Community Structure Discovery in Facebook

Emilio Ferrara

Department of Mathematics
University of Messina
V.le F. Stagno D'Alcontres n. 31, 98166, Italy
E-mail: eferrara@unime.it

Abstract: In this work we present a large-scale community structure detection and analysis of Facebook, which gathers more than 500 millions users at 2011. Characteristics of this social network have been widely investigated during the last years. Related works focus on analyzing its community structure on a small scale, usually from a qualitative perspective. In this study we consider a significant sample of the network. Data, acquired mining the Web platform, have been collected adopting two different sampling techniques. We investigated the structural properties of these samples in order to discover their community structure. Two well-known clustering algorithms, optimized for complex networks, have been here described and adopted. Results of our analysis show the emergence of a well-defined community structure inside Facebook, that is characterized by a power law distribution in the size of the communities. Moreover, the identified communities share an high degree of similarity, regardless the adopted detection algorithm.

Keywords: data mining, complex networks, community mining, community detection, complexity measures, performance measures

Reference to this paper should be made as follows: Ferrara, E. (2012) 'Community Structure Discovery in Facebook', *Int. J. Social Network Mining*, Vol. 1, No. 1, pp.67–90.

Biographical notes: Emilio Ferrara received his M.Sc. degree magna cum laude in Computer Science from the University of Messina in July 2008 and is currently a Ph.D. student of Mathematics and Computer Science at the same University. During his Ph.D. he has been a visiting student at the Vienna Technische Universitat and at the Royal Holloway University of London, and an intern at the Lixto Software GmbH. His research interests focus on the Semantic Web, in particular on topics such as adaptive algorithms for automatic extraction of information from Web sources, analysis of online social networks, social media, biological networks, folksonomies, virtual and dynamic environments.

1 Introduction

The problem of modeling and studying the structure of networks attracted the attention of a huge amount of work, in several areas, such as Social Sciences, Physics and Computer Science. Different applications of the network analysis include the study of social, biological and technological networks. During the last years the Online Social Network (OSN) phenomenon spread at an incredible high growth rate. The increasing popularity of platforms of online social-networking attracted the attention of both computer and social scientists because of unique characteristics and social importance of these networks. Several work has been conducted to analyze properties and features of OSNs both from a quantitative and a qualitative perspective.

The social role of Online Social Networks is to help people to enhance the connections among them in the context of Internet. On the one hand, these relationships are very tight over some areas of the social life of each user, such as family, colleagues, friends, and so on. On the other, the outgoing connections with other individuals not belonging to any of these categories are less likely to happen. This effect reflects in a phenomenon called *community structure*. A community is formally defined as a sub-structure present into the network that represents connections among users, in which the density of relationships within the members of the community is much greater than the density of connections among communities. From a structural perspective, this is reflected by a graph which is very sparse almost everywhere but dense in local areas, corresponding to clusters (i.e., communities).

A lot of different motivations to investigate the community structure of a network exist. From a scientific perspective, it is possible to put into evidence interesting properties or hidden information about the network itself. Moreover, intuitively individuals that belong to a same community, share some similarities and possibly have common interests or are connected by a specific relationship in the real-world. These aspects arise a lot of commercial and scientific applications; in the first category we cite, for example, marketing and competitive intelligence investigations and recommender systems. In fact, users belonging to a same community could share tastes or interests in similar products. In the latter, models of diseases propagation and distribution of information have been largely investigated in the context of social networks.

The problem of discovering the community structure of a network has been approached in several different ways. A common formulation of this problem is to find a partitioning $V = (V_1 \cup V_2 \cup \cdots \cup V_n)$ of disjoint subsets of vertices of the graph $G = (V, E)$ representing the network, in a meaningful manner. The probably most popular quantitative measure to evaluate the community structure in a network, namely *network modularity* (usually indicated by Q), has been proposed by (Girvan and Newman, 2002; Newman and Girvan, 2004). In its wider formulation (Duch and Arenas, 2005), the problem of finding a partitioning of a given network that maximizes the network modularity value is not computationally affordable because it is NP-hard. For this reason, several approximate approaches, for example based on heuristics, optimization strategies or swarm intelligence techniques, have been advanced. They are discussed in Section 2.

Two intuitive problems can be already sketched. The first one arises when partitioning the vertices into disjoint subsets, because each entity of the network could possibly belong to several different communities. The problem of overlapping communities has been already investigated in literature (Palla et al., 2005; Lee et al., 2010; McDaid and Hurley, 2010). The latter problem is represented by networks in which it makes sense that an individual does not belong to any group. In the formulation introduced above, we imposed that, regardless the overlapping communities are considered or not, each individual is required to belong at least to one group. This requirement could make sense for several networks, but is unaffordable in those cases in which some individuals could remain isolated from the rest of the network, as recently put into evidence by (Hunter et al., 2008). Such a case commonly happens in real and online social networks, as reported by recent social studies (Hampton et al., 2007).

In this work we analyze the community structure of Facebook on a large scale. We acquire two different samples of the network of relationships among the users of the social network. Each of them contains millions entities and, for this reason, we adopt two fast and efficient community detecting algorithms optimized for complex networks, working without any a-priori knowledge, in order to discover the emergent community structure.

The rest of the paper is organized as follows. Section 2 covers the background and the related work about detecting the community structure within a network, with particular attention to the specific area of Online Social Networks. Because the details about the Facebook social network structure are not public, we shortly depict in Section 3 the process of data collection, in order to make it clear how it has been possible for us to sample the Facebook social network. Two different sampling techniques are here presented and adapted to the domain of our problem. Section 4 introduces some details about two fast community detection algorithms we have adopted to detect the community structure of Facebook. Experimental results, performance evaluation and data analysis are shown in Section 5. We describe the methodology behind this work, illustrating the aspects on which we focused during our experimentation. Details related to the formulation of the problem and the choice of the solutions are illustrated. Section 6 concludes the paper, summarizing our achievements and depicting some interesting future issues.

2 Related Work

Several studies have been conducted in order to investigate the community structure of real and online social networks (Karrer et al., 2008; Shah and Zaman, 2010; Traud et al., 2011; Zhao et al., 2011). They all rely on the algorithmic background of detecting communities in a network. There are several comprehensive surveys to this problem, addressed to non practitioner readers, such as (Porter et al., 2009; Fortunato, 2010).

The related literature could be classified in two categories: i) partitioning algorithms; ii) overlapping nodes community detection algorithms.

2.1 Partitioning Algorithms

In its general formulation, the problem of finding communities in a network is intended as a data clustering problem, thus solvable assigning each vertex of the network to a cluster, in a meaningful way. There are essentially two different and widely adopted approaches to solve this problem; the first is the spectral clustering (Hagen and Kahng, 2002) which relies on optimizing the process of cutting the graph; the latter is based on the concept of *network modularity*.

