Acta Cybernetica **19** (2010) 735–748.

Evaluating Dynamically Evolving Mobile-Based Social Networks

Péter Ekler*, Tamás Lukovszki†, and Hassan Charaf‡

Abstract

The increasing capabilities of mobile phones enable them to participate in different type of web-based systems. One of the most popular systems are social networks. The phonebooks of the mobile devices also represent social relationships of the owner. This can be used for discovering additional relations in social networks. Following this line of thought, mobile-based social networks can be created by enabling a synchronization mechanism between phonebooks of the users and the social network. This mechanism detects similarities between phonebook contacts and members of the network. Users can accept or ignore these similarities. After acceptance, identity links are formed. If a member changes her or his personal detail, it will be propagated automatically into the phonebooks, via identity links after considering privacy settings. Estimating the total number of these identity links is a key issue from scalability and performance point of view in such networks. We have implemented a mobile-based social network, called Phonebookmark and examined the structure of the network during a test period of the system. We have found, that the distribution of identity links of the users follows a power law. Based on this, we propose a model for estimating the total number of identity links in the dynamically evolving network. We verify the model by measurements and we also prove the accuracy of the model mathematically. For this we use the fact, that the number of identity links of each user (and thus, the value of the random variable modeling it) is bounded linearly by the number of members N_M of the network. Then we show, that the variance of the random variable is $\Theta(N_M^{3-\beta})$, where $2 < \beta \leq 3$ is the exponent of the bounded power law distribution, i.e. for constant $c > 0$, $\Pr[X = x] = c \cdot x^{-\beta}$, if $x \leq N_M$ and $\Pr[X = x] = 0$ otherwise. The model and the results can be used in general when the distribution shows similar behavior.

Keywords: networks, social networks, power law distribution, variance

*Department of Automation and Applied Informatics, Budapest University of Technology and Economics, E-mail: peter.ekler@aut.bme.hu. Supported by the project TÁMOP 4.2.1/B-09/1/KMR-2010-0002 of the National Development Plan.

†Faculty of Informatics, Eötvös Loránd University, E-mail: lukovszki@inf.elte.hu. Supported by the project TÁMOP 4.2.1/B-09/1/KMR-2010-0003 of the National Development Plan.

‡Department of Automation and Applied Informatics, Budapest University of Technology and Economics, E-mail: hassan.charaf@aut.bme.hu. Supported by the project TÁMOP 4.2.1/B-09/1/KMR-2010-0002 of the National Development Plan.

1 Introduction

In the last decade the Internet related technologies developed rapidly. One of the most popular solutions are social network sites. Since their introduction, social network sites such as Facebook, Myspace and LinkedIn have attracted millions of users.

According to new statistics [8] Facebook has more than 500 million active users, 50% of the active users log in to Facebook every day. Other statistics show that there are more than 65 million active users currently use their mobile devices for accessing Facebook. Those mobile users are almost 50% more active on Facebook than non-mobile users.

The fact, that the phonebook of the mobile device also describe social relationships of its owner, can be used for discovering additional relations in social networks. This is beneficial for sharing personal data or other content. 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. Besides that we can see all of our contacts on the mobile phone as well as on the web interface. In addition to that if the system detects that some of my private contacts in the phonebook is similar to another registered members of the social network (i.e. may identify the same person), it can discover and suggest social relationships automatically. Accepted similarities are called identities. In the rest of this paper we refer to this solution as a *mobile-based social network*.

If a member changes some of her or his detail, it should be propagated in every phonebook to which she or he is related. In addition to that, with the help of identity links, the system can keep the phonebooks always up-to-date. In this paper we show how to calculate the expected number of identities which is a key issue from scalability point of view and we propose a model which proves the accuracy of the calculation. The model is based on power law distribution and the results can be used in general cases as well. The results were applied in Phonebookmark project at Nokia Siemens Networks.

The rest of the paper is organized as follows. Section 2 summarizes related work in the field of dynamically evolving large networks and power law distribution. Section 3 defines mobile-based social networks. Section 4 proposes a model for estimating the total number of similarities. Section 5 proves the accuracy of the model and gives an estimation for the variance of power law distribution when the random variable has an upper bound. Finally Section 6 concludes the paper and proposes further research area.

