Topological Features of Online Social Networks

Emilio Ferrara[1], Giacomo Fiumara[2]

[1]*Department of Mathematics, University of Messina, Italy*
eferrara@unime.it

[2] *Department of Physics, Informatics Section, University of Messina, Italy*
gfiumara@unime.it

Communicated by Associated editor

Abstract

The importance of modeling and analyzing Social Networks is a consequence of the success of Online Social Networks during last years. Several models of networks have been proposed, reflecting the different characteristics of Social Networks. Some of them fit better to model specific phenomena, such as the growth and the evolution of the Social Networks; others are more appropriate to capture the topological characteristics of the networks. Because these networks show unique and different properties and features, in this work we describe and exploit several models in order to capture the structure of popular Online Social Networks, such as Arxiv, Facebook, Wikipedia and YouTube. Our experimentation aims at verifying the structural characteristics of these networks, in order to understand what model better depicts their structure, and to analyze the inner *community structure*, to illustrate how members of these Online Social Networks interact and group together into smaller communities.

1. Introduction

From a scientific standpoint, the emergence of the phenomenon of Online Social Networks (OSNs), which is obtaining an unpredicted success during last years, contributed to keep high the level of attention of the research community on Social Networks.

The large dimension of these networks is a key aspect. In fact, they represent an incredible source of information on a large-scale, thus allowing studies which were impossible before by adopting classic sociological investigation techniques. Studying large-scale Online Social Networks, and their evolution, can be useful to investigate similarities and differences with real-life Social Networks. Moreover, this could help to empirically confirm,

or confute, sociological theories such as the "Small World" [1], the "Dunbar number" [2], and so on. Another advantage of studying OSNs is that connections and hierarchies among users are clearly defined, because they reflect constraints imposed by the platforms. This simplifies the process of induction of a graph of relations which reflects the structure of the Social Network. Moreover, it is possible to evaluate structural properties of large-scale networks adopting optimized algorithms and exploiting computational resources, in the light of the paradigm called Social Network Analysis (SNA), a novel branch of Computational Social Sciences.

In this work we face the problem of analyzing the topological features of some popular Online Social Networks, focusing on the characteristics of the graphs which represent these networks. To do so, we adopt some specific measures, such as the diameter and the degree distribution. Moreover, we investigate the emerging *community structure* inside these networks.

This paper is organized as follows: Section 2 focuses on related work, presenting classic literature on Social Networks, their analysis – in particular regarding Online Social Networks – and the latest works, which define directions of SNA. Section 3 introduces the key features reflected by Online Social Networks. In Section 4 we describe the generative models proposed to represent Social Networks, putting into evidence those aspects that could fit well to represent Online Social Networks and those in which they could fail. Results of our experimentation, presented in Section 5, depict the topological features of analyzed Online Social Networks. Section 6 concludes, suggesting some directions for future works.

2. Related Work

2.1. *Background in Social Networks and Models*

Literature about Social Network models is rooted in social sciences: in the sixties, Milgram and Travers [1] analyzed characteristics of real-life Social Networks, conducting several social experiments, and in conclusion, proposing the well known "Small World" model (see Section 3.1).

Another important concept, introduced by Zachary [3], is the *community structure* (see Section 3.3). He analyzed a small real-life social community (i.e., the components of a karate club), defining a model which describes the clusterization of Social Networks via cuts and fissions in sub-groups.

One of the first models has been provided by Erdős and Rényi [4] (see Section 4.1), and employs *random graphs* in order to reproduce real networks. Watts and Strogatz [5] furnished a one-parameter model that interpolates between an ordered finite dimensional lattice and a random graph (see Section 4.2). This because they empirically found that real-world Social

Networks are well connected and have a short average path length like random graphs, but they also have exceptionally large clustering coefficients, a feature which is not reflected by random graph models. Barabási and Albert [6,7] introduced different models that can be applied to friendship networks, the World Wide Web, business and commerce networks, etc., proving that they all share similar properties (see Section 4.3).

2.2. *Recent Studies and Current Trends*

Some of the current trends in the analysis of Social Networks include:

a) The investigation of the topological features of Social Networks by means of measurements, studying link symmetries, degree distributions, clustering phenomena, groups formations, etc., usually on a large-scale, by analyzing Online Social Networks [8,9].

b) The study of evolutionary aspects of Social Networks. In this context, Kumar et al. [10] defined a generative model which describes the structure of OSNs and their dynamics over the time. This model has been compared against actual data, in order to validate its reliability. Similarly, Leskovec [11], analyzed evolutionary aspects of Social Networks, trying to describe dynamic and structural features which influence the growth of communities, in particular when considering large Social Networks.

c) The definition of novel graph mining techniques, to face the computational complexity of studying large-scale social graphs with millions nodes and edges. Some authors [12,13] faced the problem of sampling from large graphs adopting different techniques, in order to establish if it is possible to avoid bias of data studying sub-graphs of Social Networks. They found that *Random Walks* and *Metropolis-Hastings* algorithms perform better, respectively, for static and dynamic graphs, concluding that samples of size of 15% of a social graph preserve the most of the properties.

d) The identification of those characteristics of the network that could suggest what nodes are more likely to be connected by trusted relationships (the *link prediction* problem) [14,15]. This is of a great interest for different commercial reasons.