The problem of minimizing the graph-cuts is NP-hard, thus an approximation of the exact solution can be obtained by using the spectral clustering (Ng et al., 2001), exploiting the eigenvectors of the Laplacian matrix of the network. We recall that the Laplacian matrix L of a given graph has components $L_{ij} = k_i\delta(i,j) - A_{ij}$, where k_i is the degree of a node i, $\delta(i,j)$ is the Kronecker delta (that is, $\delta(i,j) = 1$ if and only if $i = j$) and A_{ij} is the adjacency matrix representing the graph connections. This process can be performed using the concept of ratio cut (Wei and Cheng, 1989; Hagen and Kahng, 2002), a function which can be minimized in order to obtain large clusters with a minimum number of outgoing interconnections among them. The main limitation of the spectral clustering is that it requires in advance to define the number of communities present in the network and their size. This makes it unsuitable if one wants to discover the number and the features of existing communities in a given network. Moreover, as demonstrated by (Shah and Zaman, 2010), it does not work very well in detecting small communities within densely connected networks.

The network modularity concept can be explained as follows: let consider a network, represented by means of a graph $G = (V, E)$, which has been partitioned into m communities; its corresponding value of network modularity is

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

assuming l_s the number of edges between vertices belonging to the s-th community and d_s the sum of the degrees of the vertices in the s-th community. Intuitively, high values of Q imply high values of l_s for each discovered community; thus, detected communities are dense within their structure and weakly coupled among each other. Because the task of maximizing the function Q is NP-hard, several approximate techniques have been presented during the last years.

Let us consider the Girvan-Newman algorithm (Girvan and Newman, 2002; Newman and Girvan, 2004; Newman, 2006b). It first calculates the *edge betweenness* $B(e)$ of any given edge e in a network S, 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 vertices of S, $np(n_i, n_l)$ is the number of the shortest paths between n_i and n_l and $np_e(n_i, n_l)$ is the number of the shortest paths between n_i and n_l containing e. The GN algorithm is based on the assumption that it is possible to maximize the value of Q deleting edges with a high value of betweenness. This, because they connect vertices belonging to different

communities. Starting from this intuition, first the algorithm ranks all the edges with respect to their betweenness, thus removes the most influent, calculates the value of Q and iterates the process until a significant increase of Q is obtained. At each iteration, each connected component of S identifies a community. Its cost is $O(n^3)$, being n the number of vertices in S; intuitively, it is unsuitable for large-scale networks.

A tremendous number of improved versions of this approach has been provided in the last years, such as the fast clustering algorithm provided by (Clauset et al., 2004; Clauset, 2005), running in $O(n \log n)$ on sparse graphs; the extremal optimization method proposed by (Duch and Arenas, 2005), based on a fast agglomerative approach with $O(n^2 \log n)$ time complexity; the Newman-Leicht (Newman and Leicht, 2007) mixture model based on statistical inferences; other maximization techniques by (Newman, 2006a) based on eigenvectors and matrices.

Concluding, another different approach of partitioning is the "core-periphery", introduced by (Borgatti and Everett, 1999; Everett and Borgatti, 1999); it relies on separating a tight core from a sparse periphery.

2.2 Overlapping Nodes Community Detection

Recently, the problem of discovering the community structure in a network included the possibility of finding overlapping nodes belonging to different communities at the same time. One of the first approach has been presented by (Palla et al., 2005) and has attracted a lot of attention by the scientific community. A lot of efforts have been spent in order to advance novel possible strategies. For example, an interesting approach has been proposed by (Gregory, 2007), that is based on an extension of the Label Propagation Algorithm adopted in this work. On the other hand, an approach in which the hierarchical clustering is instrumental to find the overlapping community structure has been proposed by (Lancichinetti et al., 2009). Finally, during latest years some novel techniques have been proposed (McDaid and Hurley, 2010; Lee et al., 2010).

2.3 Our Contribution to the State-of-the-art

Our contribution can be summarized as follows. As the best of our knowledge this work presents for the first time the analysis of the community structure of a large Online Social Network by adopting quantitative techniques. To do so, first of all we collect a sample of the Facebook social network. Once data have been gathered, the emergent community structure is put into evidence and analyzed both from a quantitative and a qualitative perspective. The results of our analysis enrich the current knowledge on the features of the communities that characterize Online Social Networks and possibly it is helpful in order to better understand those processes that underly the formation of groups and communities in social networks.

3 Data Collection

Although the first purpose of this work is to analyze the community structure of the Facebook social network, some information about the process of data collection are here provided. Furnished details help in better understanding the further investigation on how the sampling methodology and the adopted techniques of information collection may affect the obtained results.

In the following, we briefly discuss the architecture of the designed sampling platform, the logic behind the process of data acquisition and preparation.

3.1 Sampling the Facebook Network

In order to collect data from a social network, several possible direction of investigation could be planned. In our case, for example, Facebook users' profile are enriched by a lot of semantic information regarding their activity and the contents they produce or share on the platform. In detail, it could be possible to extract and analyze information that characterizes the users of a given social network, on a large scale. Several interesting research topics investigate issues of great interest such as sentiment analysis, information propagation and the diffusion of consensus about common arguments among users of social networks.

Facebook, similarly to other Online Social Network services, does not provide information about the social structure of its network. Similarly, they do not diffuse personal information about users, the content they like or produce, they way they adopt the social network and their social behaviors. This, in order to protect the privacy of its users and for commercial interests. In fact, the main business of Facebook is based on exploiting personal data to provide targeted ads services to advertising companies. For all these reasons, a semantic analysis of the Facebook social network could be unfeasible without occurring in issues.

In the light of these assumptions, during the phase of extraction of data from the Facebook social network, we did not inspect, acquire or store personal information about users or their social behaviors. In particular, the aim of this work is the reconstruction of the social connections among users.

To this purpose, we designed a crawler which visited the publicly accessible friend-list Web pages of specific users, selected with respect to a particular criterion (details follow). After that, the friendship relations among discovered users have been stored in an anonymized format. At the end of this process, a partial sample of the whole graph has been obtained, which contains millions nodes and edges (i.e., users and friendship connections among them).

In this work we used the acquired data to devise, as the best of our knowledge, the first large scale community structure investigation of the Facebook social network.

3.2 The Data Collection Platform Architecture

The architecture of the designed sampling platform can be schematized as in Figure 1. We devised a Java cross-platform data mining agent, which implements the logic of the crawler, relying on the Apache HTTP Library (http://httpd.apache.org/apreq) as interface for transferring data through the

Figure 1 The data collection platform architecture

Figure 2 The logic of the Facebook crawler

Web. While running (possibly, in parallel), the agent(s) query the Facebook server(s) in order to acquire the specific friend-list Web pages of the required users. Received data are processed and collected on our server in anonymized format. Finally, at the end of the data collection process, data are post-processed, cleaned and filtered according to further requirements.