2 Related work

Huge amount of papers and popular books, such as Barabási's Linked [2] study the structure and principles of dynamically evolving large scale networks like the Internet and networks of social interactions. In [9] the authors discuss about that nowadays social networking on mobile phones is not only a buzz term for today's

enthusiasts but also provides real possibilities to the users. Many features of social processes and the Internet are governed by power law distributions. Following the terminology in [7] a nonnegative random variable X is said to have a power law distribution if $\Pr[X \geq x] = cx^{-\alpha}$, 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 1896 by Pareto [12] (see. [13]), while describing the distribution of income in the population. Zipf observed similar statistical behavior in the distribution of inhabitants in cities [14].

In [4] 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. [11] showed that the in- and out-degree distribution of the interaction graph of the studied social network applications also follow such distributions. Those distributions also approximate the degree distribution of the Gnutella network [13]. Crovella et al. [5] 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 [3] 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. [4] analyze this process rigorously and show that the degree distribution of the resulting graph follow a power law. Another model based on a local optimization process is described by Fabrikant et al. [7]. Mitzenmacher [10] gives an excellent survey on the history and generative models for power law distributions. Aiello et al. [1] studies random graphs with power law degree distribution and derives interesting structural properties in such graphs.

In our work power law distribution was also discovered in case of the number of identities. We have given a model for estimating the total number of identities and we have proven the accuracy of the model, where the variance of power law distribution was examined.

3 Mobile-based social network

Mobile-based social networks rely on the well-known social network sites, they have a similar web interface, but they add several major mobile phone-related functions to the system. Next we consider social networks as graphs. In case of general social networks, nodes are representing registered members and the edges between them represent the social relationships (e.g. friendship). Then we should notice that each member has a private mobile phone with a phonebook (Figure 1). In Figure 1, we can also observe that phonebook contacts are connected to the mobile devices "owned" by different members.

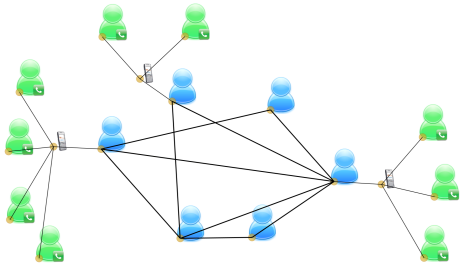

Figure 1: Basic structure of mobile-based social networks

One of the key advantages of mobile-based social networks is that they allow real synchronization between private phonebook contacts and the social network. For this a similarity detecting algorithm is needed. This algorithm is able to compare two person entries (members and private contacts, too) and determine whether they are likely similar, if so, it also proposes a probability to this detected similarity.

Figure 2 shows the graph structure when the similarity detecting algorithm has finished comparing the relevant person entries.

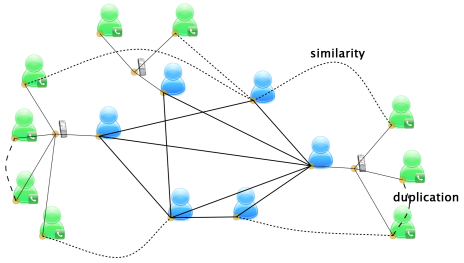

Figure 2: Detected similarities and duplications

In Figure 2 the dotted edges between members and private contacts represent detected similarities and broken lines between two private contacts illustrate possible duplications in the phonebooks. Duplications are detected as a positive side effect of the similarity detecting algorithm.

After the similarities and duplications are detected there is a semi-automatic step, the members who have phonebook contacts detected as similar to other members in the network have to decide whether detected similarities are the correct ones. In addition to that, members can also decide about the correctness of detected duplications in their phonebooks. Figure 3 shows the graph structure after some of the members have resolved the detected similarities and duplication. It can be observed that one of the private contacts of the most left member has been deleted. The other duplication link still remained on the right side because that

member has not decided about it yet.

