

The overall mean degree, considering V as the number of *total nodes* on the graph (*visited users + discovered neighbors*), is $\overline{o} = \frac{2 \cdot |E|}{|V_t|} \simeq 3.064$. The expected density of the graph is $\Delta = \frac{2 \cdot |E|}{|V_v| \cdot (|V_v| - 1)} \simeq 0.006259 \simeq 0.626\%$, considering V as the number of visited nodes. We can combine the previous equations obtaining $\Delta = \frac{\overline{d}}{|V_v| - 1}$. It means that the expected density of a graph is the average proportion of edges incident with nodes in the graph.

In our case, the value $\delta = \frac{\overline{o}}{\overline{d}} = \frac{|V_v|}{|V_t|} \simeq 0.007721 \simeq 0.772\%$, which here we introduce, represents the effective density of the obtained graph.

The distance among the effective and the expected density of the graph, which here we introduce, is computed as $\partial = 100 - \frac{\Delta * 100}{\delta} \simeq 18.94\%$.

This result means that the obtained graph is slightly more connected than expected, w.r.t. the number of unique users it contains. This consideration is also compatible with hypothesis advanced in [59]. The effective diameter of this (partial) graph is 8.75, which is compliant with the "six degrees of separation" theory [68, 88, 10, 72].

The coverage of the graph is almost complete (99.98%). The small amount of disconnected nodes can be intuitively adducted due to some collisions

caused by the hash function exploited to de-duplicate and anonymize user-IDs adopted during the data cleaning step (see Section 3.5). Some interesting considerations hold for the obtained clustering coefficient result. It lies in the lower part of the interval [0.05, 0.18] reported by [42] and, similarly, [0.05, 0.35] by [92], using the same sampling methodology. The characteristics of the collected sample are summarized in Table 2.

N. of Visited Users	N. of Discovered Neighbors	N. of Edges
63.4K	8.21M	12.58M

Avg. Deg.	Bigg. Eigenval.	Eff. Diam.	Avg. Clust. Coef.	Coverage	Density
396.8	68.93	8.75	0.0789	98.98%	0.626%

Table 2. BFS dataset description (crawling period: 08/01-10/2010)

3.4 Uniform Sampling

To acquire a comparable sample, unbiased for construction, we exploited a rejection sampling methodology. This technique has been applied to Facebook in [42], where the authors proved its correctness. Its efficiency relies on the following assumptions:

1. it is possible to generate uniform sampling values for the domain of interest;
2. these values are not sparse w.r.t. the dimension of the domain and
3. it is possible to sample these values from the domain.

In Facebook, each user is identified by a 32-bit number user-ID. Considering that user-IDs lie in the interval $[0, 2^{32} - 1]$, the highest possible number of assignable user-IDs using this system is $H \simeq 4.295e9$.

The space for names is currently filling up since the actual number of assigned user-IDs, $R \simeq 5.4e8$ roughly equals to the 540 millions of currently subscribed users[7],[8]), the two domains are comparable and the rejection sampling is viable. We generated an arbitrary number of random 32-bit user-IDs, querying Facebook for their existence (and, eventually, obtaining their friend-lists). That sampling methodology shows two advantages: i) we can statistically estimate the probability $\frac{R}{H} \simeq 12.5\%$ of getting an existing user; thus, ii) we can generate an arbitrary number of user-IDs in order to acquire a sample of the desired dimension. Moreover, the distribution of user-IDs is completely independent w.r.t. the graph structure.

---

[7] As of August 2010, http://www.facebook.com/press/info.php?statistics
[8] http://www.google.com/adplanner/static/top1000/

Description of the "Uniform" crawler

The "Uniform" sampling is another algorithm implemented in the Facebook crawler we developed. Differently w.r.t. the BFS sampler, if adopting this algorithm, it is possible to parallelize the process of extraction. This because user-IDs to be requested can be stored in different "queues". We designed the uniform sampling task starting from these assumptions: i) the number of subscribed users is $2^{29} \simeq 5.368e8$; ii) this value is comparable with the highest possible assignable number of user-IDs, $2^{32} \simeq 4.295e9$, thus iii) we can statistically assert that the possibility of querying Facebook for an existing user-ID is $\frac{2^{29}}{2^{32}} = \frac{1}{8}$ (12.5%). To this purpose, we generated eight different queues, each containing $2^{16} \simeq 65.5K \cong 63.4K$ random user-IDs (the number of visited users of the BFS sample), used to feed eight parallel crawlers.

Characteristics of the "Uniform" dataset

The uniform sampling process has been executed during the second part of August 2010. The crawler collected a sample which contains almost 8 millions friendship connections among a similar number of users. The acquired amount of nodes differs from the expected number due to the privacy policy adopted by those users who prevent their friend-lists being visited. The privacy policy aspect is discussed in Section 5.1.

The total number of visited users has been about 48.1 thousands, thus resulting in an average degree of $\overline{d} = \frac{2 \cdot |E|}{|V_v|} \simeq 326.0$, considering V as the number of *visited users*. Same assumptions, the expected density of the graph is $\Delta = \frac{2 \cdot |E|}{|V_v| \cdot (|V_v| - 1)} \simeq 0.006777 \simeq 0.678\%$.

If we consider V as the number of *total nodes* (*visited users + discovered neighbors*), the overall mean degree is $\overline{o} = \frac{2 \cdot |E|}{|V_t|} \simeq 2.025$. The effective density of the graph, previously introduced, is $\delta = \frac{|V_v|}{|V_t|} \simeq 0.006214 \simeq 0.621\%$. The distance among the effective and the expected density of the graph, is $\partial = 100 - \frac{\Delta * 100}{\delta} \simeq -9.06\%$. This can be intuitively interpreted as a slight lack of connection of this sample w.r.t. the theoretical expectation.

Some considerations hold, also comparing against the BFS sample: the average degree is slightly less (326.0 vs. 396.8), but the effective diameter is almost the double (16.32 vs. 8.75). We justify this characteristic considering that the sample could be still too small and disconnected to perfectly reflect the structure of the network. Our hypothesis is also supported by the dimension of the largest connected component, which does not contain the 5% of the sample. Finally, the clustering coefficient, less than the BFS sample (0.0471 vs. 0.0789), is still comparable w.r.t. previously considered studies [92, 42].