2.2.1. *Applications of Social Network Analysis research*

Possible applications of information acquired from Social Networks have been investigated by Staab et al. [16]: methodologies for exploiting discovered data were defined, for marketing purposes, recommendation and trust analysis, etc. Recently, several marketing and commercial studies have been applied to OSNs, in particular to discover efficient channels to distribute information [17] and users who share similar tastes and preferences in order

to suggest them useful recommendations [18]. Our study provides useful information in all these directions, identifying interesting characteristics of Online Social Networks, considering the topological features that could affect how much efficiently nodes and edges could carry information through the networks.

2.3. *Motivations of our Study*

Several motivations to study features of Online Social Networks hold. On the one hand, results reported in our study have immediate implications, for example in commercial or marketing context, in the engineering of networks, etc. On the other hand, our achievements enrich the actual panorama of experimental verification of models and theories proposed by Social Sciences about the features and the dynamics of social networks.

3. Features of Social Networks

In this section we put into evidence three key features that characterize Social Networks, i.e., i) the "Small World" effect, ii) *scale-free degree distributions* and, iii) emergence of a *community structure*. During our experimentation, we take into account these features in order to establish if Online Social Networks show these well-known characteristics.

A Social Network can be defined by means of a graph $G = (V, E)$ whose set of vertices V represents nodes of the network (i.e., the individuals), and whose set of edges E represents connections (i.e., the social ties) among nodes of the network.

Social Networks can be modeled by means of weighted/unweighted and directed/undirected graphs, depending on their characteristics. In the following, we will conveniently adopt those structures that best fit to solve the given problems. Several features that we are going to consider are related to some background key concepts regarding graph theory, such as shortest paths or diameter of a graph, that are considered as well-known and not introduced in the following for reason of briefness.

3.1. *The "Small World" Effect*

The study of the "Small World" effect on Social Networks is rooted in Social Sciences [1]. Authors put into evidence that, despite their big dimension, Social Networks usually show a common feature: there exists a relatively short path connecting any pair of nodes within the network.

In fact, a "Small World" network is represented by a graph in which most nodes are not reciprocal neighbors each other, but could be reached from each other node by a small number of hops. The diameter ℓ, that

reflects the "Small World" property, scales proportionally to the logarithm of the dimension of the network, which is formalized as

$$\ell \propto Log(|V|) \tag{1}$$

where $|V|$ means the cardinality of V. Some characteristics of many real-world networks are well-modeled by means of "Small World" networks, such as OSNs [5], Internet [19], World Wide Web [6], biological networks [20].

3.2. *Scale-free Degree Distributions*

An important feature that is reflected by several generative models of Social Networks is the degree distribution of nodes. This feature characterizes the way the nodes are interconnected each other in the Social Network. On the one hand, in a random graph (see further) the node degree (i.e., the number of edges the given node is an endpoint of) 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 from a Poisson distribution. In particular, for several large-scale networks, such as the World Wide Web [6], Internet [21], metabolic networks [22], etc., the degree distribution follows a power law

$$P(k) \sim k^{-\gamma} \tag{2}$$

This power law distribution sensibly differs from Poisson distributions, but, on the other hand, falls off more gradually than an exponential one, allowing for a few nodes of very large degree to exist. Power law based models (see Section 4.3) apparently well depict the node degree distributions of large-scale Social Networks. 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* [23].

3.3. *Emergence of a Community Structure*

Another aspect to take into account when studying Social Networks is the emergence of a *community structure*: the more this structural characteristic is evident, the higher a network tends to divide into groups of nodes

whose connections are denser among entities belonging to the given group and sparser otherwise. Not all the network models are able to represent this characteristic. For example, the Erdős-Rényi (see Section 4.1) or the Barabási-Albert models (see Section 4.3) can not meaningfully represent the concept of *community structure*, that emerges from the empirical analysis of Social Networks. The *community structure* in Online Social Networks is discussed in Section 5.2.3.

