International Conference on Mathematical and Computational Models
PSG College of Technology, Coimbatore

A SURVEY OF PROPERTIES AND MODELS OF ON-LINE SOCIAL NETWORKS

Anthony Bonato
Department of Mathematics
Ryerson University
Toronto ON
CANADA
abonato@ryerson.ca

Abstract

A number of recent studies have focused on the structure, function, and evolution of on-line social networks (OSNs) such as Facebook and Twitter. While OSNs have some characteristics in common with well-known complex networks like the web graph (such as the small world property and power law degree distributions), they exhibit important differences. For instance, OSNs exhibit a rich community structure, decreasing distances over time, and possess bad spectral expansion. Few models have been presented specifically for OSNs, and fewer still have been rigorously analyzed. We survey properties of OSNs, and summarize a recent model based on local knowledge and transitivity in social networks.

Key Words: *on-line social networks, graph models, small world property, densification power law, ILT model*

1. INTRODUCTION

Social networking sites such as Facebook, LinkedIn, Flickr, MySpace, and Twitter currently enjoy immense popularity, with an estimated half of all internet users registered on them. Recent events such as the earthquake in L'Aquila, Italy, and the political demonstrations in Iran highlight the emerging role of these networks as a vehicle for communication and the rapid dissemination of information, especially in the absence of traditional media. The study of such so-called *on-line social networks* (*OSNs*) lies in the intersection of graph theory, computer science, and social science. We may model an OSN by a graph with nodes representing users and edges corresponding to friendship links. Figure 1 shows the subgraph induced by friends of an anonymous user on Facebook (that is the set of all friends of the user and links between them), generated using the *Nexus* application.

The author acknowledges support from NSERC and MITACS grants.

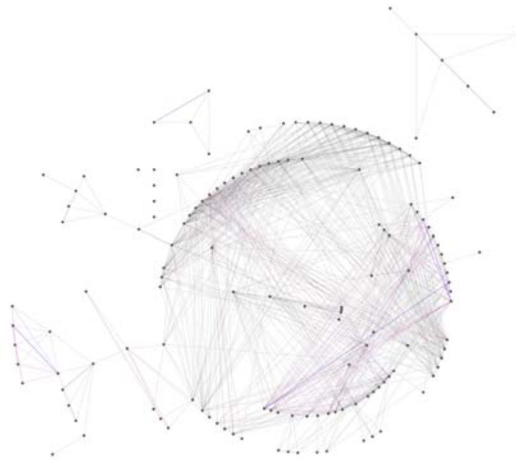

Figure 2. The subgraph induced by the friends of a Facebook user.

While OSNs gain increasing popularity among the general public, there is a parallel increase in interest in the cataloguing and modelling of their structure, function, and evolution. OSNs supply a vast and historically unprecedented record of large-scale human social interactions over time. OSNs network data is more readily accessible and measurable than in off-line social networks, and is large-scale (with often many millions of users and friendship links in a given networks). The availability of large-scale social network data has led to the observation of emergent topological properties of OSNs, and so presents challenges in the design and rigorous analysis of models simulating these properties. Graph models were successful in simulating properties of other *complex networks* like the web graph (see the books [5,8] for surveys of models of the web graph and complex networks), and it is thus, natural to propose models for OSNs. As the subject is relatively new, few models for OSNs have been posed, and there is no universal consensus of which properties such models should simulate. Further, we should expect OSNs to possess a different structure when compared other complex networks. For example, the link structure of the web graph is more organized on content, while the link structure of OSNs is organized on users.

Social network analysis has a long history, and possesses a corpus of statistical techniques (see [28], for example). We note that in off-line social networks it is more difficult to collect data, and doing so is both expensive and error-prone. Researchers are now in the enviable position of observing how OSNs evolve over time, and as such, network analysis and models of OSNs typically incorporate time as a parameter.

As with modelling complex networks, the question arises of why model OSNs in the first place. Besides the mathematical challenges, graph models can help provide understanding of the underlying mechanisms of OSNs, and hence, can serve as a predictive tool. A better comprehension of the evolution of OSNs may lead to a better insight into community structure and social influence [12] in such networks, which in turn could be used for targeted advertising, to isolating terrorist cells.