3.5 Data Preparation

During the data mining process it could happen to store redundant information. In particular, while extracting friend-lists, a crawler could save multiple

N. of Visited Users	N. of Discovered Neighbors	N. of Edges
48.1K	7.69M	7.84M

Avg. Deg.	Bigg. Eigenval.	Eff. Diam.	Avg. Clust. Coef.	Coverage	Density
326.0	23.63	16.32	0.0471	94.96%	0.678 %

Table 3. "Uniform" dataset description (crawling period: 08/11-20/2010)

instances of the same edge (i.e., a parallel edge), if both the connected users are visited; this because the graph is undirected. We adopted a hashing-based algorithm which cleans data in $O(N)$ time, removing duplicate edges. Another step, during the data preparation, is the *anonymization*: user-IDs are "encrypted" adopting a 48-bit hybrid rotative and additive hash function [77], to obtain anonymized datasets. The final touch was to verify the integrity and the congruency of data. We found that the usage of the hashing function caused occasional collisions (0.0002%). Finally, some datasets of small subgraphs (e.g., ego-networks) have been post-precessed and stored using the GraphML format [21].

4 Network Analysis Aspects

During the last years, important achievements have been reached in understanding the structural properties of several complex real networks. The availability of large-scale empirical data, on the one hand, and the advances in computational resources, on the other, made it possible to discover and understand interesting statistical properties commonly shared among different real-world social, biological and technological networks. Among others, some important examples are: the World Wide Web [6], Internet [32], metabolic networks [54], scientific collaboration networks [70, 11], citation networks [81], etc. Indeed, during the last years even the social networks are strongly imposing themselves as complex networks described by very specific models and properties. For example, some studies [68, 88, 8] proved the existence of the so called "small-world" property in complex social networks. Others [76, 2], assert that the so called "scale-free" complex networks reflect a "power-law" distribution as model for describing the behavior of node degrees. We can conclude that the topology of the networks usually provides useful information about the dynamics of the network entities and the interaction among them.

In addition to contribute to the advances in *graph theory*, the study of complex networks led to important results also in some specific contexts, such as the *social sciences*. A branch of the network analysis applied to *social sciences* is the *Social Network Analysis* (SNA). From a different perspective w.r.t. the analysis of complex networks, which mainly aims to analyze structural properties of networks, the SNA focuses on studying the nature of relationships

among entities of the network and, in the case of social networks, investigating the motivational aspect of these connections.

In this Section we will briefly describe properties and models characterizing the structure of complex networks (see Sections 4.1 and 4.2); then we will focus on defining measures and metrics proper of SNA (see Section 4.3) and, concluding, we will consider some interesting aspects regarding the visualization of related graphs (see Section 4.4).

4.1 Definitions

In this section we describe some of the common structural properties which are usually observed in several complex networks and define the way they are measured. Then, we describe concepts about the mathematical modeling of networks, including random graph models and their generalizations, the "small-world" model and its variations, and models of growth and evolution of graphs, including the preferential attachment models.

Shortest path lengths, diameter and betweenness

Let $G = (V, E)$ to be a graph representing a network; the distance d_{ij} between two nodes, labeled i and j respectively, is defined as the number of edges along the shortest path connecting them. The diameter D of the network represented by G, therefore, is defined to be the maximal distance among all distances between any pair of nodes in the network.

A common measure of distance between two nodes of the graph G, is given by the *average shortest path length* [90, 91] (also called *characteristic path length*), defined as the mean value of the geodesic distance (i.e., the shortest path) between the all-pairs node of the graph (a.k.a. the "all-pairs-shortest-path problem" [84]):

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \qquad (3)$$

The main problem adopting this definition is that ℓ diverges if the graph contains disconnected components. A possible solution to this issue is to limit the domain of the summation in Equation 3 only to the pairs of nodes belonging to the largest connected component of the graph.

Several metrics have been defined to compute the centrality of a node within a graph, such as the degree, closeness and the betweenness centrality. The latter one has been originally introduced to quantitatively evaluate the importance of an individual in a network [89, 60] and is recognized to be the most appropriate measure to reflect the concept of centrality in a network. The betweenness centrality of a node was defined by Freeman [36, 37] as in Equation 2. Correlations "betweenness-vs.-betweenness" and "betweenness-vs.-degree" have been investigated respectively by [43] and [44, 49, 26]. In

Section 4.4 we will focus on analyzing this metric calculated on the Facebook graph, for reasons further explained.

Clustering

In several networks it is shown that, if a node i is connected to a node j, which in its turn is connected to a node k, then there is a heightened probability that node i will be also connected to node k. From a social network perspective, a friend of your friend is likely also to be your friend. In terms of network topology, transitivity means the presence of a heightened number of triangles in the network, which constitute sets of three nodes connected each others [71]. The global clustering coefficient is defined by:

$$C_g = \frac{3 \times \textit{no. of triangles in } G}{\textit{no. of connected triples}} \tag{4}$$

where a *connected triple* represents a pair of nodes connected to another node. C_g is the mean probability that two persons who have a common friend are also friends together. An alternative definition of clustering coefficient C has been provided by Watts and Strogatz [91]. During our experimentation we investigated the clustering effect on the Facebook network (see Section 5.3).

The "small world" property

It well-known in literature that most large-scale networks, despite their huge size, show a common property: there exists a relatively short path which connects any pair of nodes within the network. This characteristic, the so called *small-world* property, is theoretically supported by the average shortest path length, defined by Equation 3, and it scales proportionally to the logarithm of the dimension of the network. The study of this phenomenon is rooted in *social sciences* [68, 88] and is strictly interconnected with the notion of diameter we introduced before. The Facebook social network reflects the "small world" property as discussed in Section 5.3.

Scale-free degree distributions

On the one hand, in a random graph (see further) the node degree (i.e., the number of edges the node has) is characterized by a distribution function $P(k)$ which defines the probability that a randomly chosen node has exactly k edges. Because the distribution of edges in a random graph is aleatory, the most of the nodes have approximatively the same node degree, similar to the mean degree $\langle k \rangle$ of the network. Thus, the degree distribution of a random graph is well described by a Poisson distribution law, with a peak in $P(\langle k \rangle)$. On the other hand, recent empirical results show that in the most of real-world networks the degree distribution significantly differs w.r.t. a Poisson distribution. In

particular, for several large-scale networks, such as the World Wide Web [6], Internet [32], metabolic networks [54], etc., the degree distribution follows a power-law:

$$P(k) \sim k^{-\lambda} \tag{5}$$

This power-law distribution falls off more gradually than an exponential one, allowing for a few nodes of very large degree to exist. Since these power-laws are free of any characteristic scale, such a network with a power-law degree distribution is called a scale-free network [9]. We proved that Facebook is a scale-free network well described by a power-law degree distribution, as discussed in Section 5.2.