4. Models of Social Networks

In literature, different models have been presented based on the previously discussed features. In this work we focus on the three most widely exploited modeling paradigms: 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 their ease of adoption. After the discovery of the "Small World" effect, a new class of models, namely "Small World" networks, has been introduced. Similarly, the power law degree distribution emerging from real-world Social Networks led to the modeling of the homonym networks, which are adopted to describe scale-free behaviors and other non-Poisson degree distributions.
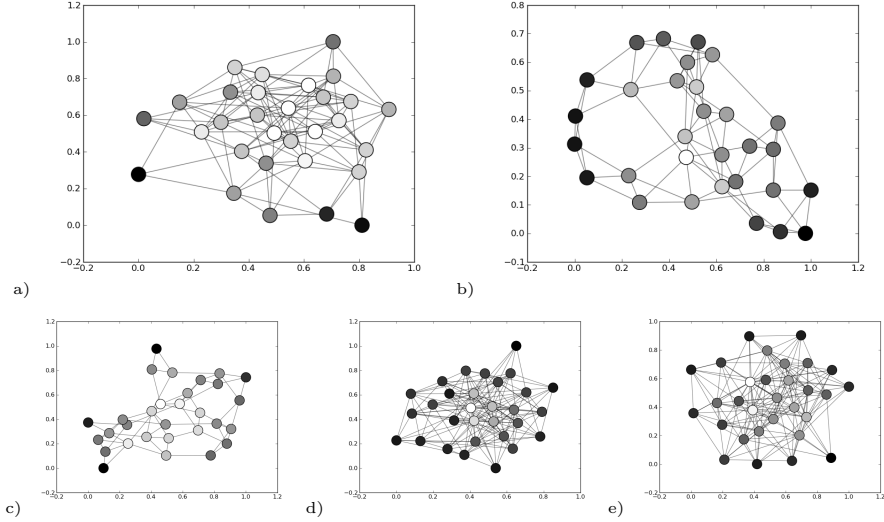

Fig. 1. Generative models of Social Networks: a) Erdős-Rényi [4]; b) Newman-Watts-Strogatz [24]; c) Watts-Strogatz [5]; d) Barabási-Albert [23]; e) Holme-Kim [25]. Lighter nodes have higher closeness centrality [26], darker nodes have lower closeness centrality. Nodes are placed according to a spring layout (Fruchterman-Reingold algorithm [27]).

4.1. *The Erdős-Rényi Model*

Erdős and Rényi [4] proposed one of the first modeling paradigm for networks, 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 (for example, random graphs have small diameters), 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.

This emerges by considering Figure 1.a, where is shown a Erdős-Rényi graph generated by adopting $n = 30$ and $p = 0.25$. The most of the nodes have similar closeness centrality (that is positively correlated to the degree), identified by the gray color in a gray-scale, and this means that all the nodes have relatively similar features (which is consistent with the formulation of the graph model, according with the Poisson distribution of node degrees). Social Networks exhibit a rather different behavior, making this model unfeasible for modern studies although it has been widely adopted in the past.

4.2. *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 coefficients, a characteristic that is not reflected by the Erdős-Rényi model or by other random graph models. Watts and Strogatz [5] 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, ranging from 0 (regular network) to 1 (random network). Focusing on two quantities, namely the characteristic path length $L(p)$ (defined as the number of edges in the shortest path between two vertices) and the clustering coefficient $C(p)$, some authors [28] found that $L \sim n/2k \geq 1$ and $C \sim 3/4$ as $p \to 0$, while $L = L_{random} ln(n)/ln(k)$ and $C = C_{random} k/n \leq 1$ as $p \to 1$. The Watts-Strogatz model is therefore suitable for explaining such properties in many real-world examples.

This model has been widely studied since the details have been published. 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.

A strong structural difference is evident between Watts-Strogatz [5], its variant Newman-Watts-Strogatz [24] presented in Figures 1.b and 1.c if compared with the Erdős-Rényi graph (Figure 1.a). First of all, it emerges that centrality of nodes is more heterogeneous, covering the whole grayscale. On the other hand, is evident, if compared with the other models, that it can not well reflect the power law distribution of node degree experimentally shown by data, even if a *community structure* is well represented (see Section 5.2.3).

4.3. *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 [7,23] 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. 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., the 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 $\gamma = 3$. 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 $\ell \sim ln(|V|)/lnln(|V|)$. Thus the model has much shorter ℓ with respect to 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 the size of the network, which is not true for real-world Social Networks.

Figures 1.d and 1.e propose two example of graphs generated by using the Barabási-Albert scale-free model [23] and a variant by Holme and Kim [25]. It is evident that the structure of these networks is much more compact than the Watts-Strogatz models (Figures 1.b and 1.c) but there are a few nodes that have a very high centrality while the most of the others have very low degrees (those depicted in dark gray). Due to the spring layout given by the Fruchterman-Reingold algorithm [27], those nodes with low degrees (belonging to the tail of the power law) are represented in peripheral positions in respect to central nodes. On the other hand, this model fails in representing a meaningful *community structure* of the network, differently to the Watts-Strogatz based models (see further).

4.4. *Discovering Communities*

Several studies have been conducted in order to investigate the *community structure* of real and Online Social Networks [29–31].

In its general formulation, the problem of finding communities in a network is intended as a clustering problem, thus solvable by 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 [32] 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 using the spectral clustering [32], exploiting the eigenvectors of the Laplacian matrix of the network. This process can be performed using the concept of ratio cut [33], 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 the number of communities present in the network and their size need to be defined in advance. This makes it unsuitable if the purpose is to discover the underlying community structure of an unknown network.

