Similarity Distribution in Phonebook-Centric Social Networks

Péter Ekler
Department of Automation and Applied Informatics
Budapest University of Technology and Economics
Magyar Tudósok Körútja 2., 1113 Budapest, Hungary
peter.ekler@aut.bme.hu

Tamás Lukovszki
Faculty of Informatics
Eötvös Loránd University
Pázmány Péter sétány 1/C, 1117 Budapest, Hungary
lukovszki@inf.elte.hu

***Abstract*—In this work we define the term of phonebook centric-social networks, describe a graph model and study structural properties of this. The key difference between phonebook-centric social networks and usual social networks is allowing synchronization between the phonebook of the mobile phone and the network. By the synchronization the goal is to identify the persons listed in the phonebook and the network, it means to find similar entries and keep the data consistent. In that way the contacts of a member of a phonebook-centric social network becomes independent from the current phone. The number of similarities is crucial in phonebook-centric social networks from scalability point of view. We found that the distribution of similarities can be very well approximated by power law distribution. This means that few users involve a huge amount of similarities while the most of them only few ones.**

Keywords-component; social networks, mobile phone, syncronization, similarities, graph structure

I. INTRODUCTION

The popularity of social networks was noticeable in the last few years. Several social networks appeared and attracted thousands or even millions of users. The online social networks Facebook [11] and Myspace [15] are among the top ten visited websites on the Internet [4]. The basic idea behind these networks was that users can manage social relationships on these networks. A social network is basically a social structure consisting of nodes that generally correspond to individuals or organizations. These nodes are connected by different type of relations. Users of social networks are able to share personal detail about themselves, talk in forums, share photos or entire galleries, play games, etc. Basically it is an environment created by the people who are using it.

Mobile phones and mobile applications are another hot topic nowadays. Both hardware and software capabilities of mobile phones have been evolving in the last decades. Yet support of mobile devices is generally marginal in most social networks, it is limited to photo and video upload capabilities and access to the social network using the mobile web browser. However, if we consider the phonebook in our mobile phone, we realize that basically it is a small part of a social network because every contact in our phonebook has some kind of relationship to us. Given an implementation that allows us to upload as well as download our contacts to and from the social networking application, we can completely keep our contacts synchronized so that we can also see all of our contacts on the mobile phone as well as on the web interface. In the rest of this paper we refer to this solution as a *phonebook-centric social network*.

In case of a phonebook-centric social network, the phonebook of the mobile phone is automatically updated with the latest information provided by the friends of its owner. This means also that the persons in the phonebook of a user also get the latest information about her or him automatically, so there is no need to notify them one by one if the phone number changes for example. In addition to that, the private contacts are also uploaded to the phonebook-centric social network. These contacts are not visible to other members of the site. However, having all of the contacts in the system has the following benefits:

- The contacts can be managed (list, view, edit, call, etc.) from a browser.
- The service notices the user if duplicate contacts are detected in its phonebook and warns about it.
- The contacts are safely backed up in case the phone gets lost.
- The contacts can be easily transferred to a new phone if the user replaces the old one.
- The phonebook can be shared between multiple phones, if one happen to use more than one phone.
- It is not necessary to explicitly search for the friends in the service, because it notices if there are members similar to the contacts in the phonebooks and warns about it.

Phonebookmark is a *phonebook-centric social network* implementation by Nokia Siemens Networks. Section 3 summarizes the exact definition of phonebook-centric social network. Before public introduction it was available for a group of general users from April to December of 2008. It had 420 registered members with more than 72000 private contacts, which is a relatively ideal number for testing the handling of similarities. During this period we have collected different type of data related to the social network which was the base of the measurements and the proposed model in this paper.