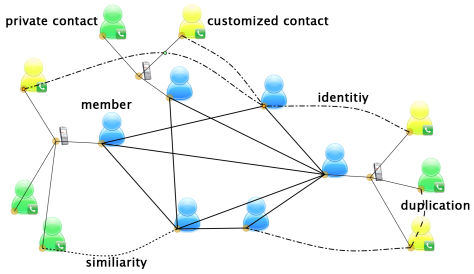

Figure 3: Resolving similarities and duplications

Figure 3 shows an example, where four of the five similarities were resolved (members found them correct) and there is still one in the system (the member has not decided yet). Resolving a similarity means that an identity link is being formed between the private contact in one's phonebook and the relevant member who represents the same person in the system. The private contacts that are linked to members via this type of identity links are called customized contacts. One of the key advantages of mobile-based social networks are these identity link, since if a member changes her or his personal detail on the web user interface (adds a new phone number, uploads a new image, changes the website address, etc), it will be automatically propagated to those phonebooks where there is a customized contact related to this member, after considering privacy issues. Additional important advantages of mobile-based social networks are:

- Private contacts can be managed (list, view, edit, delete, etc.) from a browser.
- Similarity detecting algorithm detects duplicate contacts in the phonebooks and warns about it.
- Private contacts are safely backed up in case the phone gets lost.
- Private contacts can be easily transferred to a new phone if the user replaces the old one.
- Phonebooks can be shared between multiple phones, if one happens 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 private contacts in the phonebooks and warns about it.

The described mobile-based social network architecture was actually applied in the *Phonebookmark* project at Nokia Siemens Networks. Phonebookmark covered a wide range of mobile phones with the Symbian and Java ME clients. We took

part in the implementation and before the 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 suitable number for analyzing the behavior of the network. During this period we have collected and measured different types of data related to the social network and its behavior.

We would like to highlight that the proposed mobile-based social network architecture extends the general social networks. Based on this, existing, large systems can be upgraded easier to involve mobile phones in their operation. This also indicates that it is important to examine such solutions from the performance and scalability point of view.

4 Expected value in power law based models

The additional resource requirements of mobile-based social networks depend at most from the number of identity links compared to general social networks, because the number of synchronizations depends on them. In this section we propose a model for calculating the total number of identities.

4.1 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 phonebook synchronization.

Figure 4 shows the complementary cumulative distribution function of 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 register and synchronize.

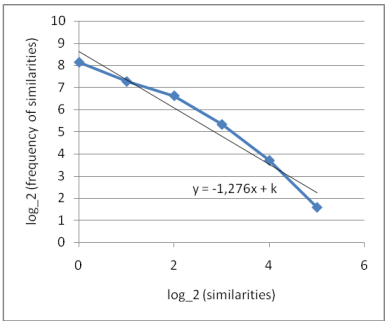

Figure 4: Distribution of similarities

We use logarithmically scaled x- and y-axis. We can see in Figure 4 that the points can be well approximated with a straight line by the least squares method,

thus the distribution of similarities can be well approximated by a power law. The exponent of the power law distribution is $\alpha = 1.276$.

According to this measurement the distribution of similarities in our case can be well approximated as follows:

$$\Pr[X \geq x] = x^{-1.276} \tag{1}$$

The evidence that the distribution of similarities follows a power law has practical consequences. The expected number of members involving at least a certain number of similarities x can be estimated by $N_M \Pr[X \geq x] = N_M x^{-1.276}$, where N_M is the number of members in the network.

4.2 Calculating the expected value

Based on the previous measurement we can estimate the total number of identities.

Theorem 1. *The number of identities N_I in a mobile-based social network can be well approximated with $N_I = N_M \frac{\varsigma(\beta-1)}{\varsigma(\beta)} P_R$, where N_M represents the number of members in the system, P_R is the acceptance rate of the similarities by the users, $\varsigma(.)$ denotes the Riemann Zeta function, $\beta = \alpha + 1$ and $\alpha > 1$ is the parameter of the distribution.*

Remark 1. $P_R \approx 0.9$ is the accuracy rate of the similarity detecting algorithm discussed in [6]. The proof is based on measurements, the importance relies on that the number of similarities has not been modeled before. The trend can be seen very well on this data set.

Proof. (Theorem 1)

We model the number of similarities generated during a member registration by a random variable X. More precisely, X models the number of similarities related to a member. First we calculate the expected number of similarities:

$$E[X] = \sum_{x=1}^{\infty} x \Pr[X = x] \tag{2}$$

Note that x starts from one, because a new member registration involves at least one similarity, because the system allows registration only by invitation, therefore the new member is already in the phonebook of the inviting member.