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 value of network modularity is

$$(3) \qquad Q = \sum_{s=1}^{m} \left[\frac{l_s}{|E|} - \left(\frac{d_s}{2|E|} \right)^2 \right]$$

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 implies 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 [20,34,35]. It first calculates the *edge betweenness* $B(e)$ of a given edge e in $G = (V, E)$, defined as

$$(4) \qquad B(e) = \sum_{n_i \in V} \sum_{n_l \in V} \frac{np_e(n_i, n_l)}{np(n_i, n_l)}$$

where n_i and n_l are vertices in V, $np(n_i, n_l)$ 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. It is possible to maximize the value of Q deleting edges with a high value of betweenness, because connecting vertices belonging to different communities. Starting from this intuition, first the algorithm ranks all the edges with respect to their betweenness, thus removing the most influential, calculates the value of Q and iterates the process until a significant increase of Q is obtained. At each iteration, each connected component of V identifies a community. Its cost is $O(n^3)$, being n the number of vertices in V; intuitively, it is unsuitable for large-scale networks.

Several variants and improvements exist; in this work, in order to discover the *community structure* of analyzed Online Social Networks, we the *Louvain method* [36]. At the best of our knowledge, does not exist a model of Social Network based on the modularity optimization, and this aspect leaves space for further investigations.

5. Results of the Experimental Evaluation

5.1. *Description of Adopted Online Social Network Datasets*

Our experimentation has been conducted on different Online Social Networks whose dataset are available online.
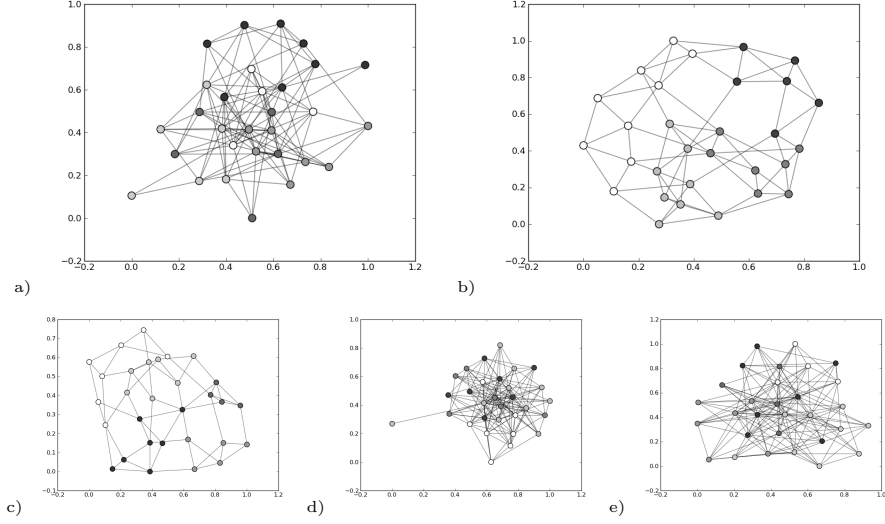
Fig. 2. *Community structure* of models of Social Networks: a) Erdős-Rényi [4]; b) Newman-Watts-Strogatz [24]; c) Watts-Strogatz [5]; d) Barabási-Albert [23]; e) Holme-Kim [25]. Nodes of the same color belong to the same community. Nodes are placed according with the spring layout given by the Fruchterman-Reingold algorithm [27].

Datasets $1 - 5$ are taken from Arxiv[a] datasets, as of April 2003, of papers in the field of, respectively: 1) "Astro Physics", 2) "Condensed Matter Physics", 3) "General Relativity and Quantum Cosmology"; 4) "High Energy Physics - Phenomenology", and 5) "High Energy Physics - Theory". Dataset 6 represents a network of scientific citations among papers belonging to the Arxiv "High Energy Physics - Theory" field. Dataset 7 illustrates the email communications among the Federal Energy Regulatory Commission members. Dataset 8 describes a sample of the Facebook friendship network, representing its social graph. Dataset 9 depicts the social graph of YouTube as of 2007. Finally, dataset 10 depicts the voting system of Wikipedia for the elections of administrators that occurred in January 2008. Adopted datasets have been summarized in Table 1.

5.2. *Topological Features*

Several measures are usually needed in order to study the topological features of Social Networks. To this purpose, for example, Carrington et al. [37] propose a list of some of them, including, amongst other, nodes/edges degree distributions, diameter, clustering coefficients, and more.

---

[a]Arxiv (http://arxiv.org/) is an Online archive for scientific preprints in the fields of Mathematics, Physics and Computer Science, amongst others.