The rest of the paper is organized as follows. Section 2 describes related work in the field of social networks, power law distributions appearing in such networks, peer-to-peer networks and the Internet and generative models leading to such distributions. Section 3 discusses the evolution of social networks and Section 4 introduces a phonebook-centric

social network implementation, called *Phonebookmark*. Section 5 shows measurements related to phonebook-centric social networks and calculate the distribution of similarities. Finally, Section 6 concludes the paper and proposes future research plans.

II. Related Work

Huge amount of papers and popular books, such as Barabási's Linked [5] study the structure and principles of dynamically evolving large scale networks like the Internet and networks of social interactions. Many features of social processes and the Internet are governed by power law distributions. Following the terminology in [12], a nonnegative random variable X is said to have a power law distribution if Pr $[X \geq x] \sim cx^{-\alpha}$, for constant $c>0$ and $\alpha>0$. In a power law distribution asymptotically the tails fall according to the power α, which leads to much heavier tails than other common models.

Distributions with an inverse polynomial tail have been first observed in 1897 by Pareto [17] (see. [14]), while describing the distribution of income in the population. In 1935 Zipf [20] and Yule [19] investigated the word frequencies in languages and based on empirical studies he stated that the frequency of the *n*-th frequent word is proportional to 1/*n*. Zipf observed similar statistical behavior in the distribution of inhabitants in cities [21].

In [8], the graph structure of the Web has been investigated and it was shown that the distribution of in- and out-degree of the web graph and the size of weekly and strongly connected components are well approximated by power law distributions. Nazir et al. [16] showed that the in- and out-degree distribution of the interaction graph of the studied Myspace applications also follow such distributions. Those distributions also approximate the degree distribution of the Gnutella network [17]. Crovella et al. [9] observed power law distributions in the sizes of files and transmission times in the Internet.

There has been a great deal of theoretical work on designing random graph models that result in a Web-like graph. Barabási and Albert [5] describe the preferential attachment model, where the graph grows continuously by inserting nodes, where new node establishes a link to an older node with a probability which is proportional to the current degree of the older node. Bollobás et al. [7] analyze this process rigorously and show the desired property. Another model based on a local optimization process is described by Fabrikant et al [12]. Aiello et al. [1] studies random graphs with power law degree distribution and derives interesting structural properties in such graphs. Mitzenmacher [14] gives an excellent survey on the history and generative models for power law distributions.

III. Evolution of Social Networks

The functionality provided by the newer social networks are more and more interesting, however it is really hard to examine the real evolution of social networks. In this section we differ them based on what type of nodes and links do they support. In this investigation we focus on social networks which somehow involve mobile phones into their functionality. The reason is that the phonebook of the mobile phone basically represents some kind of social relationships between us and our contacts.

In this section we start our examination from the first social networks which were publicly available on the web. At the end of this investigation we define phonebook-centric social networks which we considered as one of the most advanced social networks nowadays from mobile phone support point of view.

A. Simple social network (as a web-application)

The simplest social networks basically provide a convenient way for users to upload their detailed profiles and find familiars or old friends, who have also registered into the network. The structure of a simple social network is shown on Figure 1.

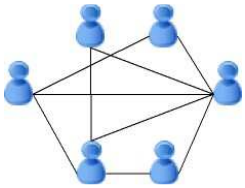

Figure 1. Simple social network

In the case of a *simple social network* nodes are the registered users and links represent social relationships between them. We denote the set of registered users by U_U and links are denoted by E_{UU}. Formally:

$$G_{SSN} = (U_U, E_{UU}), \text{ where}$$
$$E_{UU} \subseteq \{(u_U, u'_U) : u_U, u'_U \in U_U, u_U \neq u'_U\} \quad (1)$$

G_{SSN} is the (directed) graph representation of a simple social network where the points are the users and edges are social relationships marked by the users.

B. Phonebook-enabled social network

Phonebook-enabled social networks have a more advanced structure (Figure 2) because of the mobile phone support.