3.3 The Sampling Methodology

Regardless the sampling algorithm implemented, the logic of the designed mining agent is reported in Figure 2. The devised automaton requires a preparative step during which some configuration aspects are set up, such as adopted sampling algorithm, termination criterion/a to be met to conclude the execution, and maximum running time. The mining task could be started from scratch or be resumed from a previous state. Thus, during its execution, the agent(s) visit the Web pages containing the friend-list of the user to be inspected, following the directives of the sampling methodology algorithm, and extract the friendship relations represented in there. Data are stored in a compact format in order to save I/O operations and then are anonymized, in order not to store any kind of possibly private data (such as the user-IDs). The data collection process concludes if any of the defined termination criteria is met.

3.3.1 The breadth-first-search visit

The first sampling methodology has been implemented as a breadth-first-search (BFS), an uninformed traversal algorithm with the goal of visiting 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, a termination criterion is met). This sampling technique has several advantages, such as the ease of implementation and the efficiency. For these reasons it has been adopted in a variety of OSNs mining studies (Chau et al., 2007; Mislove et al., 2007; Gjoka et al., 2010; Catanese et al., 2011). On the other hand, bias of data towards high degree nodes have been proved to by (Kurant et al., 2010) to be introduced by incomplete BFS visits. We further investigate effects of this phenomenon on our results in Section 5.1.2.

During our experimentation, the adopted "seed" was the user profile used to login-on the Facebook platform. The termination criteria were: i) at least the third sub-level of friendship was completely covered; ii) the mining process exceeded 10 days of running time. The observation of a short time-limit ensured a negligible effect of evolution of the network structure. The dimension of the obtained (partial) graph of the Facebook network has been adopted as yardstick for the uniform sampling process.

3.3.2 The uniform sampling

The second sampling methodology that has been chosen is a rejection-based sampling technique, namely "Uniform" sampling. The main advantage of this technique is that it is unbiased for construction, at least in its formulation for the Facebook network. Details about its definition are provided by (Gjoka et al., 2010). The process consists of the generation of an arbitrary number of user-IDs, randomly distributed in the domain of assignment of the Facebook user-ID system. In our case, it is the space of the 32-bit numbers, thus the maximum amount of assignable user-IDs is 2^{32}, about 4 billions. As August 2010, the number of subscribed users on the Facebook platform was about 500 millions, thus the probability of randomly generating an existing user-ID was about 1/8. Thus, first we generated a number of random user-IDs, lying in the interval $[0, 2^{32} - 1]$, equal to the dimension of the BFS-sample multiplied by 8. Then, we queried Facebook for their existence. Our expectation was to obtain a sample of comparable dimensions with respect to the BFS-sample. We obtained a slightly smaller sample, due to the restrictive privacy settings imposed by some users, who configured their profile preventing the public accessibility of their friend-lists.

3.4 Description of the Acquired Datasets

All the user-IDs contained in the samples have been anonymized using a 48-bit hashing functions, in order to hide references to the users and their connections. Data have been post-processed for a cleansing step, during which all the duplicates have been removed, and the integrity and congruency of the datasets have been verified. The characteristics of the datasets are reported in Table 1. The size of both the samples is in the order of millions nodes and edges. These networks reflect some of the well-known properties of the social networks:

- The "Small world" effect: it is typical of networks with very small diameter, with respect to the number of nodes, and a huge connected component;
- The power law degree distribution: the network presents a huge number of nodes having a small number of connections and vice-versa;
- The emergence of a community structure: the aim of this work is to investigate the presence of a community structure in the considered Facebook social network.

In detail, the connected component of the "Uniform" sample includes about the 95% of the total nodes (differently with respect to the connected component of the BFS sample that covers all the network). This seems that a significantly

Dataset	No. Visited Users	No. Discovered Neighbors	No. Total Edges	Coverage
BFS	63.4K	8.21M	12.58M	98.98%
Uniform	48.1K	7.69M	7.84M	94.96%

Avg. Degree	Biggest Eigenvalue	Effective Diameter	Avg. Clustering Coefficient	Density
396.8	68.93	8.75	0.0789	0.626%
326.0	23.63	16.32	0.0471	0.678%

Table 1 BFS and Uniform datasets description (crawling period: August 2010)

small part of the users appears to be disconnected from the *giant component* that connects all the others. In this context, a small part of these users remains *isolated* in the network. Their contribution to the features of the community structure of the social network is not relevant and for this reason during our further analysis they are not taken into account. However, their existence is relevant to the study of different social phenomena, for example the *social isolation* (Hampton et al., 2007). The value of the clustering coefficient, that has been proven to be a relevant indicator of the possible emergence of a community structure inside a network, is high for both the samples. Similarly, also the value of density of connections among nodes is high.

In the light of these clues, in the following we introduce the methodology that has been adopted in order to verify the existence of a well-defined community structure inside the Facebook social network.

4 Community Structure Discovery

The detection of a community structure inside a large network is a complex and computationally expensive task. Community detection algorithms such those originally presented by (Girvan and Newman, 2002; Newman and Girvan, 2004) or by (Hagen and Kahng, 2002), are not viable solutions, respectively because too expensive for the large-scale of the Facebook sample we gathered, or because they require *a priori* knowledge. Fortunately, several optimizations have been proposed during last years. To our purposes, we adopted two fast and efficient optimized algorithms, whose performance are the best to date proposed in literature. LPA (Label Propagation Algorithm), presented by (Raghavan et al., 2007), and FNCA (Fast Network Community Algorithm), more recently described by (Jin et al., 2009), have been adopted to detect communities from the collected samples of the network. A description of their functioning follows, in particular in the context of our study.

4.1 Label Propagation Algorithm

LPA (Label Propagation Algorithm) (Raghavan et al., 2007) is a near linear time algorithm for community detection. Its functioning is very simple, considered its computational efficiency. LPA uses only the network structure as its guide, is optimized for large-scale networks, does not follows any a priori defined objective function and does not require any prior information about the communities. In addition, this technique does not require to define in advance the number

of communities present into the network or their size. Labels represent unique identifiers, assigned to each vertex of the network.

Its functioning is reported as described in (Raghavan et al., 2007):

Step 1 To initialize, each vertex is given a unique label;

Step 2 Repeatedly, each vertex updates its label with the one used by the greatest number of neighbors. If more than one label is used by the same maximum number of neighbors, one is chosen randomly. After several iterations, the same label tends to become associated with all the members of a community;

Step 3 Vertices labeled alike are added to one community.

Authors themselves proved that this process, under specific conditions, could not converge. In order to avoid deadlocks and to guarantee an efficient network clustering, they suggested to adopt an "asynchronous" update of the labels, considering the values of some neighbors at the previous iteration and some at the current one. This precaution ensures the convergence of the process, usually in few steps. (Raghavan et al., 2007) ensure that five iterations are sufficient to correctly classify 95% of vertices of the network. After some experimentation, we found that this forecast is too optimistic, thus we elevated the maximum number of iterations to 50, finding a good compromise between quality of results and amount of time required for computation.

A characteristic of this approach is that it produces groups that are not necessarily contiguous, thus it could exist a path connecting a pair of vertices in a group passing through vertices belonging to different groups. Although in our case these condition would be acceptable, we adopted the suggestion provided by the authors to devise a final step to split the groups into one or more contiguous communities. Authors proved its near linear computational cost.