Then the total number of identities N_I in a mobile-based social network can be calculated with the following formula:

$$N_I = N_M E[X] P_R. \tag{3}$$

In order to calculate $E[X]$, we need to determine $\Pr[X = x]$, which can be obtained from (1) by derivation:

$$\Pr[X = x] = c' \frac{1}{x^{\beta}}, \tag{4}$$

where $\beta = \alpha + 1$. In order to have a probability distribution,

$$\sum_{x=1}^{\infty} c' x^{-\beta} = 1. \tag{5}$$

Therefore,

$$c' = \frac{1}{\sum_{x=1}^{\infty} x^{-\beta}} = \frac{1}{\varsigma(\beta)}, \tag{6}$$

where $\varsigma(.)$ denotes the Riemann Zeta function.
Then the expected value can be calculated as:

$$\begin{aligned} E[X] &= \sum_{x=1}^{\infty} x \Pr[X = x] && (7)\\ &= \sum_{x=0}^{\infty} x \frac{1}{\varsigma(\beta)} x^{-\beta} && (8)\\ &= \frac{1}{\varsigma(\beta)} \sum_{x=1}^{\infty} x^{1-\beta} && (9)\\ &= \frac{\varsigma(\beta-1)}{\varsigma(\beta)}. && (10) \end{aligned}$$

Finally, based on (3), the expected total number of identities N_I in a mobile related social network can be estimated with the following formula:

$$N_I = N_M \frac{\varsigma(\beta-1)}{\varsigma(\beta)} P_R \tag{11}$$

For $\beta > 2$, $\varsigma(\beta-1)/\varsigma(\beta)$ is a constant.

□

During the operational period of Phonebookmark 1088 identities were detected. By applying the identity estimation model in Theorem 1, for $\beta = 2.276$, we obtain that the expected total number of identities is $N_I = 2.9196 \cdot 420 \cdot 0.9 = 1103$, which is very close to the measured number.

Since mobile-based social networks are new type of social networks, we could not perform the measurements on other databases. However in the next section we prove the accuracy of the identity model mathematically. This result can be used widely in other similar cases, since power law distribution occurs often in social networks and the Web-graph.

5 Variance of power law distribution

In the identity estimation model in Section 4 we used a random variable X which represents the number of similarities raised by a member and we showed that X

follows a power law distribution. For $\alpha \leq 2$, a power law distribution has infinite variance. Thus, for $1 < \alpha \leq 2$, the accuracy of the estimation in Theorem 1 is an issue. The law of large numbers states that the total number of identities converges to their expected value. For this, an assumption of finite variance of the variables is not necessary. However, in order to obtain much faster convergence and error probability bound, finite variance is needed. In case of finite variance also the central limit theorem can be applied.

In this section we show that the random variable X has a relevant upper bound (in our case, this upper bound is linear in the number of the members of the network). This can be used to calculate an accurate variance value, if $1 < \alpha \leq 2$. After that the central limit theorem can be used in order to obtain, that the total number of identities will be close to their expected value.

Following we highlight that identities raised by a member have a relevant upper bound, then we propose and prove a general theorem to calculate the variance of upper bounded power law distributions. Finally, we show how to apply it in case of mobile-based social networks.

Fact: If the phonebooks do not contain duplicates then the number of similarities caused by a member is at most $2(N_M - 1)$.

With other words, in the interval $[1, 2(N_M - 1)]$ the distribution of similarities follows a power law and the probability of more similarities is zero. In order to see this, note that a member can be similar to at most one private contact of each of the other $N_M - 1$ members and, for each private contact, there is at most one similar member in the network.

We show that similarities resulting from this fact has a finite variance.

Theorem 2. *Let X be a random variable with $\Pr[X = x] = c \cdot x^{-\beta}$ if $x \leq n$ and $\Pr[X = x] = 0$ otherwise, where $2 < \beta \leq 3$ and $c > 0$ is a constant. In this case the variance $\sigma^2 X$ of X is $\sigma^2 X = \Theta\left(n^{3-\beta}\right)$.*

For the proof we use two lemmatas.