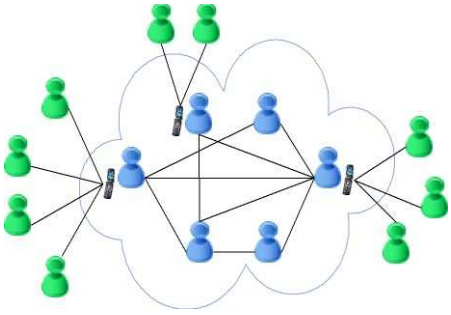

Figure 2. Phonebook-enabled social network

In a phonebook-enabled social network there are two types of nodes corresponding to members and private contacts.

Definition 1. A *member* is a registered user of the social network. Basically, members are similar to users of simple social networks. They can log into the system, find and add acquaintances, upload and share information about themselves, write forum or blog entries, etc. The key difference between members of a phonebook-enabled social network and users of a simple social network is that members can upload their contact list to the social network and maintain a backup phonebook there. We denote the set of registered members by U_M.

Definition 2. A *private contact* corresponds to a phonebook entry of a member. Each member may have multiple private contacts. However, these private contacts are not shared between members. A private contact is transferred into the system when a member synchronizes his or her phonebook with the social network. We denote the set of private contacts in the phonebooks by U_{Pc}.

In a phonebook-enabled social network the sets U_M and U_{Pc} are disjoint sets. Although, a member private contact in a phonebook may refer to the same person, they are handled separately. The main advantage of a phone-enabled social network is that the contacts in the phonebook of a member become independent from the current phone of the member.

Relationships between members are represented by the edge set E_{MM} and relationships that a private contact belongs to a member are represented by the edge set E_{MPc}, i.e.

$$E_{MM} \subseteq \{(u_M, u'_M) : u_M, u'_M \in U_M, u_M \neq u'_M\},$$
$$E_{MPc} \subseteq \{(u_M, u_{Pc}) : u_M \in U_M, u_{Pc} \in U_{Pc}\}.$$

A phonebook-enabled social network is represented by a (directed) graph

$$G_{PESN} = (U, E), \text{ where}$$
$$U = U_M \cup U_{Pc} \text{ and}$$
$$E = E_{MM} \cup E_{MPc} \quad (2)$$

C. Phonebook-centric social network

In a phonebook-enabled social network it is possible that one of our private contacts in our phonebook is similar to a member of the network, following we will refer to this as *similarity*. A similarity detection algorithm enables more advanced functionality for social networks. Such an algorithm allows us to detect and resolve similarities in the network, recommend possible relationships for the members and ensure a more intelligent behavior to the network. In addition to that this algorithm enables also to recognize *duplications* in phonebooks. We will refer to phonebook-enabled social network with similarity and duplication detection algorithm as *phonebook-centric social network* (Figure 3).

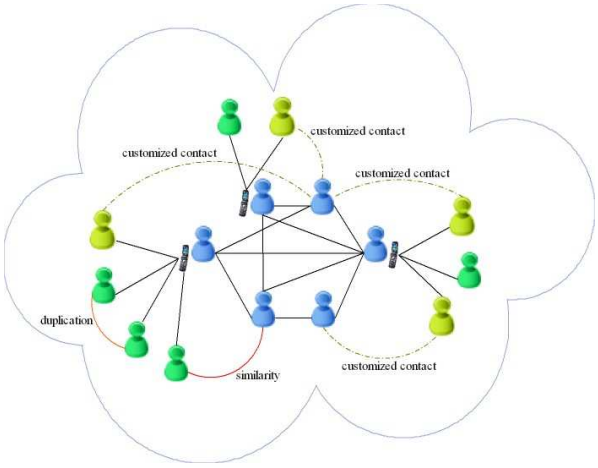

Figure 3. Phonebook-centric social network

In a phonebook-centric social network there are three types of nodes. Members and private contacts are similar to the ones defined in the previous section, however if a member resolves a similarity between another member and a private contact a new customized contact appears.