Recently, a great attention has been captured by the possibility of discovering the community structure in a network finding overlapping nodes belonging to different communities at the same time. An interesting approach has been proposed by (Gregory, 2007), that is based on an extension of the Label Propagation Algorithm previously described.

4.2 Fast Network Community Algorithm

The second efficient algorithm that has been chosen for our analysis is called FNCA (Fast Network Community Algorithm) (Jin et al., 2009). The main advantage of FNCA is that it does not require to define in advance the number of communities present into the network, or their size. This aspect makes it suitable for the investigation of the unknown community structure of a large network, such as in the case of Facebook.

FNCA is an optimization algorithm which aims to maximize the value of the network modularity function, in order to detect the community structure of a given network. The network modularity function has been introduced by (Newman and Girvan, 2004) and has been largely adopted in the last few years by the scientific community (Boccaletti et al., 2006; Du et al., 2007; Fortunato, 2010).

Given an undirected, unweighted network $G = (V, E)$, let $i \in V$ be a vertex belonging to the community $r(i)$ denoted by $c_r(i)$; the network modularity function is defined as follows

$$Q = \frac{1}{2m} \sum_{ij} \left[\left(A_{ij} - \frac{k_i k_j}{2m} \right) \times \delta \left(r(i), r(j) \right) \right] \tag{3}$$

where A_{ij} is the element of the adjacency matrix $A = (A_{ij})_{n \times n}$ representing the network, whose value is $A_{ij} = 1$ if i and j are tied by an edge, $A_{ij} = 0$ otherwise. The function $\delta(u, v)$, namely *Kronecker delta*, is equal to 1 if $u = v$ and 0 otherwise. The value k_i represents the degree of a vertex i, defined as $k_i = \sum_j A_{ij}$ while m is the maximum possible number of edges in the network, defined as $m = \frac{1}{2} \sum_{ij} A_{ij}$.

Equation 3 can be rewritten as

$$Q = \frac{1}{2m} \sum_i f_i, \qquad f_i = \sum_{j \in c_{r(i)}} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \tag{4}$$

where the function f represents the difference between actual and expected number of edges which fall within communities, from the "perspective" of each node of the network, thus indicating how strong the community structure is.

Any node of the network could evaluate the value of its f function only considering local information (i.e., information about its community). Moreover, if the local effect of relabeling a node, without changing the labels of others, is that the value of its f function increases, the global effect is that also the network modularity increases.

Given these assumptions, (Jin et al., 2009) devised a fast community detection algorithm, optimized for complex networks, adopting local information strategies. FNCA relies on the consideration that, in networks with an emergent community structure, each node should be labeled alike one of its neighbors, otherwise it is a cluster itself. Thus, each node needs to calculate its f function only for the labels of its neighbors, instead of for all the nodes of the network.

Moreover, authors put into evidence that, if the labels of neighbors of one node do not change at last iteration, the label of that node is less likely to change in the current iteration. This provides a speed-up strategy, putting nodes which satisfy this condition, in an "inactive" state, not requiring the update of their labels. Because this weak condition may fail, it is important to immediately "wake up" those nodes which do not satisfy this constraint anymore, at each iteration.

Also this algorithm, like LPA, could not converge. In our experimentation we defined a termination criterion of 50 iterations, obtaining good results also with our large-scale samples.

The time complexity of FNCA is $O(T \cdot n \cdot k \cdot c)$, where T is the number of maximum iterations, n the number of total nodes, k the average degree of all nodes, and c the average community size at the end the algorithm execution. Furthermore, with the support of the analysis in literature (Leskovec et al., 2009), for large-scale networks, FNCA is a near linear algorithm.

Algorithm	No. Communities	Q	Time (s)
BFS (8.21 M vertices, 12.58 M edges)			
FNCA	50,156	0.6867	5.97e+004
LPA	48,750	0.6963	2.27e+004
Uniform (7.69 M vertices, 7.84 M edges)			
FNCA	40,700	0.9650	3.77e+004
LPA	48,022	0.9749	2.32e+004

Table 2 Results of the Community Detection on Facebook

5 Experimentation

The experimental results obtained by using LPA and FNCA on the Facebook network are reported in Table 2. Both the algorithms show good performance while applied to this network. A very compact community structure has been highlighted by using both the algorithms. In detail, resulting values of Q are almost identical with respect to the considered sample; moreover, the number of detected communities is very similar.

5.1 Methodology of Investigation

By analyzing the obtained community structures we considered the following aspects: i) the distribution of the dimensions of the obtained clusters (i.e. the number of members constituting each detected community), and, ii) the qualitative composition of the communities and the degree of similarity among different sample sets (i.e., those obtained by using different algorithms and sampling techniques).

5.1.1 Community distribution: Uniform sample

The analysis of the community structure of a network from a quantitative perspective may start with the study of the distribution of the dimension of the communities. Our investigation started considering first the "Uniform" sample, which is known to be unbiased for construction. Results obtained are adopted to investigate the possible bias introduced by the BFS sampling technique, as discussed in the following.

As depicted by Figures 3 and 4, results obtained by using the two different algorithms on the "Uniform" sample are interesting and deserve explanations. In detail, analytical results (as reported in Table 2) and figures put into evidence that both the algorithms identified a similar amount of communities, which is reflected by almost identical values of network modularity in the two different sets. Moreover, identified communities are themselves, the most of the times, of the same dimensions, regardless the adopted community detection algorithm.

These aspects lead us to advance different hypothesis on the characteristics of the community structure of Facebook detected on the unbiased "Uniform" sample. The first consideration regards the distribution of the size of the communities. Both the distributions obtained by using the LPA and the FNCA algorithm show a characteristic power law distribution. This is emphasized by Figures 3 and 4,

which represents the distributions of the dimension of communities obtained by using, respectively, FNCA and LPA, applied on the "Uniform" sample. In Figure 3, the clusters size distribution obtained by using FNCA is fitted to a power law function ($\gamma = 0.45$) which effectively approximates its behavior. Similarly, Figure 4 represents the clusters size distribution produced by LPA, which gives results in a shorter interval (i.e., [0,500] with respect to [0,1000] used in Figure 3), well fitting to a power law function ($\gamma = 0.37$). A first consideration is that the communities detected by using the LPA algorithm appears to be slightly displaced to bigger values with respect to those represented by the FNCA in the first quartile, while the number of communities greater than 400 members quickly decreases.

These results permit us to draw two conclusions:

- On a large scale, to the best of our knowledge, this is the first experimental analysis that proves that the size of the communities emerging in an Online Social Networks follows a well-defined power law distribution. This result is novel and validates the hypothesis, proved on a small scale on several real-world social networks (for example, (Traud et al., 2011)), that not only the degree distribution follows a scale-free behavior, but even the processes of aggregations among individuals of an Online Social Network can be effectively described by communities whose dimensions follow a power law. This results in the following point.