4.2 Networks models

Concepts such as the short paths length, the clustering and the scale-free degree distribution have been recently applied to rigorously model the networks. Three main modeling paradigms exist: i) random graphs, ii) "small world" networks and, iii) power-law networks. Random graphs represent an evolution of the Erdős-Rényi model, and are widely used in several empirical studies, because of the ease of adoption. After the discovery of the clustering property, a new class of models, namely "small world" networks, has been introduced. Similarly, the power-law degree distribution definition led to the modeling of the homonym networks, which are adopted to describe scale-free behaviors, focusing on the dynamics of the network in order to explain phenomena such as the power-law tails and other non-Poisson degree distribution, empirically shown by real-world networks.

The Erdős-Rényi model

Erdős and Rényi [30, 31] proposed one of the first models of network, the random graph. They defined two models: the simple one consists of a graph containing n vertices connected randomly. The commonly adopted model, indeed, is defined as a graph $G_{n,p}$ in which each possible edge between two vertices may be included in the graph with the probability p (and may not be included with the probability $(1 - p)$).

Although random graphs have been widely adopted because their properties ease the work of modeling networks (e.g., random graphs have a small diameter), they do not properly reflect the structure of real-world large-scale networks, mainly for two reasons: i) the degree distribution of random graphs follows a Poisson law, which substantially differs from the power-law distribution shown by empirical data; ii) they do not reflect the clustering phenomenon, considering all the nodes of the network with the same "weight", and reducing, de facto, the network to a giant cluster.

The Watts-Strogatz model

The real-world social networks are well connected and have a short average path length like random graphs, but they also have exceptionally large clustering coefficient, which is not reflected by the Erdős-Rényi model or by other random graph models. In [91], Watts and Strogatz proposed a one-parameter model that interpolates between an ordered finite dimensional lattice and a random graph. Starting from a ring lattice with n vertices and k edges per vertex, each edge is rewired at random with probability p [91].

The model has been widely studied since the details have been published. Its role is important in the study of the small-world theory. Some relevant theories, such as Kleinberg's work [57, 56], are based on this model and its variants. The disadvantage of the model, however, is that it is not able to capture the power-law degree distribution as presented in most real-world social networks.

The Barabási-Albert model

The two previously discussed theories observe properties of real-world networks and attempt to create models that incorporate those characteristics. However, they do not help in understanding the origin of social networks and how those properties evolve.

The Barabási-Albert model suggests that two main ingredients of self-organization of a network in a scale-free structure are *growth* and *preferential attachment*. These pinpoint to the facts that the most of networks continuously grow by the addition of new nodes which are preferentially attached to existing nodes with large numbers of connections (again, "rich gets richer"). The generation scheme of a Barabási-Albert scale-free model is as follows: (i) *Growth*: let p_k to be the fraction of nodes in the undirected network of size n with degree k, so that $\sum_k p_k = 1$ and therefore the mean degree m of the network is $\frac{1}{2} \sum_k k p_k$. Starting with a small number of nodes, at each time step, we add a new node with m edges linked to nodes already part of the system; (ii) *preferential attachment*: the probability $\prod_i$ that a new node will be connected to the node i (one of the m already existing nodes) depends on the degree k_i of the node i, in such a way that $\prod_i = k_i \sum_j k_j$.

Models based on preferential attachment operates in the following way. Nodes are added one at a time. When a new node u has to be added to the network it creates m edges (m is a parameter and it is constant for all nodes). The edges are not placed uniformly at random but *preferentially*, i.e., probability that a new edge of u is placed to a node v of degree $d(v)$ is proportional to its degree, $p_u(v) \propto d(v)$. This simple behavior leads to power-law degree tails with exponent $\lambda = 3$. Moreover it also leads to low diameters. While the model captures the power-law tail of the degree distribution, it has other properties that may or may not agree with empirical results in real networks. Recent analytical research on average path length indicate that

$\ell \sim ln(N)/lnln(N)$. Thus the model has much shorter l w.r.t. a random graph. The clustering coefficient decreases with the network size, following approximately a power-law $C \sim N^{-0.75}$. Though greater than those of random graphs, it depends on network size, which is not true for real-world social networks.

4.3 Social Network Analysis

In the previous section we discussed about the network as a complex system: in particular, complex network theory, by its graph theoretical approach, does not explain the network by its elements' behavior, but it deals with a whole organism that evolves by means of its single components. In this section, instead, we approach the study of that components: SNA deals with the study of the actors involved in a network. It is an approach based on the analysis of the behavior of single entities which are part of the network and govern its evolution. The single components have the possibility of choosing their own connections without considering the network as a whole structure but only w.r.t. individual characteristics. Indeed, almost the totality of social network models deals with the external information on the actors. Social network analysts often use these additional information to explain network formation. This is the principal difference between social network analysis and complex network theory. In the latter we often disregard the additional information on the single nodes because the attention is pointed out toward a more structural method, namely a *systemic approach.*

Metrics

Metrics allow analysts to systematically dissect the social world, creating a basis on which to compare networks, track changes over the time and determine the relative position of individuals and clusters within the network [51].

One of the primary uses of *graph theory* in social network analysis is the identification of the most important actors in a social network [89]. The *degree centrality* is a measure of the degree of an actor in a network. An actor with a high degree centrality is "where the action is" in the network. Thus, this measure focuses on the most visible actors in the network. This actors can be recognized by others as a major channel of relational information. In contrast, actors with low degrees are peripheral in the network.

A second view of centrality is based on *closeness* or distance. This measure focuses on how close an actor is to all the other actors in the network. This idea of centrality based on closeness is inversely related to the distance. As a node grows farther apart in distance from other nodes, its centrality will decrease, since there will be more lines in the geodesics linking that node to the rest [33]. Finally, interactions between two non adjacent actors might depend on the other actors, especially those lying on the paths between them.