Definition 3. A *customized contact* is created from a private contact when a member is similar to a private contact and the owner member of the private contact marks them as similar person. In this way the owner can edit this contact in her or his phonebook but if the referred member changes her or his profile, the change will be propagated to the customized contact. However this propagation will take effect only if the owner member has not edited that specific profile detail yet. Following we will refer to this propagate mechanism as *customization*. The set of customized contacts is denoted by U_C.

Beside of the edge sets E_{MM} and E_{MPc}, a phonebook-centric social network contains a few more type of edges: the set E_{MoC} of edges between members and their customized contacts and the set E_{MC} of edges between customized contacts and the referred members. Formally,

$$E_{MoC} = \left\{ \begin{matrix} (u_{Mo}, u_C) : u_{Mo} \in U_M, u_C \in U_C, \exists\, u'_M \in U_M, \\ u'_M \neq u_{Mo}, (u_{Mo}, u'_M) \in E_{MM}, (u'_{Mo}, u_C) \in E_{MC} \end{matrix} \right\}$$
$$E_{MC} = \left\{ \begin{matrix} (u_M, u_C) : u_M \in U_M, u_C \in U_C, \exists\, u'_M \in U_M, \\ u'_M \neq u_M, (u_M, u'_M) \in E_{MM}, (u'_M, u_C) \in E_{MoC} \end{matrix} \right\} \quad (3)$$

In addition, there are two more type of edges. The set E_S of edges indicate similarities between private contacts and members of the network and the set E_D of edges indicate (potential) duplications between private contacts of a member. Formally,

$$E_S = \left\{ \begin{matrix} (u_{pc}, u_m) : u_m \in U_M, u_{pc} \in U_{PC}, (u_{pc}, u_m) \notin E_{MPc}, \\ \exists (u_{pc}, u'_m) \in E_{MPc}, \exists (u'_m, u_m) \in E_{MM}, u'_m \in U_M \end{matrix} \right\}$$
$$E_D = \left\{ \begin{matrix} (u_{pc}, u'_{pc}) : u_{pc}, u'_{pc} \in U_{PC}, \\ u_{pc} \neq u'_{pc}, \exists \left((u_{pc}, u_m), (u'_{pc}, u_m) \right) \in E_{MPc}, \\ u_m \in U_M \end{matrix} \right\} \quad (4)$$

A phonebook-centric social network is represented by a graph:

$$G_{PBSN} = (U, E), \text{ where}$$
$$U = U_M \cup U_{Pc} \cup U_C \text{ and}$$
$$E = E_{MPc} \cup E_{MoC} \cup E_{MC} \cup E_D \cup E_S \quad (5)$$

The number of edges in E_{MC} is a key point in phonebook-centric social networks from the performance point of view because of the customization mechanism. In this work we propose a model how to calculate the distribution of E_{MC} edges and investigate the structure of phonebook-centric social networks including similarities. The study is based on measurements using a phonebook-centric social network implementation, called *Phonebookmark*.

IV. PHONEBOOKMARK

A. Phonebookmark functions

Phonebookmark implements all phonebook-centric social network features. We have described its similarity handling algorithm in [10], which allows also to detect duplications in phonebooks and handles even similar names like 'Joe' and 'Joseph'. *Phonebookmark* contains also other popular social networking features in order to increase its popularity like photo sharing, instant messaging, forum, blog and a general search engine. In addition to that it supports a Java ME-based mobile client which basically allows for members to keep their contacts up-to-date via a synchronization mechanism to the social network.

B. Resolving duplications and similarities

In order to understand the proposed model for calculating the distribution of similarities in phonebook-centric social networks, first we have to get familiar with the similarity resolving mechanism.

Phonebookmark provides a semi-automatic similarity detecting and resolving mechanism. First it detects similarities and calculates similarity weight values which are used to calculate probability values regarding to the accurate of the detected similarities. *Phonebookmark* uses this probability also to determine the proper order of multiple similarities (Figure 4).

