Publ. Math. Debrecen
81/3-4 (2012), 251–270
DOI: 10.5486/PMD.2012.4635

Phonebook-centric social networks – dealing with similarities

By PÉTER EKLER (Budapest) and TAMÁS LUKOVSZKI (Budapest)

Abstract. The popularity of social networks is increasing rapidly. The capabilities of mobile phones enable them to participate in social network applications. In addition to the relations between the users of the social network, the phonebooks in the mobile phones also define social relations. The key idea behind *Phonebook-centric social networks* is that Phonebook-centric social networks also provide a synchronization mechanism between phonebooks of users and the social network. Similarities appear in the system when a contact in a phonebook is similar to a member of the social network. The load and the scalability of the system mainly depend on the number of similarities. By analyzing the data of the phonebook-centric social network implementation *Phonebookmark* we experience that the distribution of in- and out-degrees and of the similarities follow a power law. Based on these facts we propose a model, how to estimate the total number of similarities. We verify the accuracy of our estimation empirically and theoretically. For the empirical test we use the data of *Phonebookmark* collected during a time period of eight months. Finally, we prove an $O(N_M)$ upper bound on the total number of similarities, with high probability, i.e. with probability $1 - O(N_M^{-\gamma})$, where $\gamma > 1$ is an arbitrarily chosen constant and N_M is the number of members of the network.

1. Introduction

The popularity of social networks was noticeable in the last few years. Several social networks appeared and attracted millions of users. The online social networks FACEBOOK [14] and MYSPACE [18] are among the top ten visited websites of the Internet [4]. The basic idea behind these networks is that users can

Mathematics Subject Classification: 91D30, 90B18, 60–08.
Key words and phrases: social network, mobile computing, power law, scalability.
This project is supported by the New Hungary Development Plan (Project ID: TÁMOP-4.2.1/B-09/1/KMR-2010-0002 and TÁMOP-4.2.1/B-09/1/KMR-2010-0003).

maintain social relationships on these networks. The nodes of a social network correspond to individuals or organizations. The edges between the nodes represent social 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. Since the phonebook of the mobile device also describe the social relationships of its owner, discovering additional relations in social networks 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 so that we can also see all of our contacts on the mobile phone as well as on the web interface. We refer to this solution as a *phonebook-centric social network*.

One of the key features of phonebook-centric social networks is that the phonebook of the mobile phone is automatically updated with the latest information provided by the friends of its owner. This also means 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. For this, discovering and handling similarities is a key issue. This mechanism is resource intensive. The load and the scalability of the system mainly depend on the number of similarities in the network.

1.1. Our contributions. In this work we investigate the structure of the similarities and other characteristic parameters of the network and we also show a phonebook-centric social network implementation, called *Phonebookmark*. By analyzing the data of *Phonebookmark* we experience that the distribution of in- and out-degrees and of the similarities follow a power law. We propose a model, how to estimate the total number of similarities. We verify the accuracy of our estimation empirically and theoretically. For the empirical test we use the data of *Phonebookmark* collected during a time period of eight months. Finally, we prove an $O(N_M)$ upper bound on the total number of similarities, with high probability, i.e. with probability $1 - O(N_M^{-\gamma})$, where $\gamma > 1$ is an arbitrarily chosen constant and N_M is the number of members of the network.

1.2. Outline of the 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 dynamically evolving networks and generative models leading to those distributions. Section 3 presents the structure of phonebook-centric social networks. Section 4 introduces the phonebook-centric social network implementation Phonebookmark. Section 5 shows the results of measurements related to phonebook-centric social networks. Section 6 discusses about estimating the total number of similarities and shows by measurements the accuracy of the estimation. In Section 7 we prove an upper bound on the total number of similarities. Section 8 concludes the paper.

2. 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 networks of social interactions and the Internet. Many features of social processes and the Internet are governed by power law distributions. Following the terminology in [13] a nonnegative random variable X is said to have a power law distribution if $\Pr[X \geq x] = cx^{-\alpha}$, for constants $c > 0$ 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 which lead usually to the normal law.

Distributions with an inverse polynomial tail have been first observed in 1897 by Pareto [19] (see e.g. [17]), while describing the distribution of income in the population. In 1935 Zipf [23] and in 1944 Yule [22] 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 [24].