One of the principle aims of the present article is to survey empirical properties of OSNs (see Section 2), and in addition, provide a master list of their key graph theoretical properties (see Section 3). Although this list is by no means exhaustive, it points to the main challenges in the modelling of OSNs. While no one model captures all the observed properties of OSNs, in Section 4 we survey a recent model (the Iterated Local Transitivity model) which simulates many of them. We finish with a brief discussion of future work on modelling OSNs.

2. EMPIRICAL PROPERTIES OF OSNs

There are now a number of studies on the mining of OSNs for various graph theoretical properties. We highlight findings from these studies in this section.

Large-scale. Basic statistics of networks include their *order* (that is, the number of nodes), *size* (that is, number of edges), and *degrees of nodes* (that is, the number of edges incident with a node; for directed graphs, in-degrees and out-degrees are defined analogously). OSNs are examples of massive or large-scale networks, with order and size in the millions; further, some users have disproportionately high degrees. According to the statistics on Facebook's website, there are over 250 million nodes in the Facebook network, with an average degree of 120 (giving a size of approximately 3^{10}). LinkedIn's website claims their OSN has at least 43 million nodes, with a new user joining every second. Each of the nodes of Twitter corresponding to celebrities Ashton Kutcher, Ellen Degeneres, Britney Spears, and the news agency CNN have degree over 2 million.

Small world property. The *small world property*, introduced by social scientists Watts and Strogatz [29], is a central notion in the study of complex networks, and has roots in the work of the Milgram [26] on short paths of friends connecting strangers in the United States. The *diameter* of a graph is a longest shortest path connecting two nodes (and ∞ if the graph is disconnected). Define the *average distance* of an undirected graph $G = (V(G),E(G))$ of order n as

$$L(G) = \frac{\frac{1}{2}\sum_{x,y\in V(G)} d(x,y)}{\binom{n}{2}},$$

where $d(x,y)$ represents the distance between nodes x and y. Let $N(x)$ be the neighbour set of a node x, and let $e(x,t)$ be the number of edges in the subgraph induced by $N(x)$ in G. For a node x in $V(G)$ with degree at least 2 define the *clustering coefficient* of x as

$$c(x) = \frac{e(x,t)}{\binom{\deg(x)}{2}}.$$

By convention, $c(x) = 0$ if the degree of x is at most 1. The *clustering coefficient* of G, written $C(G)$, is the average of the clustering coefficients over all nodes of G. The small world property demands a low average distance of $O(\log\log n)$ (or a diameter of $O(\log n)$), and a higher clustering coefficient than found in a random graph $G(n,p)$ with the same number of nodes and approximately same average degree.

Adamic et al. [1] provided an early study of an OSN at Stanford University, and found that the network has the small world property. Similar results were found in [2] which studied Cyworld, MySpace, and Orkut, and in [27] which examined data collected from four OSNs: Flickr, YouTube, LiveJournal, and Orkut. In the latter study, the average distances and clustering coefficients of the OSNs were found to be lower and higher, respectively, than those of the web graph. Low diameter (of 6) and high clustering coefficient were reported in the Twitter friendship graph by Java et al. [20] (see also [21]).

Power laws. Power laws indicate an undemocratic nature of networks: many nodes have low degree, but a small proportion of nodes have substantially higher degree. Given an undirected graph G of order n and a non-negative integer k, we define

$$N_{k,t} = |\{x \in V(G) : \deg_G(x) = k\}|.$$

The *degree distribution of* G is the sequence ($N_{k,n}$: $0 \leq k \leq \Delta(G)$), where $\Delta(G)$ is the maximum degree of G. We say that the degree distribution of G follows a *power law* (or G is a *power law graph*) if for each degree k (or a sufficient range of degrees),

$$\frac{N_{k,n}}{n} \sim k^{-\beta}$$

for a fixed real constant $\beta > 1$. In power law graphs, if we plot the log-log plot of the degree distribution (number of nodes of give degree versus degree), then we obtain an (approximate) straight line with slope $-\beta$. If G is directed graph, then we may define power laws for the in- and out-degree distributions in the obvious way. Power laws were observed over a decade ago in subgraphs sampled from the web graph, and are ubiquitous properties of complex networks (see Chapter 2 of [5]).