These other actors potentially might have some control over the interactions between the two non-adjacent actors. The important idea here is that an actor is central if it lies between other actors on their geodesics. This implies that, to have a large *betweenness centrality* [36], the actor must be between many of the actors via their geodesics [91]. Although this centrality has gained popularity because of its generality, this index assumes that all geodesics are equally likely when estimating the critical probability that an actor fall on a particular geodesic. It also ignores the fact that if some actors on the geodesics have large degrees, then the geodesics containing these expansive actors are more likely to be used as shortest paths than other geodesics. Also it would be more realistic to consider betweenness counts which focus on paths other than geodesics. Information centrality generalizes the notion of betweenness centrality so all paths between actors, with weights depending on their lengths, are considered when calculating the betweenness counts.

4.4 Visualizing Social Networks

One of the key elements that characterize modern social network analysis is the visual representation. Looking at a network graph may provide an overview of the structure of the network, calling out cliques, communities, and key participants. Drawings of relational structures like social networks are only useful if they "effectively convey information to the people that use them" [28, 29, 19]. Network visualization is often as frustrating as appealing. Network graphs can rapidly get too dense and large to make out any meaningful patterns. Many obstacles like vertex occlusions and edge crossings make creating well-organized and readable network graphs challenging. There is an upper limit on the numbers of vertices and edges that can be displayed in a bounded set of pixels; typically only a few hundred or thousand vertices can be meaningfully and distinctly represented on average-sized computer screens.

A key reference for better-quality network visualization is the so-called Netviz Nirvana guidelines [79]. Several graph layout algorithms can be used, including variants of the "spring embedder" such as the Harel-Koren [52] fast multi-scale method, the popular Fruchterman-Reingold [38] force-directed algorithm and more scalable gravitational N-Body approaches (see Figures 3 and 4), such as those implemented in *LogAnalysis* [23] and *NodeXL* [51]. The results of applying these algorithms vary depending on the size and topology of the network.

LogAnalysis presents social networks using a familiar node-link representation, where nodes represent members of the system and links represent the articulated "friendship" links between them. It integrates statistics proposed by Perer and Shneiderman [78]: overall network metrics (i.e., number of nodes, number of edges, density, diameter), node rankings (i.e., degree, betweenness and closeness centrality), edge rankings (i.e., weight, betweenness centrality), edge rankings in pairs and cohesive subgroups. Network members are presented using both their self-provided name, ID (e.g., the Facebook user-ID)

and, if available, a representative photograph (e.g., the Facebook profile picture). The networks are presented as egocentric networks. Users can expand the display by selecting nodes to make visible others' friends as well. Analysts can also explore a network by focusing on one node, the node's neighbors, and the ties among them and can interactively increase the depth of the neighborhood by dragging a slider bar.
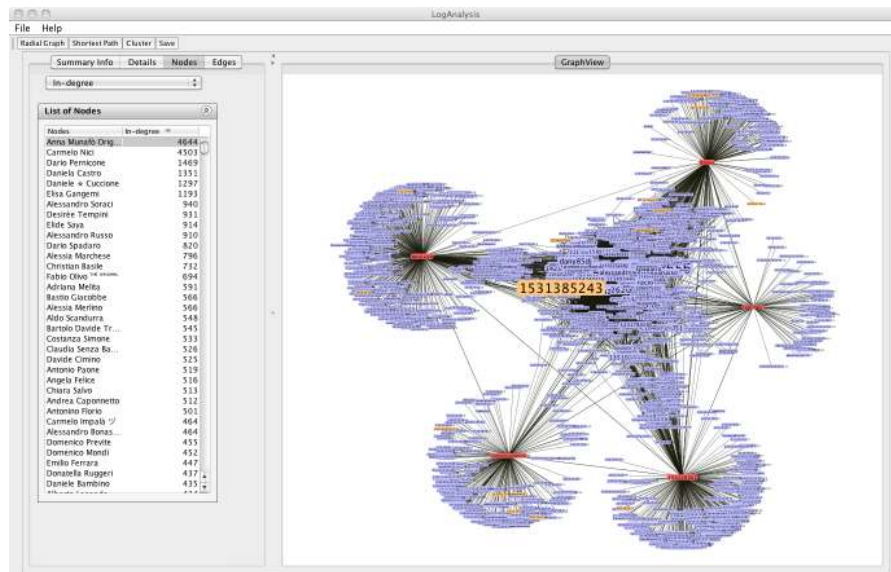

Fig. 3. N-Body approach with LogAnalysis

Betweenness centrality in Facebook

The analysis of large ego-networks led us to another interesting consideration on the behavior of the *betweenness centrality* (BC); it goes as follows. Clearly, the more a node is central and *important,* the higher its BC. One could suppose that this measure is directly interconnected with the degree of a node, or with other measures of centrality (e.g., the Pagerank). In Figure 4 we show the behavior of this metric evaluated on an ego-network of about 25 thousands of nodes. The following consideration holds: the node which covers the most important position in this network (vertex "8478", in red) does not show "special" properties (e.g., its degree and its Pagerank are lower than the most of the other nodes). However, it appears in more than the double of shortest paths w.r.t. the other nodes of the network. Similar considerations hold for other nodes (vertices "24221", "5851", "9453", in yellow, and "11661", "24853, in green) in this particular ranking. Intuitively, nodes with high BC represent a potential efficient way of connection among peripheral nodes.

It is known that the BC distribution follows a power-law $p(g) \sim g^{-\eta}$ for scale-free networks [43]. Similarly to the degree exponent case, in general, the BC exponents increase for node and link sampling and decrease for snowball sampling as the sampling fraction gets lower. The correlation between degree and BC of nodes [12], shown in Figure 10 (Section 5.4), could explain the same direction of changes of degree and BC exponents.