Lemma 1. *Let X be a random variable with $\Pr[X = x] = c \cdot x^{-\beta}$ if $x \leq n$ and $\Pr[X = x] = 0$ otherwise, where $2 < \beta \leq 3$ and $c > 0$ is a constant. In this case the variance is $\sigma^2 X = O\left(n^{3-\beta}\right)$.*

Proof. From the *Steiner formula*, the variance is calculated as $\sigma^2 X = E[X^2] - (E[X])^2$. $E[X]$ was defined previously. For $\beta > 2$, $E[X]$ is a finite constant, i.e. $E[X] = \Theta(1)$. Thus we only need to calculate $E[X^2]$. By definition:

$$E[X^2] = \sum_{x=1}^{\infty} x^2 \Pr[X = x] \tag{12}$$

$$= \sum_{x=1}^{n} x^2 \Pr[X = x] \tag{13}$$

$$= \sum_{x=1}^{n} x^2 c x^{-\beta} \tag{14}$$

$$= c \sum_{x=1}^{n} x^{2-\beta}, \tag{15}$$

where $c = \frac{1}{\sum_{x=1}^{n} x^{-\beta}}$ and $\Pr[X = x] = cx^{-\beta}$. Then $\sum_{x=1}^{\infty} \Pr[X = x] = 1$.

Let $y = \frac{1}{c} E[X^2]$. Following we show an upper estimation for y. In order to do so, we create an upper estimation model for the function of y by using the powers of $1/2$. Let $z = 2^{\frac{1}{2-\beta}}$, then:

$$\frac{1}{c} z^2 \Pr[x = z^i] = \frac{1}{\left(2^{i \frac{1}{2-\beta}}\right)^{\beta-2}} = \frac{1}{2^i} \tag{16}$$

Figure 5 illustrates how we performed the estimation, with the $f1$ function.

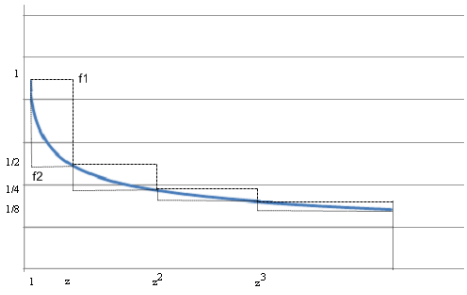

Figure 5: Staged estimation function

Now we are able to give an upper bound on y.

$$
\begin{aligned}
y &\leq \sum_{i=0}^{\log_z n} \left(z^{i+1} - z^i\right) 2^{-i} && (17)\\
&= \sum_{i=0}^{\log_z n} (z-1) z^i 2^{-i} && (18)\\
&= (z-1) \sum\nolimits_{i=0}^{\log_z n} \left(\frac{z}{2}\right)^i && (19)\\
&= (z-1) \left(\frac{\left(\frac{z}{2}\right)^{\log_z n + 1} - 1}{\frac{z}{2} - 1} \right) && (20)\\
&= (z-1) \left(\frac{\frac{z}{2} n^{\frac{1}{\log_{z/2} z}} - 1}{\frac{z}{2} - 1} \right) && (21)\\
&= (z-1) \left(\frac{\frac{z}{2} n^{\frac{1}{1+\log_{z/2} 2}} - 1}{\frac{z}{2} - 1} \right). && (22)
\end{aligned}
$$

The explanation to the last step:

$$\log_{z/2} z = \log_{z/2} 2\frac{z}{2} = 1 + \log_{z/2} 2 \tag{23}$$

To continue, first we have to check the following calculation. Remember that $z = 2^{\frac{1}{2-\beta}}$. Then

$$\log_{z/2} 2 = \frac{\log_2 2}{\log_2 z/2} = \frac{1}{\log_2 \left(\frac{2^{\frac{1}{2-\beta}}}{2} \right)} = \frac{1}{\frac{1}{\beta-2} - 1} = \frac{\beta - 2}{3 - \beta} \tag{24}$$

Therefore,

$$n^{\frac{1}{1+\log_{z/2} 2}} = n^{\frac{1}{1+\frac{\beta-2}{3-\beta}}} = n^{3-\beta} \tag{25}$$

This way y looks as:

$$y \leq (z-1) \left(\frac{\frac{z}{2} n^{\beta-3} - 1}{\frac{z}{2} - 1} \right) \tag{26}$$

Next we show that the variance by applying the *Steiner formula* and the previous calculations is $O(n^{3-\beta})$.