Kumar et al. [22] studied the evolution of the OSNs Flickr and Yahoo!360, and found that these networks exhibit power-law degree distributions. Golder et al. [19] analyzed the Facebook network by studying the messaging pattern between friends with a sample of 4.2 million users, and the degree distribution of the Facebook OSN was reported to follow a power law. Power law degree distributions for both the in- and out-degree distributions were documented in Flickr, YouTube, LiveJournal, and Orkut [27], as well as in Twitter [20]. In [27], the in- and out-degree distributions for each OSN studied were found to be similar (a property not shared, for example, by the web graph).

Densification power laws and shrinking distances. Recent work by Leskovec et al. [23] highlights two additional properties of complex networks. A graph G with e_t edges and n_t nodes satisfies a *densification power law* if there is a constant a in (1,2) such that e_t is proportional to n_t^a. In particular, the average degree grows to infinity with the order of the network (in contrast to say the preferential attachment model, which generates graphs with constant average degree). In [23], densification power laws were reported in several real-world networks such as the physics citation graph and the internet graph at the level of autonomous systems. Another striking property found in such networks (and also in OSNs; see [22]) is that distances in the networks (measured by either diameter or average distance) decreases with time. Standard models for complex networks such as preferential attachment or copying models have logarithmically or sublogarithmically growing diameters and average distances with time. Various models (such as the Forest Fire [23] and Kronecker multiplication [24]) were proposed simulating power law degree distribution, densification power laws, and decreasing distances. Note that both of these properties are dynamic, in as much as they require snapshots of the network over time.

Component structure. In [22], it was reported that in Flickr and Yahoo!360, users fall into one of three categories: singletons, the giant component, and the middle region. The *singletons* are simply isolated nodes in the network. The *giant component* represents a dense core of low (in fact, shrinking) diameter, and this subgraph contains nodes with high degree. The *middle region* is the remainder of the OSN, and consists of various isolated communities with a *star-like structure* (that is, high degree nodes joined to lower degree ones), who interact with one another but not with the overall network.

Bad spectral expansion. Social networks often organize into separate clusters in which the intra-cluster links are significantly higher than the number of inter-cluster links. In particular, social networks contain communities (characteristic of social organization), where tightly knit groups correspond to the clusters [17]. As a result, it is reported in [14] that social networks possess bad spectral expansion properties realized by small gaps between the first and second eigenvalues of their adjacency matrices.

3. MODELS FOR OSNs

Although modelling was conducted extensively for real-world complex networks such as the web graph (see [5]), models of OSNs have only recently been introduced, and very few have been rigorously analyzed. A model *simulates* a property if graphs generated by the model satisfy the property (with high probability if the model is stochastic). Based on the discussion from Section 2, models for OSNs should simulate the following properties.

1. *On-line*: the order and size of the graph changes over time.
2. Small world property.
3. Power law degree distributions.
4. Shrinking distances and densification power law.
5. Component structure.
6. Bad spectral expansion.

Complex network models generate graphs which typically satisfy properties 1-3 (see [5]). For example, the preferential attachment model (which is one of the earliest and best known models for complex networks; see [3,4]) simulates properties 1-3. Properties 5 and 6 are more specific to social networks. The Forest Fire and Kronecker multiplication models [23,24] simulate properties 1-4. A model was introduced in [22] which simulates property 5. No one model is known which simulates all these six properties.

We survey a deterministic model introduced in [6], called *Iterated Local Transitivity* (*ILT*), for OSNs and other complex networks which dynamically simulates several of their properties (specifically, all but properties 3 and 5). The central idea behind the ILT model is what sociologists call *transitivity*: if *u* is a friend of *v*, and *v* is a friend of *w*, then *u* is a friend of *w* (see, for example, [15,28,30]). In its simplest form, transitivity gives rise to the notion of *cloning*, where *u* is joined to all of the neighbours of *v*. In the ILT model, given some initial graph as a starting point, nodes are repeatedly added over time which clone each node, so that the new nodes form an independent set. Local knowledge is an important feature of social and complex networks. The ILT model uses only local knowledge in its evolution, in that a new node only joins to neighbours of an existing node.

We now give a precise formulation of the model. The ILT model generates finite, simple, undirected graphs (G_t: $t \geq 0$) over a countably infinite sequence of discrete time-steps. The only parameter of the model is the initial graph G_0, which is any fixed finite connected graph. Assume that for a fixed $t \geq 0$, the graph G_t has been constructed. To form G_{t+1}, for each node x in $V(G_t)$ add its *clone* x', such that x' is joined to x and all of its neighbours at time t. Note that the set of new nodes at time

$t+1$ form an independent set of cardinality $|V(G_t)|$. See Figure 2 for the graphs generated from the 4-cycle over the time-steps $t = 1, 2, 3$, and 4.