- Our analysis puts into evidence that, even on a large scale that is well represented by an OSN such as Facebook, people tend to aggregate principally in a large amount of small communities instead of in very large communities. This, at the same time, in our opinion demystifies some theoretical approaches of community detection such as the *core-periphery* (Borgatti and Everett, 1999; Everett and Borgatti, 1999).

- A rejection-based sampling methodology (such as the "Uniform" sampling) is appropriate to describe the community structure emerging on a large sample of a Online Social Network. Differently with respect to other approaches, it seems to preserve those characteristics that influence the distribution of the friendship relations thus well representing the community structure of large networks.

5.1.2 Community distribution: BFS sample

Results obtained by analyzing the BFS distribution, show partially different characteristics. Figures 5 and 6 show the cluster dimensions distribution by using, respectively, FNCA and LPA applied to the BFS sample. Both these distributions show some fluctuations if compared to the power law distribution adopted as possible fitting function. By using the FNCA (see Figure 5), the peak of the distribution is represented by those communities constituted of 10–30 members, then it sharply slopes depicting a first fluctuation around clusters of dimension around a hundred of members, and a second minor fluctuation around the three hundreds of members. A similar behavior is shown by the LPA algorithm (see Figure 6).

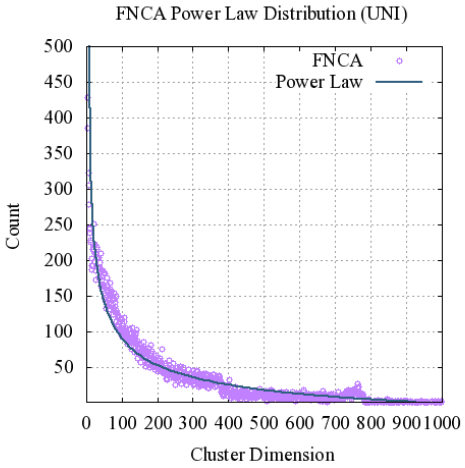

Figure 3 FNCA Power law distributions on the "Uniform" sample

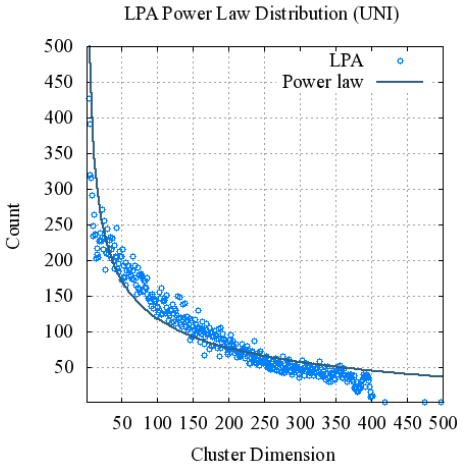

Figure 4 LPA Power law distributions on the "Uniform" sample

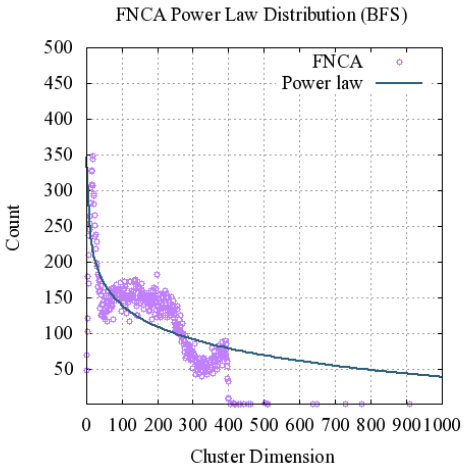

Figure 5 FNCA Power law distribution on the BFS sample

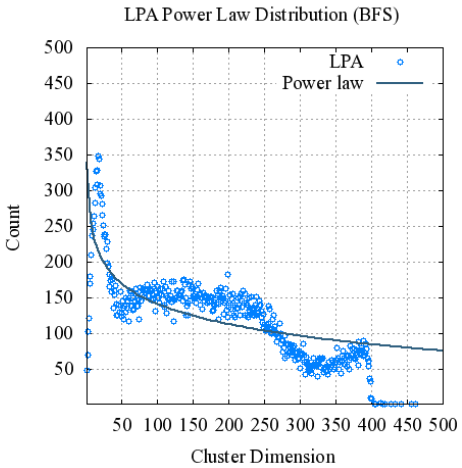

Figure 6 LPA Power law distribution on the BFS sample

The differences in the behavior between the BFS and "Uniform" samples distributions reflect accordingly with the adopted sampling techniques. In fact, in related works (Gjoka et al., 2010; Kurant et al., 2010), has been put into evidence the influence of the adopted sampling methods on the characteristics of the obtained sets, in particular focusing on the possible bias introduced by using the BFS algorithm, towards high degree nodes, if the BFS visit is incomplete (such as in our case, in which a sample of the whole network has been collected).

We could draw the conclusion that the adoption of the BFS sampling technique is not appropriate in the case one would to investigate the community structure of a large network whose complete sampling is not feasible, for example because of

constraints imposed by the network itself or by its dimension (such in the case of Facebook). On the other hand, the BFS sampling has been proved to be effective and efficient in the opposite cases.

5.1.3 Overlapping Rate between Distributions

The idea that two different algorithms could produce different community structures is not counterintuitive, but in our case we have some indications that the obtained results could share an high degree of similarity. To this purpose, in the following we investigate the similarities among the community structures obtained by using the two different algorithms, FNCA and LPA. This is represented by the overlapping rate calculated considering the distributions of the community dimensions from a quantitative perspective. To do so, we adopt a divergence measure, called *Kullback-Leibler divergence*, that is defined as

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{5}$$

where P and Q represent, respectively, the probability distribution that characterizes the behavior of the LPA and the FNCA community sizes, calculated on a given sample. In detail, let i be a given size such that $P(i)$ and $Q(i)$ represent the probability that a community of size i exists in the distribution P and Q. Intuitively, the KL divergence is helpful if one would like to calculate how much different is a distribution with respect to a given one.

In particular, being the KL divergence defined in the interval $0 \leq D_{KL} \leq \infty$, the smaller the value of KL divergence between two distributions, the more similar they are. In the light of this assumption, we calculated the pairwise KL divergences between the distributions discussed above, finding the following results:

- On the "Uniform" sample:
  - $D_{KL}(LPA||FNCA) = 0.007722$
  - $D_{KL}(FNCA||LPA) = 0.007542$
- On the BFS sample:
  - $D_{KL}(LPA||FNCA) = 0.003764$
  - $D_{KL}(FNCA||LPA) = 0.004292$

The values found by adopting the KL divergence put into evidence a strong correlation between the distributions calculated by using the two different algorithms on the two different samples.

From a graphical standpoint, we put into evidence the correlation found by means of the KL divergence, as follows. In Figures 7 and 8 a semi-logarithmic scale has been adopted. In Figure 7, we plotted together the distributions depicted in Figures 3 and 4 that represent the community structure of the "Uniform" sample. Similarly, Figure 8 shows the distributions presented in Figures 5 and 6 regarding the BFS sample.