We can conclude that the study of the betweenness centrality in Facebook is fundamental for all those aspects related to discovering central nodes of the network, and that BC is a numerical property for applications (e.g., for marketing purposes, broadcasting news, etc.).
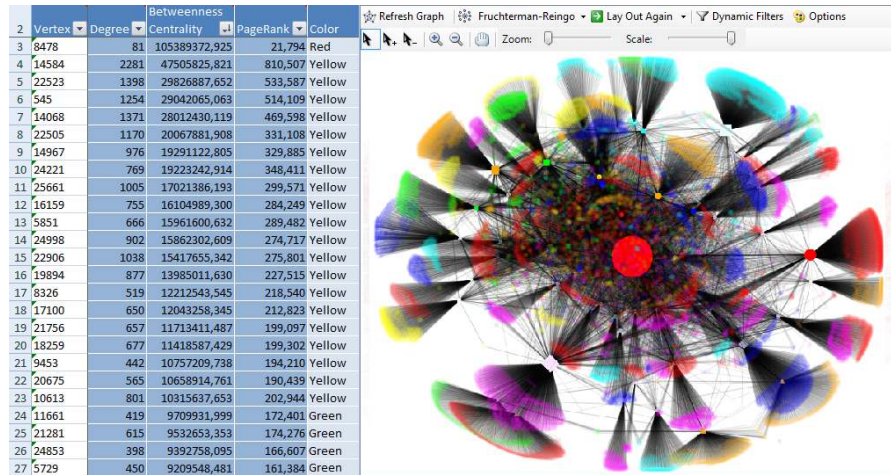
	Vertex	Degree	Betweenness Centrality	PageRank	Color
3	8478	81	105389372,925	21,794	Red
4	14584	2281	47505825,821	810,507	Yellow
5	22523	1398	29826887,652	533,587	Yellow
6	545	1254	29042065,063	514,109	Yellow
7	14068	1371	28012430,119	469,598	Yellow
8	22505	1170	20067881,908	331,108	Yellow
9	14967	976	19291122,805	329,885	Yellow
10	24221	769	19223242,914	348,411	Yellow
11	25661	1005	17021386,193	299,571	Yellow
12	16159	755	16104989,300	284,249	Yellow
13	5851	666	15961600,632	289,482	Yellow
14	24998	902	15862302,609	274,717	Yellow
15	22906	1038	15417655,342	275,801	Yellow
16	19894	877	13985011,630	227,515	Yellow
17	8326	519	12212543,545	218,540	Yellow
18	17100	650	12043258,345	212,823	Yellow
19	21756	657	11713411,487	199,097	Yellow
20	18259	677	11418587,429	199,302	Yellow
21	9453	442	10757209,738	194,210	Yellow
22	20675	565	10658914,761	190,439	Yellow
23	10613	801	10315637,653	202,944	Yellow
24	11661	419	9709931,999	172,401	Green
25	21281	615	9532653,353	174,276	Green
26	24853	398	9392758,095	166,607	Green
27	5729	450	9209548,481	161,384	Green

Refresh Graph · Fruchterman-Reingo · Lay Out Again · Dynamic Filters · Options · Zoom · Scale

Fig. 4. Betweenness centrality and clustering effect in a ego-network of 25K nodes

5 Experimental Work

We describe some interesting experimental results as follows. To compute the community profile of a network and its node centrality measures, such as degree and betweenness, we have adopted the Stanford Network Analysis Platform (SNAP) [62], a general purpose network analysis library.

5.1 Privacy Settings

We investigated the adoption of restrictive privacy policies by users: our statistical expectation using the "Uniform" crawler was to acquire $8 \cdot \frac{2^{16}}{2^3} \simeq 65.5K$ users. Instead, the actual number of collected users was 48.1K. Because of privacy settings chosen by users, the discrepancy between the expected number

of acquired users and the actual number was about 26.6%. In other words, only a quarter of Facebook users adopt privacy policies which prevent other users (except for those in their friendship network) from visiting their friend-list.

5.2 Degree distribution

A first description of the network topology of the Facebook friendship graph can be obtained from the degree distribution. According to Equation 5, a relatively small number of nodes exhibit a very large number of links. An alternative approach involves the Complementary Cumulative Distribution Function (CCDF)

$$\wp(k) = \int_k^\infty P(k')dk' \sim k^{-\alpha} \sim k^{-(\gamma-1)} \tag{6}$$

When calculated for a complete graph, CCDF shows up as a straight line in a log-log plot, while the exponent of the power-law distribution only varies the height (not the shape) of the curve.

In Figure 5 is plotted the degree distribution, as obtained from the BFS and "Uniform" sampling techniques. The limitations due to the dimensions of the cache which contains the friend-lists, upper bounded to 400, are evident. The BFS sample introduces an overestimate of the degree distribution in the left and the right part of the curves. The CCDF is shown, for the same sample, in Figure 6.
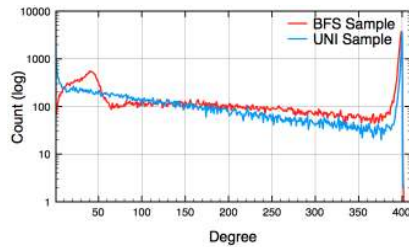

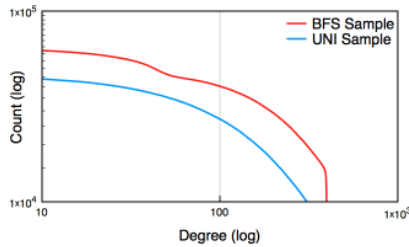**Fig. 5.** Degree distribution   **Fig. 6.** CCDF degree distribution

5.3 Diameter and Clustering Coefficient

It is well-known that most real-world graphs exhibit a relatively small diameter. A graph has diameter D if every pair of nodes can be connected by a path of length of at most D edges. The diameter D may be affected by outliers (again, the small-world phenomenon). A robust measure of the pairwise distances between nodes in a graph is the effective diameter, which is the minimum number of links (steps/hops) within which some fraction (or quantile q,

say q = 0.9) of all connected pairs of nodes can reach each other. The effective diameter has been found to be small for large real-world graphs, like Internet and the Web, real-life and OSNs [7, 68, 64].

The hop-plot package extends the notion of diameter by plotting the number of reachable pairs $g(h)$ within h hops, as a function of the number of hops h [76]. It gives us a sense of how quickly the neighborhoods of nodes expand with the number of hops. In Figure 7 the number of hops necessary to connect any pair of nodes is plotted as a function of the number of pairs of nodes. As a consequence of the more "compact" structure of the graph, the BFS sample shows a faster convergence to the asymptotic value listed in Table 2.

The clustering coefficient of a node is the ratio of the number of existing links over the number of possible links between its neighbors.