Figure 4. Handling multiple similarities

After a detected similarity is being selected, *Phonebookmark* provides a user interface where the details of the two people can be merged. Here the user can choose whether to resolve or ignore the similarity, which is the base of the semi-automatic behavior (Figure 5).

Figure 5. Semi-automatic similarity resolution

Phonebookmark is also able to handle different versions of the same first name, because of the semi-automatic similarity resolution mechanism. If a user resolves a similarity, a mechanism checks if the first names of the two users are different and stores them in a database table. Later if these first names were found similar several times, the algorithm handles them as similar first names. Table 1 illustrates a snapshot from the 'similarname' database table.

TABLE I. SIMILAR NAMES DATABASE TABLE

name1	name2	count
Joe	Joseph	10
Kathrine	Kate	7
Samantha	Sam	2

The test data of Phonebookmark had 420 registered members with more than 72000 private contacts. During the operational period the algorithm detected 2000 similarities and users have resolved more than 90% of these, which is an encouraging number for analyzing the distribution of similarities and propose a model for it.

V. MEASUREMENTS AND RESULTS

A. Distribution of in- and out-degrees

The first experiment we conducted was to analyze the distribution of in- and out-degree of the members in *Phonebookmark*. As expected, the distribution of both follows a power law.

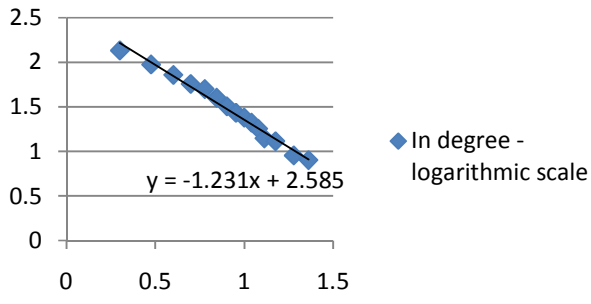

Figure 6. Log-log plot of the distribution of in-degree in *Phonebookmark*

Figure 6 illustrates the in degree of *Phonebookmark* on logarithmic scale. The *x*-axis represents the base 10 logarithm of in-degrees, while the *y*-axis is the base 10 logarithm of the number of nodes which have at least that amount of incoming edges, thus the amounts are aggregated. According to [4], a nonnegative random variable X is said to have a power law distribution in the following case:

$$\Pr[X \geq x] \sim cx^{-\alpha} \quad (6)$$

for constant c>0 and α>0. In a power law distribution asymptotically the tails fall according to the power α. Such a distribution leads to much heavier tails than other common models, such as exponential distributions.

According to the previous statements if X has a power law distribution, then in a log-log plot of $\Pr[X \geq x]$, also known as the *complementary cumulative distribution function*, asymptotically the behavior will be a straight line. This provides a simple empirical test for whether a random variable has a power law given an appropriate sample. In this case the gradient of the line in the log-log plot is the α parameter of the given power law distribution:

$$\ln(\Pr[X \geq x]) = -\alpha \ln(x) + \ln(c) \quad (7)$$

The out-degree of *Phonebookmark* is depicted Figure 7 using logarithmically scaled *x*- and *y*-axis.

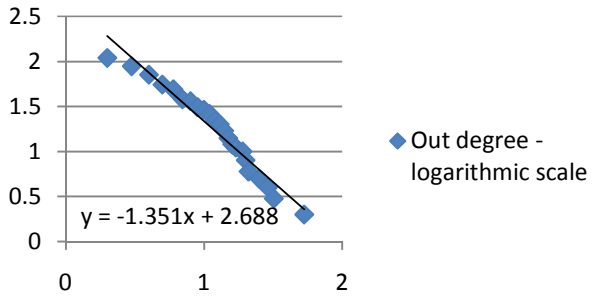

Figure 7. Log-log plot of the distribution of out degree in *Phonebookmark*

The *x*-axis represents the number of out degrees and the *y*-axis the number of nodes which have at least that amount of outgoing edges. Note that both in- and out-degrees on *Phonebookmark* follows power law distribution, thus it is similar to general social networks.