By analyzing the distribution of the community sizes of the "Uniform" set, it emerges a perfectly linear behavior, that characterizes both the FNCA and the

LPA results. This agrees with the power law distributions previously emphasized, that well depict the behavior of the emergent community structure in that sample. Additionally, the two distributions are almost overlapping. A similar consideration holds for the BFS sample. Even though the distributions suffers of the spikes previously discussed, a strong correlation between them has been put into evidence both by the KL divergence and by the graphical representation. These indications gave us the opportunity of investigating from a qualitative perspective the characteristics of the community structure of Facebook.

In detail, a different consideration regarding the qualitative analysis on the similarity of the two different community structures is provided in the next Section. That kind of investigation aims at evaluating what members constitute the communities detected by adopting the algorithms previously introduced. Our findings prove that, regardless the adopted community detection algorithm, the communities discovered, not only are characterized by similar distributions of sizes, but are also mainly constituted of the same members. This finding proves that the emergent community structure in Facebook is well characterized and defined, according with the quantitative results we discussed above.

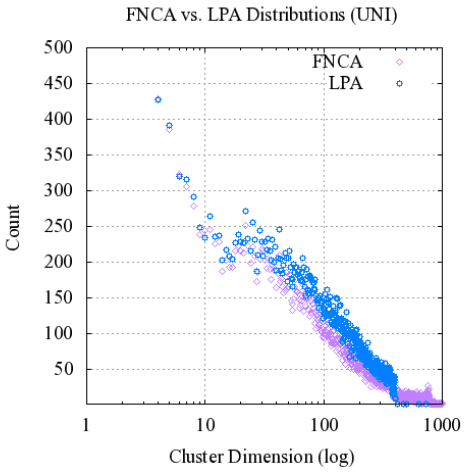

Figure 7 FNCA vs. LPA (UNI)

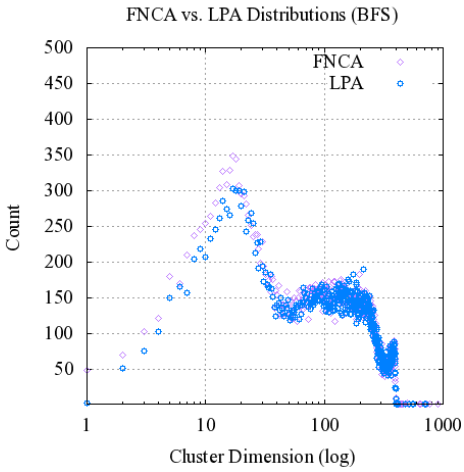

Figure 8 FNCA vs. LPA (BFS)

5.1.4 Community structure similarity

In this Section we introduce the methodology of investigation of the similarity among different community structures. A community structure is represented by a list of vectors which are identified by a "community-ID"; each vector contains the list of user-IDs (in anonymized format) of the users belonging to that specific community; an example is depicted in Table 3.

In order to evaluate the similarity of the community structures obtained by using the two algorithms, FNCA and LPA, a coarse-grained way to compare two

Community-ID	List of Members
community-ID$_1$	{user-ID$_a$; user-ID$_b$; ...; user-ID$_c$}
community-ID$_2$	{user-ID$_i$; user-ID$_j$; ...; user-ID$_k$}
...	{...}
community-ID$_N$	{user-ID$_x$; user-ID$_y$; ...; user-ID$_z$}

Table 3 Representation of community structures

sample sets would be by adopting a simple measure of similarity such as the *Jaccard coefficient*, defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

where A and B represent the two community structures. While calculating the intersection of the two sets, communities differing even by only one member would be considered different, while an high degree of similarity among them could be envisaged.

A more convenient way to compute the similarity among these sets is to evaluate the *Jaccard coefficient* at the finest level, comparing each vector of the former set against all the vectors in the latter set, in order to "match" the most similar ones. Under these assumptions, the *Jaccard coefficient* could be rewritten in its vectorial formulation as

$$J(\mathbf{v},\mathbf{w}) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \tag{7}$$

where M_{11} represents the total number of shared elements between vectors $\mathbf{v}$ and $\mathbf{w}$, M_{01} represents the total number of elements belonging to $\mathbf{w}$ and not belonging to $\mathbf{v}$, and, finally M_{10} the vice-versa. The result lies in $[0,1]$. The more two compared communities are similar, or, in other words, the more the constituting members of two compared communities are overlapping, the higher the value of the *Jaccard coefficient* computed this way.

An almost equivalent way to compute the similarity with an high degree of accuracy would be by applying the *Cosine similarity*, among each possible couple of vectors belonging to the two sets. The *Cosine similarity* is defined as

$$cos(\Theta) = \frac{A \cdot B}{||A||\ ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{8}$$

where A_i and B_i represent the binary frequency vectors computed on the list members over i.

Once matched the most similar pairs of communities between the two compared sets, the mean degree of similarity is computed by

$$\sum_{i=1}^{N} \frac{max\,(J(\mathbf{v},\mathbf{w})_i)}{N} \qquad \text{and} \qquad \sum_{i=1}^{N} \frac{max\,(cos(\Theta)_i)}{N} \tag{9}$$

where $max(J(\mathbf{v}, \mathbf{w})_i)$ and $max\,(cos(\Theta)_i)$ mean the highest value of similarity chosen among those calculated combining the vector i of the former set A with all the vectors of the latter set B, respectively adopting the *Jaccard coefficient* and the *Cosine similarity*. We obtained the results as in Table 4.

		Degree of Similarity FNCA vs. LPA			
Metric	Dataset	In Common	Mean	Median	Std. D.
J	BFS	2.45%	73.28%	74.24%	18.76%
	Uniform	35.57%	91.53%	98.63%	15.98%

Table 4 Similarity degree of community structures

As intuitively deducible by analyzing Figures 7 and 8, not only the community structures calculated by using the two different algorithms, FNCA and LPA, follow similar distributions with respect to the dimensions, but also the communities themselves are constituted mostly by the same members or by a congruous amount of common members.

From results it emerges that both the algorithms produce a faithful and reliable clustering representing the community structure of the Facebook network. Moreover, while the number of identical communities between the two sets obtained by using BFS and "Uniform" sampling, is not so high (i.e., respectively, $\simeq$2% and $\simeq$35%), on the contrary the overall mean degree of similarity is very high (i.e., $\simeq$73% and $\simeq$91%).

Considered the way we compute the mean similarity degree, as in Equation 9, this is due to the high number of communities which differ only for a very small number of components. Finally, the fact that the median, which identifies the second quartile of the samples is, respectively, $\simeq$75% and $\simeq$99% demonstrates the strong similarities of the produced result sets.

All these considerations graphically emerge by analyzing Figures 9 and 10, in which the higher the degree of similarity, calculated by using the *Jaccard coefficient*, the denser the distribution, in particular in the first quartile, becoming obvious for values near to 1. The unbiased characteristics of the "Uniform" sample reflect also in Figure 9, in which the similarity degree of the community structure is evident because the most of the values lie on the boundary zone near to 1. The degree of similarity of the community structure of the BFS sample, shown in Figure 10, appears more distributed all over the second half of the distribution, becoming denser in the first quartile.