Given a network $G = (V, E)$, a clustering coefficient, C_i, of node $i \in V$ is:

$$C_i = 2|\{(v, w)|(i, v), (i, w), (v, w) \in E\}|/k_i(k_i - 1) \qquad (7)$$

where k_i is the degree of node i. It can be interpreted as the probability that any two nodes that share a common neighbor have a link between them. The clustering coefficient of a node represents how well connected its neighbors are. For any node in a tightly-connected mesh network, the clustering coefficient is 1. The clustering coefficient of a network is the mean clustering coefficient of all nodes.

Often, it is insightful to examine not only the mean clustering coefficient (see Section 4.1), but its distribution. In Figure 4 it is possible to clearly identify the clustering effect of a Facebook subgraph, visually enhanced by applying several iterations of the Fruchterman-Reingold algorithm [38]. Figure 8 shows the average clustering coefficient plotted as a function of the node degree for the two sampling techniques. As a consequence of the more systematic approach of extraction, the distribution of the clustering coefficient of the BFS sample shows a smooth behavior.
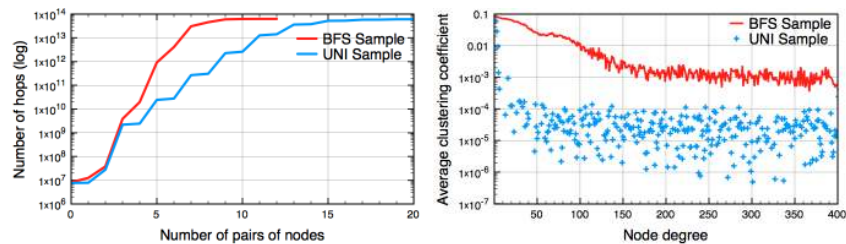

Fig. 7. Hops and diameter   **Fig. 8.** Clustering coefficient

The following considerations hold for the diameter and hops: the BFS sample may be affected by the "wavefront expansion" behavior of the visiting algorithm, while the "Uniform" sample may still be too small to represent

a faithful estimate of the diameter (this hypothesis is supported by the dimension of the largest connected component which does not cover the whole graph, as discussed in the next paragraph). Different conclusions can be derived for the clustering coefficient property. It is important to observe that the average values of the two samples fluctuate in the same interval reported by recent similar studies on OSNs (i.e., [0.05, 0.18] by Wilson et al. [92], [0.05, 0.35] by Gjoka et al. [42]), thus confirming that this property is preserved by both the adopted sampling techniques.

5.4 Connected Components

A connected component is a maximal set of nodes where for each pair of nodes there exists a path connecting them. Analogously, directed graphs show weakly and strongly connected components.

As shown in Tables 2 and 3, the largest connected components cover the 99.98% of the BFS graph and the 94.96% of the "Uniform" graph. In Figure 9, the scattered points in the left part of the plot have a different meaning for each sampling techniques. In the "Uniform" case, the sampling picked up disconnected nodes. In the BFS case, disconnected nodes are meaningless, as they are due to some collisions of the hashing function during the de-duplication phase of the data-cleaning step. This interpretation is supported by their small number (29 collisions over 12.58 millions of hashed edges) involving only the 0.02% of the total edges. However, the quality of the sample is not affected.

These conclusions are confirmed in Figure 10, where the betweenness centrality is plotted as a function of the degree. In the right part of the plot the betweenness centrality shows a linearly proportional behavior w.r.t. the degree. In our opinion, this implies a high degree of connectedness of the sample, since a high value of betweenness centrality is related to a high value of the degree of the nodes.
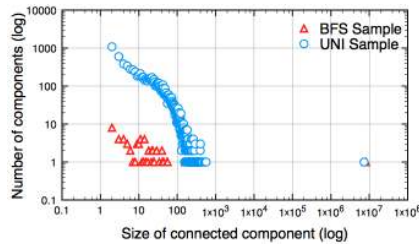

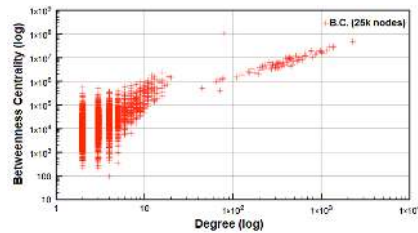**Fig. 9.** Connected components

Fig. 10. Betweenness centrality vs degree (on a ego-network of 25K nodes)

6 Conclusions

The success of OSNs and the growth of their user base is of great interest to both Social and Computer Science. Extraction and analysis of OSN data describing social networks poses both a technological challenge and an interpretation challenge. We have presented here our long-term research project on social network analysis and our two implemented systems: the ad hoc Facebook crawler and the LogAnalysis tool for analysis and visualization.

The ad hoc Facebook crawler has been developed to comply with the increasingly-strict terms of Facebook end-user license, i.e., to create large, fully anonymous samples that can be employed for scientific purposes. Two different sampling techniques have been implemented in order to explore the graph of friendships of Facebook, since the BFS visiting algorithm is known to introduce a bias in case of an incomplete visit.

Analysis of such large samples was tackled using concepts and algorithms typical of the *graph theory*, namely users were represented by nodes of a graph and relations among users were represented by edges. Our *LogAnalysis* tools supports OSN analysis and gives a graphical visualization of key graph theory and social network analysis concepts: degree distribution, diameter, centrality metrics, clustering coefficient computation and eigenvalues distribution. Future developments involve the implementation of different sampling techniques (e.g., Monte Carlo Random Walks) in order to speed-up the data extraction process and the evaluation of network metrics.

Acknowledgments

References

1. Adamic, L., Adar, E.: Friends and Neighbors on the Web. Social Networks **25**(3), 211–230 (2003)
2. Adamic, L., et al.: Power-law distribution of the world wide web. Science **287**(5461), 2115 (2000)
3. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering **17**(6), 734–749 (2005)
4. Ahn, Y., Han, S., Kwak, H., Moon, S., Jeong, H.: Analysis of topological characteristics of huge online social networking services. In: Proceedings of the 16th International Conference on the World Wide Web, pp. 835–844. ACM (2007)
5. Aiello, L.M., Barrat, A., Cattuto, C., Ruffo, G., Schifanella, R.: Link creation and profile alignment in the aNobii social network. In: Proceedings of the 2nd IEEE International Conference on Social Computing, pp. 249–256 (2010)