In several papers (e.g. [5]) a power law distribution is referred as:

$$\Pr[X = x] \sim c' x^{-\beta} \quad (8)$$

(8) can be obtained by the derivation of the right hand side of (6), where β=α+1 and *c'*=α**c*, see for example [1],[2].

It is interesting to observe that in [5] is reported that the degree distribution of the collaboration graph of movie actors follows a power law with $\beta = 2.3 \pm 0.1$ (i.e. to $\alpha = 1.3 \pm 0.1$) in which range the distribution of our social network falls.

Another famous example for power law in degree distribution is the distribution of in- and out-degree distribution of the web graph reported in [8], where the exponents are $\beta = 2.09$ for in degree and $\beta = 2.72$ for out degree ($\alpha = 1.09$ and 1.72 respectively).

B. *Distribution of similarities*

Based on the database and database logs of *Phonebookmark* we managed to measure the distribution of similarities raised by a member during registration and first phonebook synchronization. Figure 8 shows the number of similarities, where the *x*-axis is the number of similarities and the *y*-axis means how many people arises at least that amount of similarities when registers and synchronizes.

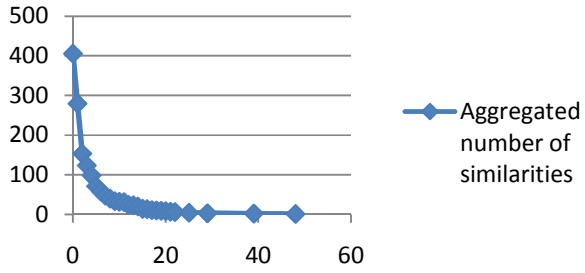

Figure 8. Number of similarities

For example, we can see that 32 people arises at least 10 similarities. The shape of this function is quite similar to a heavy tail function which is not unusual in case of internet and social network related distributions. Figure 8 illustrates the previous function on a logarithmic scale.

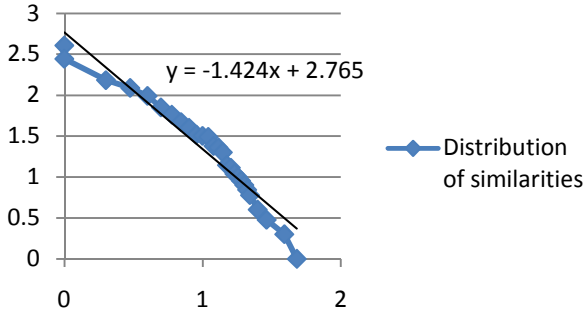

Figure 9. Distribution of out similarities in *Phonebookmark*

We can see on Figure 9 that the shape of the function is close to a linear, thus the distribution is power law [2]. We illustrated the linear on the figure, thus we can read that the α parameter in the power law distribution is 1.4249 and $\beta = 2.4249$. According to the previous measurement the distribution of similarities in our case is:

$$\Pr[X \geq x] \sim x^{-1.4249} \quad (9)$$

Consider the number of edges in E_{MC} (number of similarity edges). The evidence that the distribution of the similarities follows a power law has practical consequences. The expected number of users involving at least a certain number of similarities x can be estimated by $N * \Pr[X \geq x] \sim N * x^{-1.4249}$, where N is the number of members in the network.

VI. Conclusion and Future Work

Phonebook-centric social networks belong to the new generation of social networks. They provide several new features, which have several interesting research implications as well.

In this paper we have studied the evolution of social networks and we have defined phonebook-centric social networks. We have introduced *Phonebookmark* which is a phonebook-centric social network implementation with several additional features. *Phonebookmark* was available for a group of general users from April to December of 2008. It had 420 registered members with more than 72000 private contacts. Based on measurements from this period we have analyzed the in- and out-degree distribution of *Phonebookmark* and experienced that these distributions follow a power law.