Table 1. Datasets and results: $d(q)$ is the effective diameter, γ and σ, resp., the exponent of the power law node degree and community size distributions, Q the *network modularity*.

no.	Network	no. nodes	no. edges	Dir.	Type	$d(q)$	γ	σ	Q	Ref
1	CA-AstroPh	18,772	396,160	No	Collaborat.	5.3	2.23	1.50	0.628	[12]
2	CA-CondMat	23,133	186,932	No	Collaborat.	7.9	2.65	1.49	0.731	[12]
3	CA-GrQc	5,242	28,980	No	Collaborat.	8.9	2.12	1.48	0.861	[12]
4	CA-HepPh	12,008	237,010	No	Collaborat.	6.6	1.71	1.46	0.659	[12]
5	CA-HepTh	9,877	51,971	No	Collaborat.	8.4	2.63	1.46	0.768	[12]
6	Cit-HepTh	27,770	352,807	Yes	Citation	6.5	3.28	1.48	0.658	[12]
7	Email-Enron	36,692	377,662	Yes	Collaborat.	5.4	1.84	1.48	0.615	[12]
8	Facebook	63,731	1,545,684	Yes	Online Com.	6.8	2.91	1.48	0.634	[9]
9	Youtube	1,138,499	4,945,382	Yes	Online Com.	7.6	2.05	−	0.447	[9]
10	Wiki-Vote	7,115	103,689	Yes	Collaborat.	4.5	1.38	−	0.418	[12]

In this work, the following features have been investigated: i) node degree distribution; ii) diameter and hops; iii) *community structure*.

5.2.1. *Degree distribution*

The first interesting feature we analyzed is the degree distribution, which reflects in the topology of the network. The literature refers that Social Networks are usually described by power law degree distributions, $P(k) \sim k^{-\gamma}$, where k is the node degree and $\gamma \leq 3$.

The degree distribution can be represented by using some distribution functions: one of the most commonly used is the Complementary Cumulative Distribution Function (CCDF)

$$(5) \qquad \wp(k) = \int_k^\infty P(k')dk' \sim k^{-\alpha} \sim k^{-(\gamma-1)}.$$

In Figure 3 we show the degree distribution and the correspondingly CCDF evaluated on our Online Social Networks. For those networks that are directed, the out-degree is represented. All the graphics are represented by using a log-log scale, in order to put into evidence the scale-free behavior shown in these networks. In particular, for each of these distributions we estimated the value of γ in Equation 2. Values of γ are reported in Table 1.

Online Social Networks can be classified into two categories: i) networks that are properly described by a power law distribution; ii) networks that show some fluctuations with respect to those power law distributions that best fit to the real data. We discuss these two categories separately.

The networks that are well described by a power law distribution, such as those depicting datasets 7–10, are all Online Communities (i.e., networks of individuals connected by social ties such as a friendship relations) characterized by a fact: the most of the users are rather inactive and, thus,

they have a few connections with others members of the network. This phenomenon shapes a very consisting and long tail and a short head of the power law distribution of the degree (as graphically depicted by the respective plots in Figure 3).

The former category, instead, includes the co-authors networks (datasets 1–5) that are collaboration networks and a citation network (dataset 6). The plotting of these data against the power law distribution that try to best fit them show some fluctuation, in particular in the head of the distribution, in which, apparently, the behavior of the distribution is not properly described. The rationale behind this phenomenon lies into the intrinsic features of these networks, that are slightly different in respect to Online Communities.

For example, regarding the co-authors networks that represent collaborations among Physics scientists, the most of the papers are characterized by the following behavior, that can be inferred from the analysis of the real data: the number of co-authors tends to increase up to 3 (in average), then it slowly slopes down to a dozen, and finally it quickly decreases.

A similar interpretation holds even for the citation network, that is usually intended as a network in which there is a very small number of papers that gather the most of the citations, and a huge number of papers that have few or even no citations at all. This is a well-known phenomenon, called the "first-mover effect" [38].

Intuitively, from a modeling perspective, the only viable solution to capture these scale-free degree distribution behaviors would be by means of a Barabási-Albert *preferential attachment* model. On the one hand, by using this model it would be possible to reproduce the power law degree distribution of all the Online Social Networks depicted above. Similarly, even the "Small World" effect that describes networks with small diameters would be captured. On the other hand, this model would fail in depicting the *community structure* of those networks, whose existence has been put into evidence, both in this study (see Section 5.2.3) and in other works [20,39,40].

5.2.2. *Diameter and hops*

Most real-world Social Networks exhibit a relatively small diameter, but the diameter is susceptible to outliers. A more reliable measure of the pairwise distances between nodes in a graph is the *effective diameter*, i.e., the minimum number of hops in 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 networks, like Internet and the Web [6], real and Online Social Networks [7].

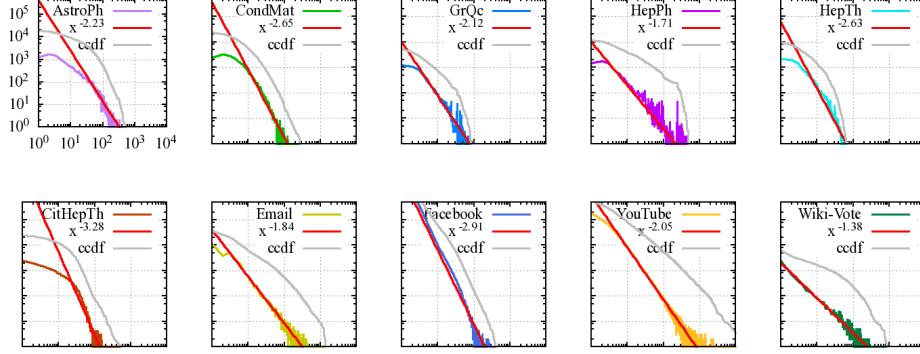A hop-plot extends the notion of diameter by plotting the number of

Fig. 3. Node degree distributions (*log–log* scale): for each network we plot the data, the best fitting power law function and *complementary cumulative distribution function* (all the plots use the same scale of the first one).

reachable pairs $g(h)$ within h hops, as a function of the number of hops h. It gives us a sense of how quickly neighborhoods of nodes expand with the number of hops.