6. Albert, R.: Diameter of the World Wide Web. Nature **401**(6749), 130 (1999)
7. Albert, R., Barabási, A.: Statistical mechanics of complex networks. Reviews of Modern Physics **74**(1), 47–97 (2002)
8. Amaral, L., Scala, A., Barthélémy, M., Stanley, H.: Classes of small-world networks. Proceedings of the National Academy of Sciences **97**(21), 11,149 (2000)
9. Barabási, A., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509 (1999)
10. Barabási, A., Crandall, R.: Linked: The new science of networks. American Journal of Physics **71**, 409 (2003)
11. Barabási, A., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications **311**(3-4), 590–614 (2002)
12. Barthelemy, M.: Betweenness Centrality in Large Complex Networks. European Physical Journal B **38**, 163–168 (2004)
13. Batagelj, V., Doreian, P., Ferligoj, A.: An optimizational approach to regular equivalence. Social Networks **14**(1-2), 121–135 (1992)
14. Blondel, V., Gajardo, A., Heymans, M., Senellart, P., van Dooren, P.: A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching. SIAM Review **46**(4), 647–666 (2004)
15. Blondel, V., Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics **10**, P10,008 (2008)
16. Boldi, P., Rosa, M., Santini, M., Vigna, S.: Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks. In: Proceedings of the 20th International Conference on World Wide Web. ACM Press (2011)
17. Boldi, P., Vigna, S.: The WebGraph framework I: Compression techniques. In: Proceedings of the 13th International World Wide Web Conference, pp. 595–601. ACM Press (2004)
18. Borgatti, S.P., Everett, M.G.: Models of core/periphery structures. Social Networks **21**(4), 375–395 (2000)
19. Boyer, J.M., Myrvold, W.J.: On the Cutting Edge: Simplified O(n) Planarity by Edge Addition. Journal of Graph Algorithms and Applications **8**(3), 241–273 (2004)
20. Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE Transactions on Knowledge and Data Engineering **20**(2), 172–188 (2008)
21. Brandes, U., Eiglsperger, M., Herman, I., Himsolt, M., Marshall, M.: GraphML progress report structural layer proposal. In: Graph Drawing, pp. 109–112. Springer (2002)
22. Catanese, S., De Meo, P., Ferrara, E., Fiumara, G.: Analyzing the facebook friendship graph. In: Proceedings of the 1st International Workshop on Mining the Future Internet, pp. 14–19 (2010)
23. Catanese, S., Fiumara, G.: A visual tool for forensic analysis of mobile phone traffic. In: Proceedings of the 2nd ACM Workshop on Multimedia in Forensics, pp. 71–76. ACM (2010)
24. Chau, D., Pandit, S., Wang, S., Faloutsos, C.: Parallel crawling for online social networks. In: Proceedings of the 16th International Conference on the World Wide Web, pp. 1283–1284 (2007)
25. Clauset, A., Newman, M., Moore, C.: Finding community structure in very large networks. Physical Review E **70**(6), 066,111 (2004)

26. Crucitti, P., Latora, V., Marchiori, M.: A topological analysis of the Italian electric power grid. Physica A **338**, 92–97 (2004)
27. De Meo, P., Ferrara, E., Fiumara, G.: Social Networking and Community Behavior Modeling, chap. Finding Similar Users in Facebook. IGI Publisher (2011)
28. Di Battista, G., Eades, P., Tamassia, R., Tollis, I.: Algorithms for drawing graphs: An annotated bibliography. Computational Geometry **4**(5), 235–282 (1994)
29. Di Battista, G., Eades, P., Tamassia, R., Tollis, I.: Graph drawing: algorithms for the visualization of graphs. Prentice Hall (1998)
30. Erdős, P., Rényi, A.: On random graphs. Publicationes Mathematicae **6**(26), 290–297 (1959)
31. Erdős, P., Rényi, A.: On the Evolution of Random Graphs. In: Publication of the Mathematical Institute of the Hungarian Academy of Sciences (1960)
32. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: ACM SIGCOMM Computer Communication Review, vol. 29, pp. 251–262. ACM (1999)
33. Faust, K., Wasserman, S.: Centrality and Prestige: A Review and Synthesis. Journal of Quantitative Anthropology **4**(1985), 23–78 (1992)
34. Ferrara, E., Fiumara, G., Baumgartner, R.: Web Data Extraction, Applications and Techniques: A Survey. Tech. Report (2010)
35. Fortunato, S.: Community detection in graphs. Physics Reports **486**, 75–174 (2010)
36. Freeman, L.: A set of measures of centrality based on betweenness. Sociometry **40**(1), 35–41 (1977)
37. Freeman, L.: Centrality in social networks conceptual clarification. Social networks **1**(3), 215–239 (1979)
38. Fruchterman, T., Reingold, E.: Graph drawing by force-directed placement. Software: Practice and Experience **21**(11), 1129–1164 (1991)
39. Garton, L., Haythornthwaite, C., Wellman, B.: Studying online social networks. Journal of Computer-Mediated Communication **3**(1) (1997)
40. Ghosh, R., Lerman, K.: Predicting influential users in online social networks. In: Proceedings of KDD workshop on Social Network Analysis (SNA-KDD) (2010)
41. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the National Academy of Science **99**(12), 7821–7826 (2002)
42. Gjoka, M., Kurant, M., Butts, C., Markopoulou, A.: Walking in Facebook: a case study of unbiased sampling of OSNs. In: Proceedings of the 29th Conference on Information Communications, pp. 2498–2506. IEEE (2010)
43. Goh, K., Kahng, B., Kim, D.: Universal behavior of load distribution in scale-free networks. Physical Review Letters **87**(27), 278,701 (2001)
44. Goh, K., Oh, E., Kahng, B., Kim, D.: Betweenness centrality correlation in social networks. Physical Review E **67**(1), 17,101 (2003)
45. Goldenberg, A., Zheng, A., Fienberg, S., Airoldi, E.: A survey of statistical network models. Foundations and Trends in Machine Learning **2**(2), 129–233 (2010)
46. Golub, G., Loan, C.V.: Matrix Computations. Johns Hopkins University Press (1996)
47. Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: Proceedings of the 2005 Workshop on Privacy in the Electronic Society, pp. 71–80. ACM (2005)