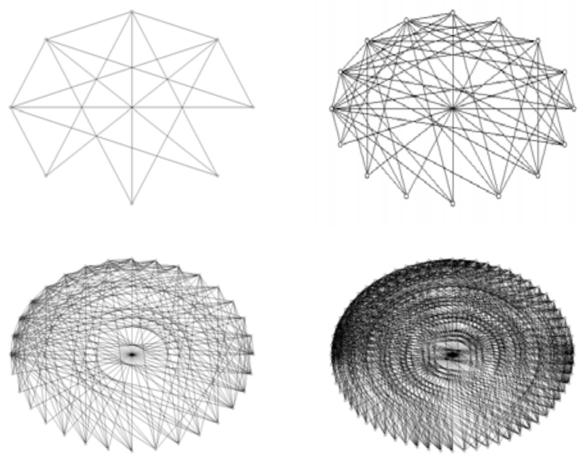

Figure 2. The evolution of the ILT model with $G_0 = C_4$ for $t = 1,2,3,4$.

We state our main results on the ILT model (with proofs found in [6]). We first demonstrate that the model exhibits a densification power law. We write $\deg_t(x)$ for the degree of a node at time t, n_t for the order of G_t, and e_t for its size. It is straightforward to see that $n_t = 2^t n_0$. Define the *volume* of G_t by

$$vol(G_t) = \sum_{x \in V(G_t)} \deg_t(x) = 2e_t.$$

Theorem 2.1. [6] For $t > 0$, the average degree of G_t equals

$$\left(\frac{3}{2}\right)^t \left(\frac{vol(G_0)}{n_0} + 2\right) - 2.$$

Observe that Theorem 2.1 supplies a densification power law with exponent

$$a = \frac{\log 3}{\log 2} \approx 1.58.$$

Nodes in graphs generated by the ILT model have (on average) short distances between them, even if the initial graph has large average distance.

Theorem 2.2. [6] For $t > 0$,

$$L(G_t) = \frac{4^t \left(\sum_{x,y \in G_0} d(x,y) + (e_0 + n_0)(1 - (3/4)^t) \right)}{4^t n_0^2 - 2^t n_0} = \Theta(1).$$

For many initial graphs G_0 (such as large cycles) Theorem 2.2 demonstrates that the average distance actually decreases over time, a property observed in OSNs and other complex networks (see [22,23]).

The clustering coefficient of the graph at time t generated by the ILT model is estimated and shown to tend to 0 slower than a $G(n,p)$ random graph with the same average degree.

Theorem 2.3. [6] For $t > 0$,

$$C(G_t) = n^{\log_2(7/8) + o(1)}.$$

Observe that $C(G_t)$ tends to 0 as t becomes large. In contrast, for a random graph $G(n,p)$ with comparable average degree $pn = \Theta\left(n^{\log_2(3/2)}\right)$, the clustering coefficient is $p = \Theta\left(n^{\log_2(3/4)}\right)$ which tends to zero much faster than $C(G_t)$.

The ILT model generates graphs with bad expansion properties as indicated by the spectral gap of both their normalized Laplacian and adjacency matrices. For regular graphs, the eigenvalues of the adjacency matrix are related to several important graph properties, such as in the expander mixing lemma. The normalized Laplacian matrix of a graph, introduced by Chung [7] relates to important graph properties even in the case where the underlying graph is not regular. Let A denote the adjacency matrix of G, and D denote the diagonal matrix with degrees of the vertices of G along the diagonal. Then the *normalized Laplacian* matrix of G is

$$L = I - D^{-1/2} A D^{1/2}.$$

Let $0 = \lambda_1 \le \cdots \le \lambda_{n-1} \le 2$ denote the eigenvalues of L. The *spectral gap* of the normalized Laplacian of G is

$$\lambda(G) = \max\{|\lambda_1 - 1|, |\lambda_{n-1} - 1|\}.$$

Spectral properties of random power law graphs were studied in [9,18]; in [9] it was found that for random power law graphs (with certain parameters) that

$$\lambda(G) \le (1 + o(1)) \frac{4}{\sqrt{d}},$$

where d is the expected average degree. For the graphs G_t generated by the ILT model, we observe that the spectra behaves quite differently and, in fact, the spectral gap has a constant order. The following theorem suggests a significant spectral difference between graphs generated by the ILT model and random graphs.