$$\begin{aligned}
\sigma^2 X &= E[X^2] - (E[X])^2 && (27)\\
&\leq \frac{1}{c} y - \Theta(1) && (28)\\
&= \frac{1}{c}(z-1)\left(\frac{\frac{z}{2}n^{\beta-3}-1}{\frac{z}{2}-1}\right) - \Theta(1) && (29)\\
&\leq \frac{1}{c}\frac{(z-1)\,z}{z-2} n^{3-\beta} - \Theta(1) && (30)\\
&= O\left(n^{3-\beta}\right). && (31)
\end{aligned}$$

□

Lemma 2. *Let X be a random variable with $\Pr[X = x] = c \cdot x^{-\beta}$ if $x \leq n$ and $\Pr[X = x] = 0$ otherwise, where $2 < \beta \leq 3$ and $c > 0$ is a constant. In this case the variance is $\sigma^2 X = \Omega\left(n^{3-\beta}\right)$.*

Proof. We use the notations of the previous lemma. We give a lower bound on y using function $f2$ is shown on Figure 5.

$$y \geq \sum_{i=0}^{\log_z n} \left(z^{i+1} - z^i\right) 2^{-(i+1)}, \tag{32}$$

which is the half of the upper bound given in (17). Then by following the steps of the proof of the previous lemma we obtain that

$$y \geq \frac{z-1}{2}\left(\frac{\frac{z}{2}n^{\beta-3}-1}{\frac{z}{2}-1}\right) = \Omega(n^{3-\beta}). \tag{33}$$

Therefore,

$$\sigma^2 X = \Omega(n^{3-\beta}). \tag{34}$$

□

Proof. (Theorem 2) The proof is straightforward by applying Lemma 1-2:
$\sigma^2 X = \Theta\left(n^{3-\beta}\right)$, because $\sigma^2 X = O\left(n^{3-\beta}\right)$ and $\sigma^2 X = \Omega\left(n^{3-\beta}\right)$.

□

Theorem 2 can be applied in case of mobile-based social networks, when X represents the number of similarities raised by a member and the upper bound is $n = 2(N_M - 1)$.

6 Conclusion and future work

Social network sites are becoming more and more important in everyday life. Phonebook-centric social networks enable to manage online and mobile relationships within one system. The key mechanism of such networks is a similarity handling algorithm which detects similarities between members of the network and phonebook entries.

The number of identities is a key parameter from scalability point of view. In this paper we have shown how to calculate the expected number of identities and we have proven the accuracy of that calculation. However the results can be used generally in case of power law distributions where the random variable has an upper bound.

Further work includes additional measurements and extending the model with phonebook duplication handling.

References

[1] Aiello, W., Chung, F. R. K., and Lu, L. A random graph model for massive graphs. In *Proc. 32nd Symposium on Theory of Computing STOC*, pages 171–180, 2000.

[2] Barabási, A.-L. Linked: How everything is connected to everything else. *Perseus Publishing*, 2002.

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

[4] Bollobás, B., Riordan, O., Spencer, J., and Tusnady, G. Random structures and algorithms. *IEEE Internet Computing Journal*, 18:279–290, 2001.

[5] Crovella, M. E., Taqqu, M. S., and Bestavros, A. 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, 1998.

[6] Ekler, P. and Lukovszki, T. Similarity distribution in phonebook-centric social networks. In *Proceedings of 5th International Conference on Wireless and Mobile Communications (ICWMC 2009)*, Cannes, France, 2009.

[7] Fabrikant, A., Koutsoupias, E., and Papadimitriou, C. H. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. In *Proceedings of the 29th International Colloquium on Automata, Languages and Programming (ICALP)*, pages 110–122, 2002.

[8] Facebook statistics. http://www.facebook.com/press/info.php?statistics, August 2010.

[9] Forstner, B. and Kelnyi, I. *Mobile Peer to Peer A Tutorial Guide*, chapter Mobile Social networking - Beyond the Hype, pages 161–190. Number ISBN 978-0-470-69992-8. Wiley, 2009.

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

[11] Nazir, Atif, Raza, Saqib, and nee Chuah, Chen. Unveiling facebook: A measurement study of social network based applications.

[12] Pareto, V. Course d'economie politique profess a l'universit de lausanne. 3, 1896.

[13] Ripeanu, M., Foster, I., and Iamnitch, A. Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing Journal*, 6:50–57, 2002.

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

Received 13th August 2010