In Figure 4 the number of hops necessary to connect any pair of nodes is plotted as a function of the number of pairs of nodes, for each given network. As a consequence of the *compact structure* of these networks (highlighted by the scale-free distributions and the "Small World" effect, discussed above), diameters show a fast convergence to asymptotic values listed in Table 1.

From a modeling standpoint, as for the degree distributions, the previous considerations hold true. Both the Watts-Strogatz and the Barabási-Albert models could efficiently depict the "Small World" feature of these Online Social Networks, and, most importantly, empirical data verify the so called "Six degrees of separation" theory, that is strictly related with the "Small World" formulation. In fact, it is evident that, regardless of the large scale of the networks analyzed, the effective diameters are really small (and, in average, close to 6), which is proved for real-world Social Networks [5,41].

5.2.3. *Community Structure*

From the perspective of the models representing the *community structure* of a network, we can infer the following information: from Figure 2.a, where the *community structure* of a Erdős-Rényi model is represented, the result appears random, according to the formulation of the model and its expected behavior when the calculated *network modularity Q* function (Equation 3) is analyzed. From Figures 2.b-c, at a first glance, it emerges that the *community structure* of Watts-Strogatz models is very regular and there is
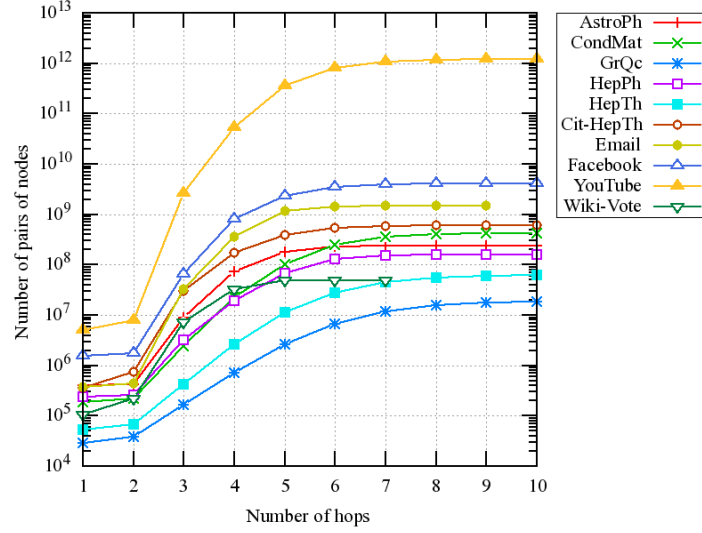
Fig. 4. Effective diameters (*log-normal* scale): the number of pairs of nodes reachable is plotted against the number of hops required, with $q = 0.9$.

a balance between communities with tighter connections and those with weaker connections. This reflects the formulation of the model but does not well depict the *community structure* represented by scale-free networks. Finally, Figures 2.d-e appear more compact and densely connected, features that are not reflected by experimental data.

Even if well-representing the "Small World" effect and the power law distribution of degrees, the Barabási-Albert model and its variants appear inefficient to represent the *community structure* of Online Social Networks.

From our experimental analysis on real datasets, by analyzing the obtained *community structures* by using the *Louvain method* [36], we focus on the study of the distribution of the dimensions of the obtained communities (i.e., the number of members constituting each detected community) [31].

Recently Fortunato and Barthelemy [42], put into evidence a resolution limit while adopting *network modularity* as maximization function for the community detection. In detail, authors found that modularity optimization may fail in the discovery of communities whose size is smaller than a given threshold. This value is strictly correlated to the characteristics of the given network. This resolution limit results in the creation of large communities incorporating an important part of the nodes of the network. In practice, in some particular cases it is possible that the clustering process produces a small number of communities of big sizes. This would possibly affect results in two ways: i) enlarging the tail of the power law distribution of the size of

15

the community, or ii) producing a not significant clustering of the network.

Because the clustering algorithm adopted (i.e., the *Louvain method*) is a modularity maximization technique, we investigated the effect of the resolution limit on our datasets. We found that in two cases (i.e., for the datasets 9–10) the clustering obtained was biased from the resolution limit and we excluded these networks from our analysis.

In the following we investigate the behavior of the distribution of the size of the communities in our networks.