5.1.5 Resolution limit and outliers

Recently (Fortunato and Barthélemy, 2007), in the context of detecting communities by adopting the *network modularity* as maximization function, a resolution limit has been put into evidence. In particular, in (Fortunato and Barthélemy, 2007), authors found that modularity optimization could fail in the detection of communities smaller than a given threshold that is strictly dependent on the characteristics of the studied network. This results in another effect, that is the creation of large communities that include a large part of the nodes of the network, without affecting the global value of network modularity.

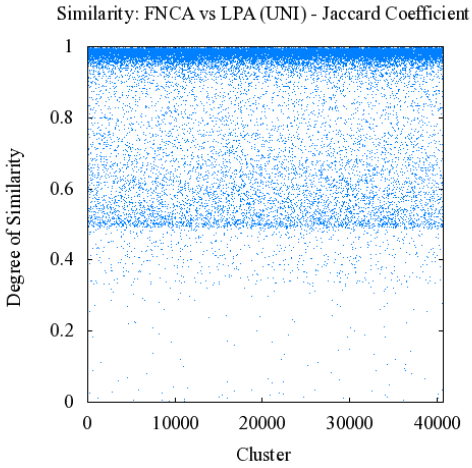

Figure 9 Jaccard FNCA vs. LPA (UNI)

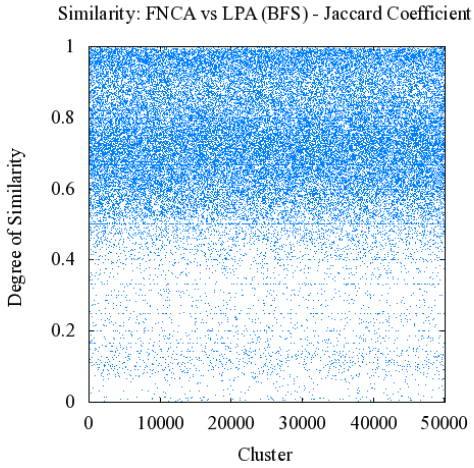

Figure 10 Jaccard FNCA vs. LPA (BFS)

Starting from the assumption that the characteristics of the network could be affected by the adopted sampling methodology, we investigated the effect of the resolution limit put into evidence by (Fortunato and Barthélemy, 2007) on our datasets. The results of our analysis respectively on the BFS and the "Uniform" samples could be discussed separately. On the former dataset, a small number of communities whose dimensions exceed those obtained in the distributions previously discussed, have been identified. Possibly, these large communities have been identified because of the resolution limit suffered by FNCA and LPA, that are both based on the modularity maximization.

Table 5 reports the amount of outliers, i.e., those communities that largely exceed the average dimension and are suspected of suffering from the problem of the resolution limit. From this analysis, it seems that LPA suffers the resolution limit slightly less than FNCA. In fact, a smaller number of outliers has been found by using the LPA method, with respect to the adoption of the FNCA algorithm, in the context of the BFS sample.

On the contrary, the "Uniform" sample apparently does not cause any problem of resolution limit, proving once again that this sample could be considered as ground-truth. By using the FNCA on the "Uniform" sample, a large number of communities whose dimension is slightly greater than one thousand members appear, that is coincident to the final part of the tail of the power law distribution, depicted in Figure 3. The LPA method even in the "Uniform" sample provides possibly the most reliable results, without incurring in any possible effect of resolution limit.

5.2 Discussion of Results

A summary of the results achieved with our analysis of the Facebook network can be drawn as follows.

- First of all, in this paper we put into evidence, as the best of our knowledge for the first time on a large scale, that the community structure of an online

		Amount with respect to Number of Members				
Set	Alg.	$\geq$ 1K	$\geq$ 5K	$\geq$ 10K	$\geq$ 50K	$\geq$ 100K
BFS	FNCA	4	1	2	1	1
	LPA	1	0	2	0	1
UNI	FNCA	81	0	0	0	0
	LPA	0	0	0	0	0

Table 5 The effect of the resolution limit on our datasets

social network presents a clear power law distribution of the dimension of the communities.

This result is independent with respect to the algorithm adopted to discover the community structure, and even (but in a less evident way) with respect to the sampling methodology adopted to collect the datasets. In detail, the former consideration is supported by the evidence that, by using both LPA and FNCA (two state-of-the-art well-known algorithm to detect the community structure of large networks), the dimension of the discovered communities fits well to a power law distribution with exponent approximately equal to 0.4. The latter result appears clear if we take in consideration the "Uniform" sampling methodology, that produced a dataset that could be adopted as ground-truth for forthcoming studies. On the other hand, this is the first experimental work that proves on a large scale the hypothesis, theoretically found by (Kurant et al., 2010), of the possible bias towards high degree nodes introduced by the BFS sampling methodology for incomplete visits of large graphs.

- As far as concerns the qualitative analysis of results obtained during our experimentation, we put into evidence that the Facebook social networks not only is characterized by a clear community structure, and that the size of the communities follow a power law distribution, but also that this community structure is qualitatively well defined. In fact, regardless the algorithm adopted in order to detect the underlying communities, it emerges that those communities detected by adopting LPA and FNCA share an high degree of similarity.

  In detail, if one admit to accept a small degree of dissimilarity (say, ϵ) among the components of the communities identified by using two different algorithms, it appears that the results show an high degree of similarity both by considering the "Uniform" and the BFS samples. In the former case, the maximum value of ϵ that represents the difference among any pair of communities, is not more than approximately the 8%. In the latter, this ϵ must be raised up to the 26%. This means that the communities identified by the two methods share at least approximately a 92% of members in common for the case of the "Uniform" samples, and a 74% for the BFS case.

- As for the community detection algorithm, we found that the LPA method has been proved to be a good choice among the heuristic methods based on local information in order to discover the underlying community structure of a large network. Results compared against another well-known similar

method, called FNCA, seems to be slightly better, in particular if we consider the well-known problem of the resolution limit (Fortunato and Barthélemy, 2007) that affect the process of community detection on a large scale.

In detail, LPA works well even in the context of a sample like the BFS that could possibly be affected by some bias towards high degree nodes (as recently put into evidence by (Kurant et al., 2010)). Results provided by the two algorithms on the "Uniform" sample are comparable. Even though the computational cost of these two techniques is very similar, we experienced that the LPA method performs slightly better that FNCA on our datasets.

For all these reasons, we can conclude that the more appropriate approach in order to study the community structure of large networks is to adopt a rejection-based sampling methodology (such as the "Uniform" sampling). Moreover, both the LPA method and the FNCA technique appear appropriate to this purpose. Recently, a novel high performance approach for the community detection, called "Louvain method" (Blondel et al., 2008) has been presented. The performance of this technique in the context of the community structure discovery on large scale Online Social Networks such as Facebook deserves further investigation.