As a main contribution of this paper we showed a graph based description for phonebook-centric social networks and we have proposed a model, that shows, that the distribution of similarities generated by members follows power law distribution.

Future work plans include analyzing the size of phonebooks of members and propose an extended the model for calculating the number of similarities based on this analysis. Additional plan is to examine duplications of phonebooks which basically a straight consequence of a similarity detecting mechanism.

Acknowledgment

The authors would like to express their thanks to Balázs Bakos, Zoltán Ivánfi for their support from Nokia Siemens Networks.

References

[1] L. A. Adamic. Zipf, power-law, Pareto -- a ranking tutorial. http://www.hpl. hp.com/research/idl/papers/ranking, 2000.

[2] L. A. Adamic, B.A. Huberman. Zipf's law and the Internet. *Glottometrics*, Volume 3, pages: 143-150, 2002.

[3] W. Aiello, F. R. K. Chung, L. Lu. A random graph model for massive graphs. In: *Proc. STOC 2000*: pages: 171-180, 2000.

[4] Alexa. http://www.alexa.com/site/ds/top_sites. May 2009.

[5] A.-L. Barabási, R. Albert: Emergence and scaling in random networks. *Science*, Volume 286, pages: 509-512, 1999.

[6] A.-L. Barabási. Linked: How Everything Is Connected to Everything Else. Perseus Publishing, 2002.

[7] B. Bollobás, O. Riordan, J. Spencer, G. Tusnady. The degree sequence of a scale-free random graph process. *Random Structures and Algorithms*, Vol. 18(3), pages: 279 – 290, 2001.

[8] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proc. of the 9th international World Wide Web conference on Computer networks*, 2000.

[9] M. E. Crovella, M. S. Taqqu and A. Bestavros. Heavy-Tailed Probability Distributions in the World Wide Web. In: R.J. Adler, R.E. Feldman, M.S. Taqqu (eds.), A Practical Guide To Heavy Tails. 1, pages: 3--26. Chapman and Hall, New York. 1998.

[10] P. Ekler, Z. Ivánfi, K. Aczél. Similarity Management in Phonebook-centric Social Networks. *The Fourth International Conference on Internet and Web Applications and Services* (ICIW 2009), May 2009.

[11] Facebook. http://www.facebook.com. May 2009.

[12] A. Fabrikant, E. Koutsoupias, and C. H. Papadimitriou. Heuristically Optimized Trade-offs: A New Paradigm for Power Laws in the Internet. In *Proc. of ICALP*, pages: 110-122. Springer-Verlag LNCS, 2002.

[13] B. A. Huberman, L. A. Adamic. Growth dynamics of the World-Wide Web. *Nature*, Vol. 401, page 131, 1999.

[14] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions, *Internet Mathematics*, Volume 1, pages: 225-251, 2001.

[15] Myspace. http://www.myspace.com. May 2009.

[16] Nazir, S. Raza and C.-N. Chuah. Unveiling Facebook: A measurement Study of Social Network Based Applications. In: *Proc. ACM Internet Measurement Conference (IMC*), 2008.

[17] V. Pareto. Course d'economie politique professé à l'université de Lausanne, 3 volumes, 1896-7.

[18] M. Ripeanu, I. Foster, and A. Iamnitch. Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design. *IEEE Internet Computing Journal*, Volume 6, 2002.

[19] G. U. Yule. Statistical study of literary vocabulary, Cambridge University Press, 1944.

[20] G. K. Zipf. The Psycho-Biology of Language. An Introduction to Dynamic Philology. Houghton Mifflin, Boston, MA, 1935

[21] G. K. Zipf. Human behavior and the principle of least effort, Addison-Wesley, 1949.

.