In Figure 5 on the x-axis we plot the size of the community, and on the y-axis the probability P(x) of finding a community of the given size into the network. For each distribution we provide the best fitting power law function with a given exponent σ (that always ranges into the interval [1.4,1.5]) that well approximates the behavior of the community size.

In the figure the data are plotted as points and it is possible to highlight some communities whose size is larger than that expressed by the expected power law function (plotted as a red line), that constitute the heavy tail of the power law distribution. The results depicted show that, inside large Online Social Networks, there is a high probability of finding a high number of communities that contain few individuals and a relatively low probability of finding communities constituted by a large number of members. This confirms that individuals are more likely to aggregate in small communities, such as those representing family, friends, colleagues, etc., rather than in large communities [30,43].

Moreover, from Figure 5 we can put into evidence that large Online Communities, for example Facebook and the scientific collaboration networks, show a very tight *community structure* (a fact proved also by the high values of *network modularity*, reported as Q in Table 1). For example, regarding the collaboration networks, intuitively, we interpret this fact considering that scientists usually co-authoring different works, with different persons, work on papers signed only by a small amount of co-authors. It is very likely that these co-authors tend to group together (for example, if they co-authored several works) in the corresponding scientific communities.

On the other hand, for some networks such as the citation network and the email network, Figure 5 shows that it exists an important amount of communities constituted by a large amount of individuals, constituting the heavy long tail of the power law distribution. Also this aspect has an intuitive explanation. In fact, if we consider a network of scientific citations, there is a small amount of papers with a huge number of citations (which are very central in the topology of the network and, thus, are aggregated in same communities) and the most of the others that have very few citations, that forming small communities among each other (or single entities).
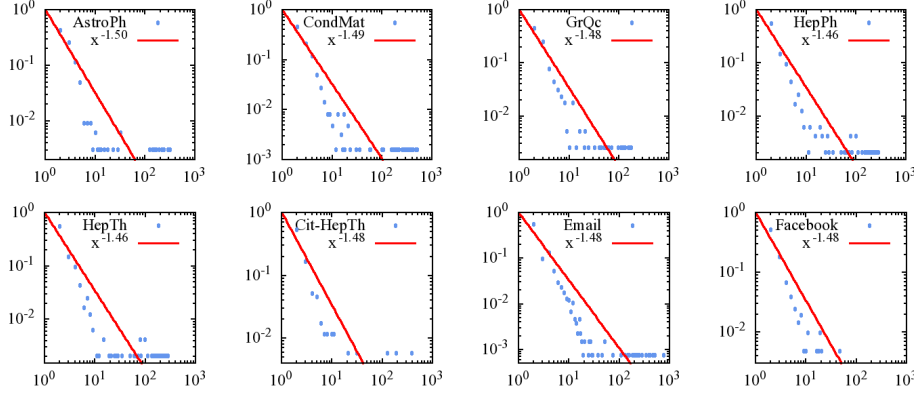
Fig. 5. *Community structure* analysis (*log–log* scale): the probability $P(x)$ of finding a community of a given size into the network is plotted against the size of the community. In red, the best fitting power law distribution functions are depicted.

6. Conclusions

In this paper we put into evidence those models which try to efficiently and faithfully represent the topological features of a family of complex networks, called Online Social Networks.

Several models have been presented in literature and we focused our attention on the three main exploited models, i.e., i) Erdős-Rényi random graphs, ii) Watt-Strogatz and, iii) Barabási-Albert preferential attachment. Each model, even if well describes some specific characteristics, fails in faithfully representing all the three main features we identified, that characterize Online Social Networks, namely i) "Small World" effect, ii) *scale-free degree distributions* and, finally, iii) emergence of a *community structure*.

We analyzed the topological features of several real-world Online Social Networks, fitting real data to models and putting into evidence what characteristic are preserved and what could not faithfully be represented by using these models.

As for future work, our main aim is to provide a generative model that, at the best of our knowledge, for first would try to faithfully represent the three principal features of Online Social Networks we have identified.

Acknowledgements

REFERENCES

1. J. Travers and S. Milgram, An experimental study small world problem, *Sociometry*, vol. 32, no. 4, pp. 425–443, 1969.

2. R. Dunbar, *Grooming, gossip, and the evolution of language.* Harvard University Press, 1998.

3. W. Zachary, An information flow model for conflict and fission in small groups, *J. of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

4. P. Erdős and A. Rényi, On random graphs, *Publicationes Mathematicae*, vol. 6, no. 26, pp. 290–297, 1959.

5. D. Watts and S. Strogatz, Collective dynamics of small-world networks, *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

6. R. Albert, H. Jeong, and A. Barabási, Diameter World Wide Web, *Nature*, vol. 401, no. 6749, p. 130, 1999.

7. R. Albert and A. Barabási, Statistical mechanics of complex networks, *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

8. Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, Analysis of topological characteristics of huge online social networking services, in *16th International Conference on the World Wide Web*, pp. 835–844, 2007.

9. A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, Measurement and analysis of online social networks, in *7th ACM conference on Internet measurement*, pp. 29–42, 2007.