Theorem 2.4. [6] For $t > 0$,

$$\lambda(G_t) > \frac{1}{2}.$$

Theorem 2.4 represents a drastic departure from the good expansion found in $G(n,p)$ random graphs, where $\lambda(G) = o(1)$ [9]. We note that as in the Laplacian case, there is also a small spectral gap in the adjacency matrix of G_t (see [6]).

We close this section with an observation about the ILT model which points to a potential new direction of research in OSNs. In a graph G, a set S of nodes is a *dominating set* if every node not in S has a neighbour in S. The *domination number* of G, written $\gamma(G)$, is the minimum cardinality of a dominating set in G. In some sense, dominating sets act as "gateways" in the network: from them one can broadcast across the network in one time-step. In the ILT model, the domination number is a constant independent of time.

Theorem 2.5. [6] For all $t > 0$,

$$\gamma(G_t) = \gamma(G_0).$$

In terms of OSNs, Theorem 2.5 suggests that users in the network can spread information or influence with relative ease no matter how large the graph becomes. The theorem contrasts with the results of [11] on the domination number of the preferential attachment model, where it was shown that with probability tending to 1 as $t \to \infty$, a graph G_t generated by that model at time t satisfies

$$\gamma(G_t) = \Theta(|V(G_t)|).$$

We hope to verify that OSNs have small domination number. Computing the domination number in general graphs is NP-hard [16]. Hence, determining the precise domination number of a large-scale OSN is likely difficult, but upper bounds (using known heuristics) are more tractable.

4. FUTURE WORK

The modelling of OSNs is in its infancy, and more attention in the future should be on the dynamic nature of these networks (that is, the arrival and departure from the network, user interaction, and growth rates of the networks; see [31]). There is an emerging corpus of empirical OSN data as summarized by the six properties in Section 3. The ILT model described in Section 3 satisfies many (but not all, such as power law degree distributions) of these properties. Duplication models for biological networks [10] generate power law graphs and share some similarities in common with the ILT model. A randomized mixture of duplication and cloning may provide a more realistic model for the structure and function of OSNs. Potential candidates for more realistic models include randomizing the cloning process in the ILT model so that

i) only nodes with high degree are cloned; or
ii) edges incident with low degree nodes are deleted.

A randomized ILT model was presented in the full version of [6] with *tuneable* (that is, varies with choice of parameters) densification power law exponent, and satisfying properties of the deterministic model (such as decreasing distances and bad expansion). In this model, edges are placed randomly amongst new nodes, with the motivation that some new users to OSN may already be friends outside the network. When they join an OSN, they seek each other out and become friends there.

Some OSNs, such as Twitter, are better represented by directed graphs (with some users following others, resulting in a one-way communication between them). It would be useful to devise and analyze a natural model for OSNs which generates directed graphs.

In geometric graph models, vertices are identified with points in a metric space, and edges are introduced by a mixture of probabilistic rules and proximity in the space. For instance, we may consider the web graph as embedded in so-called *topic space*, where web pages with related content are closer together via a prescribed metric. The discovery of empirically observed properties often leads to new models; for example, the presence of sparse cuts in the web graph led to the consideration of geometric models (see Chapter 4 of [5] for an overview of such models). We may envision OSNs as embedded in a *social space* whose dimensions quantify user traits such as interests or geography; for instance, nodes representing users from the same country would be closer in social space. A promising direction would be to develop a geometric graph model for OSNs, and use it for community detection and clustering. A first step in this direction was given in [25], which introduced a rank-based geometric model for social networks.

4. REFERENCES

[1] L.A. Adamic, O. Buyukkokten, E. Adar, A social network caught in the web, *First Monday* **8** (2003).

[2] Y. Ahn, S. Han, H. Kwak, S. Moon, H. Jeong, Analysis of topological characteristics of huge on-line social networking services, In: *Proceedings of the 16th International Conference on World Wide Web*, 2007.

[3] A. Barabási, R.Albert, Emergence of scaling in random networks, *Science* **286** (1999) 509-512.