48. Guimera, R., Amaral, L.N.: Functional cartography of complex metabolic networks. Nature **433**(7028), 895–900 (2005)
49. Guimera, R., Mossa, S., Turtschi, A., Amaral, L.: The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. Proceedings of the National Academy of Sciences **102**(22), 7794 (2005)
50. Han, J., Kamber, M.: Data Mining: Concepts and Techniques - 2nd Edition. Morgan Kaufmann Publishers (2006)
51. Hansen, D., Smith, M., Shneiderman, B.: Analyzing social media networks with NodeXL: Insights from a Connected World. Elsevier (2010)
52. Harel, D., Koren, Y.: A fast multi-scale method for drawing large graphs. In: Proceedings of the Conference on Advanced Visual Interfaces, pp. 282–285 (2000)
53. Jeh, G., Widom, J.: SimRank: a measure of structural-context similarity. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543 (2002)
54. Jeong, H., Tombor, B., Albert, R., Oltvai, Z., Barabási, A.: The large-scale organization of metabolic networks. Nature **407**(6804), 651–654 (2000)
55. Kim, M., Han, J.: CHRONICLE: A Two-Stage Density-Based Clustering Algorithm for Dynamic Networks. In: Proceedings of the International Conference on Discovery Science, Lecture Notes in Computer Science, pp. 152–167. Springer (2009)
56. Kleinberg, J.: The small-world phenomenon: an algorithm perspective. In: Proceedings of the 32nd Symposium on Theory of Computing, pp. 163–170. ACM (2000)
57. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46**(5), 604–632 (1999)
58. Kumar, R., Novak, J., Tomkins, A.: Structure and evolution of online social networks. Link Mining: Models, Algorithms, and Applications pp. 337–357 (2010)
59. Kurant, M., Markopoulou, A., Thiran, P.: On the bias of breadth first search (bfs) and of other graph sampling techniques. In: Proceedings of the 22nd International Teletraffic Congress, pp. 1–8 (2010)
60. Latora, V., Marchiori, M.: A measure of centrality based on network efficiency. New Journal of Physics **9**, 188 (2007)
61. Leicht, E., Holme, P., Newman, M.E.J.: Vertex similarity in networks. Physical Review Part E **73**(2), 026,120 (2006)
62. Leskovec, J.: Stanford Network Analysis Package (SNAP). URL `http://snap.stanford.edu/`
63. Leskovec, J., Faloutsos, C.: Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 631–636 (2006)
64. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 177–187 (2005)
65. Leskovec, J., Lang, K., Mahoney, M.: Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th International Conference on the World Wide Web, pp. 631–640. ACM (2010)
66. Mathioudakis, M., Koudas, N.: Efficient identification of starters and followers in social media. In: Proceedings of the International Conference on Extending Database Technology, pp. 708–719. ACM (2009)

67. McCown, F., Nelson, M.: What happens when Facebook is gone? In: Proceedings of the 9th Joint Conference on Digital Libraries, pp. 251–254. ACM (2009)
68. Milgram, S.: The small world problem. Psychology Today **2**(1), 60–67 (1967)
69. Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, pp. 29–42. ACM (2007)
70. Newman, M.: Scientific collaboration networks. I. Network Construction and Fundamental Results. Physical Review E **64**(1), 16,131 (2001)
71. Newman, M.: The Structure and Function of Complex Networks. SIAM Review **45**(2), 167 (2003)
72. Newman, M., Barabasi, A., Watts, D.: The structure and dynamics of networks. Princeton University Press (2006)
73. Newman, M., Leicht, E.: Mixture models and exploratory analysis in networks. Proceedings of the National Academy of Sciences **104**, 9564–9 (2007)
74. Onnela, J., Reed-Tsochas, F.: The spontaneous emergence of social influence in online systems. Proceedings of the National Academy of Science **107**, 18,375 (2010)
75. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043), 814–818 (2005)
76. Palmer, C., Steffan, J.: Generating network topologies that obey power laws. In: Global Telecommunications Conference, vol. 1, pp. 434–438. IEEE (2002)
77. Partow, A.: General Purpose Hash Function Algorithms. URL `http://www.partow.net/programming/hashfunctions/`
78. Perer, A., Shneiderman, B.: Balancing systematic and flexible exploration of social networks. IEEE Transactions on Visualization and Computer Graphics pp. 693–700 (2006)
79. Perer, A., Shneiderman, B.: Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In: Proceeding of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 265–274. ACM (2008)
80. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proceedings of the National Academy of Science **101-9**, 2658–2663 (2004)
81. Redner, S.: How popular is your paper? An empirical study of the citation distribution. The European Physical Journal B **4**(2), 131–134 (1998)
82. Romero, D., Galuba, W., Asur, S., Huberman, B.: Influence and passivity in social media. In: Proceedings of the 20th International Conference Companion on World Wide Web, pp. 113–114. ACM (2011)
83. Romero, D., Kleinberg, J.: The Directed Closure Process in Hybrid Social-Information Networks, with an Analysis of Link Formation on Twitter. In: Proceedings of the 4th International Conference on Weblogs and Social Media (2010)
84. Seidel, R.: On the all-pairs-shortest-path problem. In: Proceedings of the 24th Symposium on Theory of Computing, pp. 745–749. ACM (1992)
85. Snasel, V., Horak, Z., Abraham, A.: Understanding social networks using formal concept analysis. In: Proceedings of the Web Intelligence/IAT Workshops, pp. 390–393. IEEE (2008)

86. Snasel, V., Horak, Z., Kocibova, J., Abraham, A.: Reducing social network dimensions using matrix factorization methods. In: International Conference on Advances in Social Network Analysis and Mining., pp. 348–351. IEEE (2009)
87. Song, X., Chi, Y., Hino, K., Tseng, B.: Identifying opinion leaders in the blogosphere. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management, pp. 971–974. ACM (2007)
88. Travers, J., Milgram, S.: An experimental study of the small world problem. Sociometry **32**(4), 425–443 (1969)
89. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge University Press (1994)
90. Watts, D.: Small worlds: the dynamics of networks between order and randomness. Princeton University Press (2004)
91. Watts, D., Strogatz, S.: Collective dynamics of small-world networks. Nature **393**(6684), 440–442 (1998)
92. Wilson, C., Boe, B., Sala, A., Puttaswamy, K., Zhao, B.: User interactions in social networks and their implications. In: Proceedings of the 4th European Conference on Computer Systems, pp. 205–218. ACM (2009)
93. Ye, S., Lang, J., Wu, F.: Crawling Online Social Graphs. In: Proceedings of the 12th International Asia-Pacific Web Conference, pp. 236–242. IEEE (2010)