6 Conclusion

In this work we presented, as the best of our knowledge, the first large-scale community structure investigation of the Facebook social network. We described the process of data collection, adopting two different techniques of sampling in order to obtain comparable data, to discover bias and to validate our assumptions. We adopted two fast and efficient algorithms already presented in literature, specifically optimized to detect the community structure on large-scale networks, such as Facebook, consisting of millions entities. A very strong community structure emerges by our analysis, and several characteristics have been highlighted by our experimentation, such as a typical power law distribution of the dimension of the clusters. We finally investigated the degree of similarity of the different community structures obtained by using the two algorithms on the respective samples, putting into evidence strong similarities.

A large-scale qualitative analysis of such a huge social network is challenging. For example, it is not trivial to infer motivations underlying groups formation. In order to enhance the accuracy of qualitative studies of these phenomena, visual representations of data on community structure, interconnections among clusters and community members would help. As for future work a useful contribution would consist of designing a tool capable of providing meaningful graphical representations of the community structure of a complex network. Several tools have been already devised, capable of efficiently represent complex networks for visual analysis purposes, and the extension of their features to represent the community structure of complex networks would be worthy of further efforts.

Another interesting research field that is acquiring relevant attention during last years regards the overlapping community structure detection. In fact, in a social network may exist nodes which belong to multiple, different communities. In our opinion, this aspect is worthy of further investigations, in particular in the

context of Facebook. A possible approach could be, for example, by comparing sets, representing the community structure of the network, obtained by using both standard algorithms for community detection and overlapping community detection algorithms.

Acknowledgements

We would like to thank the Editor and the anonymous Referees whose comments helped us to greatly improve the quality of the work.

References

Blondel, V., Guillaume, J., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, P10008.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D., 2006. Complex networks: Structure and dynamics. Physics Reports 424 (4-5), 175–308.

Borgatti, S., Everett, M., 1999. Models of core/periphery structures. Social Networks 21, 375–395.

Catanese, S., De Meo, P., Ferrara, E., Fiumara, G., Provetti, A., 2011. Crawling facebook for social network analysis purposes. In: Proceedings of the International Conference on Web Intelligence, Mining And Semantics. pp. 52:1–52:8.

Chau, D., Pandit, S., Wang, S., Faloutsos, C., 2007. Parallel crawling for online social networks. In: Proceedings of the 16th International Conference on World Wide Web. ACM, pp. 1283–1284.

Clauset, A., 2005. Finding local community structure in networks. Physical Review E 72 (2), 026132.

Clauset, A., Newman, M., Moore, C., 2004. Finding community structure in very large networks. Physical Review E 70 (6), 066111.

Du, N., Wu, B., Pei, X., Wang, B., Xu, L., 2007. Community detection in large-scale social networks. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, pp. 16–25.

Duch, J., Arenas, A., 2005. Community detection in complex networks using extremal optimization. Physical Review E 72 (2), 027104.

Everett, M., Borgatti, S., 1999. Peripheries of cohesive subsets. Social Networks 21, 397–407.

Fortunato, S., 2010. Community detection in graphs. Physics Reports 486 (3-5), 75–174.

Fortunato, S., Barthélemy, M., 2007. Resolution limit in community detection. Proceedings of the National Academy of Sciences 104 (1), 36.

Girvan, M., Newman, M., 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99 (12), 7821.

Gjoka, M., Kurant, M., Butts, C., Markopoulou, A., 2010. Walking in facebook: a case study of unbiased sampling of osns. In: Proceedings of the 29th Conference on Information Communications. IEEE, pp. 2498–2506.

Gregory, S., 2007. An algorithm to find overlapping community structure in networks. Knowledge Discovery in Databases: PKDD 2007, 91–102.

Hagen, L., Kahng, A., 2002. New spectral methods for ratio cut partitioning and clustering. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 11 (9), 1074–1085.

Hampton, K., Sessions, L., Her, E., Rainie, L., 2007. Social isolation and new technology. PEW Research Center 4.

Hunter, D., Goodreau, S., Handcock, M., 2008. Goodness of fit of social network models. Journal of American Statistics Association 103 (481), 248–258.

Jin, D., Liu, D., Yang, B., Liu, J., 2009. Fast Complex Network Clustering Algorithm Using Agents. In: Proceedings of the 8th International Conference on Dependable, Autonomic and Secure Computing. pp. 615–619.

Karrer, B., Levina, E., Newman, M., 2008. Robustness of community structure in networks. Physical Review E 77 (4), 046119.

Kurant, M., Markopoulou, A., Thiran, P., 2010. On the bias of BFS (Breadth First Search). In: Proceedings of the 22nd International Teletraffic Congress. IEEE, pp. 1–8.

Lancichinetti, A., Fortunato, S., Kertész, J., 2009. Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics 11, 033015.

Lee, C., Reid, F., McDaid, A., Hurley, N., 2010. Detecting highly overlapping community structure by greedy clique expansion. In: Proceedings of the 4th Workshop on Social Network Mining and Analysis. ACM.

Leskovec, J., Lang, K., Dasgupta, A., Mahoney, M., 2009. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Mathematics 6 (1), 29–123.

McDaid, A., Hurley, N., 2010. Detecting highly overlapping communities with model-based overlapping seed expansion. In: 2010 International Conference on Advances in Social Networks Analysis and Mining. IEEE, pp. 112–119.

Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, B., 2007. Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. ACM, pp. 29–42.

Newman, M., 2006a. Finding community structure in networks using the eigenvectors of matrices. Physical Review E 74 (3), 036104.

Newman, M., 2006b. Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103 (23), 8577.

Newman, M., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical review E 69 (2), 026113.

Newman, M., Leicht, E., 2007. Mixture models and exploratory analysis in networks. Proceedings of the National Academy of Sciences 104 (23), 9564.

Ng, A., Jordan, M., Weiss, Y., 2001. On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems 14: Proceeding of the 2001 Conference. pp. 849–856.

Palla, G., Derényi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435 (7043), 814–818.

Porter, M., Onnela, J., Mucha, P., 2009. Communities in networks. Notices of the American Mathematical Society 56 (9), 1082–1097.

Raghavan, U., Albert, R., Kumara, S., 2007. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E 76 (3), 036106.

Shah, D., Zaman, T., 2010. Community detection in networks: The leader-follower algorithm. In: Proceedings of the Workshop on Networks Across Disciplines: Theory and Applications. pp. 1–8.

Traud, A., Kelsic, E., Mucha, P., Porter, M., 2011. Comparing Community Structure to Characteristics in Online Collegiate Social Networks. SIAM Review, 1–17.

Wei, Y., Cheng, C., 1989. Towards efficient hierarchical designs by ratio cut partitioning. In: Proceedings of the IEEE International Conference on Computer-Aided Design. pp. 298–301.

Zhao, Y., Levina, E., Zhu, J., 2011. Community extraction for social networks. In: Proceedings of the 2011 Joint Statistical Meetings.