[4] B. Bollobás, O. Riordan, J. Spencer, G. Tusnády, The degree sequence of a scale-free random graph process, *Random Structures and Algorithms* **18** (2001) 279-290.

[5] A. Bonato, *A Course on the Web Graph*, American Mathematical Society Graduate Studies Series in Mathematics, Providence, Rhode Island, 2008.

[6] A. Bonato, N. Hadi, P. Horn, P. Pralat, C. Wang, Dynamic models of on-line social networks, In: *Proceedings of the 6th Workshop on Algorithms and Models for the Web Graph (WAW'09),* 2009.

[7] F.R.K. Chung, *Spectral Graph Theory*, American Mathematical Society, Providence, Rhode Island, 1997.

[8] F.R.K. Chung, L. Lu, *Complex Graphs and Networks*, American Mathematical Society, Providence, Rhode Island, 2006.

[9] F.R.K. Chung, L. Lu, V. Vu, The spectra of random graph with given expected degrees, *Internet Mathematics* **1** (2004) 257-275.

[10] F.R.K. Chung, G. Dewey, D.J. Galas, L. Lu, Duplication models for biological networks, *Journal of Computational Biology* **10** (2003) 677-688.

[11] C. Cooper, R. Klasing, M. Zito, Lower bounds and algorithms for dominating sets in web graphs, *Internet Mathematics* **2** (2005) 275-300.

[12] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, S. Suri, Feedback effects between similarity and social influence in on-line communities, In: *Proceedings of the 14th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2008.
[13] H. Ebel, J. Davidsen, S. Bornholdt, Dynamics of social networks, *Complexity* **8** (2003) 24-27.
[14] E. Estrada, Spectral scaling and good expansion properties in complex networks, *Europhys. Lett*. **73** (2006) 649-655.
[15] O. Frank, Transitivity in stochastic graphs and digraphs, *Journal of Mathematical Sociology* **7** (1980) 199-213.
[16] M.R. Garey, D.S. Johnson. Strong NP-completeness results: motivation, examples, and implications, *Journal of the ACM* **25** (1978) 499–508.
[17] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, *Proceedings of the National Academy of Sciences* **99** (2002) 7821-7826.
[18] C. Gkantsidis, M. Mihail, A. Saberi, Conductance and congestion in power-law graphs, In: *Proceedings of the 2003 ACM SIGMETRICS International Conference on Measurement Modeling of Computer Systems*, 2003.
[19] S. Golder, D. Wilkinson, B. Huberman, Rhythms of social interaction: messaging within a massive on-line network, In: *Proceedings of 3rd International Conference on Communities and Technologies*, 2007.
[20] A. Java, X. Song, T. Finin, B. Tseng, Why we twitter: understanding microblogging usage and communities, In: *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop*, 2007.
[21] B. Krishnamurthy, P. Gill, M. Arlitt, A few chirps about Twitter, In: *Proceedings of The First ACM SIGCOMM Workshop on Online Social Networks*, 2008.
[22] R. Kumar, J. Novak, A. Tomkins, Structure and evolution of on-line social networks, In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.
[23] J. Leskovec, J. Kleinberg, C. Faloutsos, Graphs over time: densification laws, shrinking diameters and possible explanations, In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.
[24] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, Realistic, mathematically tractable graph generation and evolution, using Kronecker multiplication, In: *Proceedings of European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2005.
[25] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, A. Tomkins, Geographic routing in social networks, *Proceedings of the National Academy of Sciences* **102** (2005) 11623-11628.
[26] S. Milgram, The small world problem, *Psychology Today* **2** (1967) 60-67.
[27] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, B. Bhattacharjee, Measurement and analysis of on-line social networks, In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, 2007.
[28] J.P. Scott, *Social Network Analysis: A Handbook*, Sage Publications Ltd, London, 2000.
[29] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* **393** (1998) 440-442.
[30] H. White, S. Harrison, R. Breiger, Social structure from multiple networks, I: Blockmodels of roles and positions, *American Journal of Sociology* **81** (1976) 730-780.
[31] W. Willinger, R. Rejaie, M. Torkjazi, M.Valafar, M. Maggioni, Research on online social networks: time to face the real challenges, In: *Proceedings of The 2nd Workshop on Hot Topics in Measurement and Modeling of Computer Systems*, 2009.