10. R. Kumar, J. Novak, and A. Tomkins, Structure and evolution of online social networks, *Link Mining: Models, Algorithms, and Applications*, pp. 337–357, 2010.

11. J. Leskovec, J. Kleinberg, and C. Faloutsos, Graphs over time: densification laws, shrinking diameters and possible explanations, in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 177–187, ACM, 2005.

12. J. Leskovec and C. Faloutsos, Sampling from large graphs, in *12th International Conference on Knowledge Discovery and Data Mining*, pp. 631–636, 2006.

13. M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou, Walking in Facebook: a case study of unbiased sampling of OSNs, in *29th Conference on Information Communications*, pp. 2498–2506, 2010.

14. J. Golbeck and J. Hendler, Inferring binary trust relationships in web-based social networks, *Transactions on Internet Technology*, vol. 6,

no. 4, pp. 497–529, 2006.

15. D. Liben-Nowell and J. Kleinberg, The link-prediction problem for social networks, *J. American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

16. S. Staab, P. Domingos, P. Mike, J. Golbeck, L. Ding, T. Finin, A. Joshi, A. Nowak, and R. Vallacher, Social networks applied, *IEEE Intelligent Systems*, vol. 20, no. 1, pp. 80–93, 2005.

17. D. Kempe, J. Kleinberg, and É. Tardos, Maximizing the spread of influence through a social network, in *9TH ACM international conference on Knowledge discovery and data mining*, pp. 137–146, 2003.

18. P. De Meo, A. Nocera, G. Terracina, and D. Ursino, Recommendation of similar users, resources and social networks in a social internetworking scenario, *Information Sciences*, vol. 181, no. 7, pp. 1285–1305, 2011.

19. R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Dynamical and correlation properties internet, *Physical Review Letters*, vol. 87, no. 25, p. 258701, 2001.

20. M. Girvan and M. Newman, Community structure in social and biological networks, *PNAS*, vol. 99, no. 12, p. 7821, 2002.

21. M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationships internet topology, in *ACM SIGCOMM Computer Communication Review*, vol. 29, pp. 251–262, 1999.

22. H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A. Barabási, The large-scale organization of metabolic networks, *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

23. A. Barabási and R. Albert, Emergence of scaling in random networks, *Science*, vol. 286, no. 5439, p. 509, 1999.

24. M. Newman and D. Watts, Renormalization group analysis small-world network model, *Physics Letters A*, vol. 263, no. 4-6, pp. 341–346, 1999.

25. P. Holme and B. Kim, Growing scale-free networks with tunable clustering, *Physical Review E*, vol. 65, no. 2, p. 026107, 2002.

26. G. Sabidussi, The centrality index of a graph, *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.

27. T. Fruchterman and E. Reingold, Graph drawing by force-directed placement, *Software: Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

28. P. Holland and S. Leinhardt, Transitivity in structural models of small groups., *Comparative Group Studies*, 1971.

29. B. Karrer, E. Levina, and M. Newman, Robustness of community structure in networks, *Phy. Rev. E*, vol. 77, no. 4, p. 46119, 2008.

30. M. Porter, J. Onnela, and P. Mucha, Communities in networks, *Notices American Mathematical Society*, vol. 56, no. 9, pp. 1082–1097, 2009.

31. S. Fortunato, Community detection in graphs, *Physics Reports*, vol. 486, pp. 75–174, 2010.

32. A. Ng, M. Jordan, and Y. Weiss, On Spectral Clustering: Analysis and an algorithm, in *Advances in Neural Information Processing Systems 14*, 2001.

33. L. Hagen and A. Kahng, New spectral methods for ratio cut partitioning and clustering, *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 11, no. 9, pp. 1074–1085, 2002.

34. M. Newman and M. Girvan, Finding and evaluating community structure in networks, *Physical review E*, vol. 69, no. 2, p. 26113, 2004.

35. M. Newman, Modularity and community structure in networks, *PNAS*, vol. 103, no. 23, p. 8577, 2006.

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

37. P. Carrington, J. Scott, and S. Wasserman, *Models and methods in social network analysis*. Cambridge Univ Pr, 2005.

38. M. Newman, The first-mover advantage in scientific publication, *Europhysics Letters*, vol. 86, p. 68001, 2009.

39. M. Newman, The structure of scientific collaboration networks, *PNAS*, vol. 98, no. 2, p. 404, 2001.

40. J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney, Statistical properties of community structure in large social and information networks, in *17th international conference on World Wide Web*, pp. 695–704, 2008.

41. A. Barrat and M. Weigt, On the properties of small-world network models, *The European Physical J. B-Condensed Matter and Complex Systems*, vol. 13, no. 3, pp. 547–560, 2000.

42. S. Fortunato and M. Barthélemy, Resolution limit in community detection, *PNAS*, vol. 104, no. 1, p. 36, 2007.

43. E. Ferrara, Community structure discovery in facebook, *International Journal of Social Network Mining (In press)*, 2011.