In [7] the graph structure of the Web has been investigated and it was shown that the distribution of in- and out-gree of the web graph and the size of weekly and strongly connected components are well approximated by power law distributions. Nazir et al. [20] showed that the in-and out-degree distribution of the interaction graph of the studied Facebook applications also follow such distributions. Those distributions also approximate the degree distribution of the Gnutella network [21]. 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 [6] describe the preferential attachment model, where the graph grows continuously by inserting nodes, such that a 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. [13]. Mitzenmacher [17] gives an excellent survey on the history and generative models for power law distributions. Aiello et al. [3] studies random graphs with power law degree distribution and derives interesting structural properties in such graphs.

The first experimental steps towards demonstrating the unique capabilities of mobile phones by bringing per-to-peer technology to mobile devices have been taken with the implementations of popular content sharing protocols, Gnutella and BitTorrent, for mobile phones [16]. In another paper [12] it was shown that even mainstream phones can be involved into BitTorrent network by implementing MobTorrent, which is Java ME based BitTorrent client. In [11] it is investigated how to involve these mobile clients into a hybrid content sharing system in order to increase their efficiency.

3. The structure of social networks

This section examines social networks based on what type of nodes and links do they support. We focus mainly 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.

We begin the 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.

3.1. Simple social network (like a web-application). The simplest social networks basically provide a convenient way for users to upload their detailed profiles and find relatives or old friends, who have also registered into the network. The structure of a simple social network is shown on Figure 1.

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

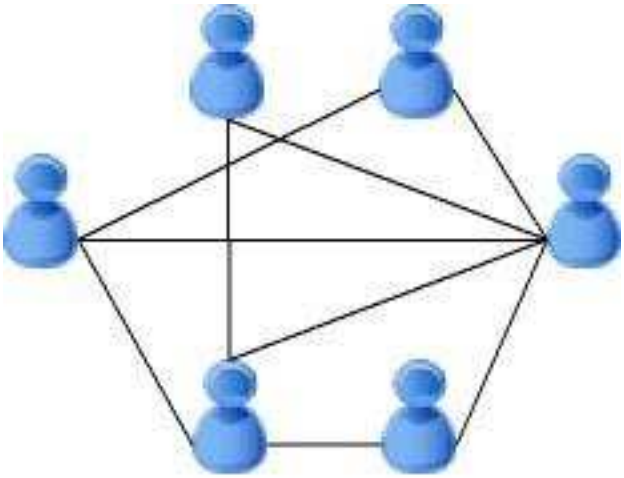


Figure 1. Simple social network

by U_U and links are denoted by E_{UU}. Then a simple social network is represented by a (directed) graph:

$$G_{SSN} = (U_U, E_{UU}) \text{ with} \tag{1}$$

$$E_{UU} \subseteq \{(u_U, u'_U) : u_U, u'_U \in U_U,\ u_U \neq u'_U\} \tag{2}$$

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

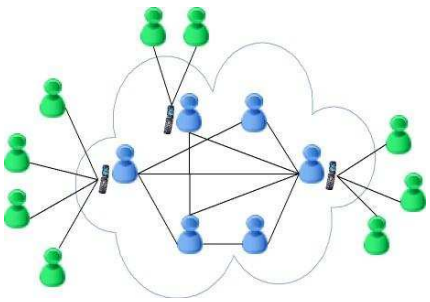


Figure 2. Phonebook-enabled social network

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 private contact in a phonebook may refer to the same person, they are handled separately. The main advantage of phonebook-enabled social networks 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.

$$E_{MM} \subseteq \{(u_M, u'_M) : u_M, u'_M \in U_M, u_M \neq u'_M\} \tag{3}$$

$$E_{MPc} \subseteq \{(u_M, u_{Pc}) : u_M \in U_M,\ u_{Pc} \in U_{Pc}\} \tag{4}$$

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

$$G_{PESN} = (U, E), \quad \text{where} \tag{5}$$

$$U = U_M \cup U_{Pc} \tag{6}$$

$$E = E_{MM} \cup E_{MPc} \tag{7}$$

3.3. 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, we refer to this as a *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. In addition to that this algorithm enables also to recognize duplications in phonebooks. We refer to such a social network

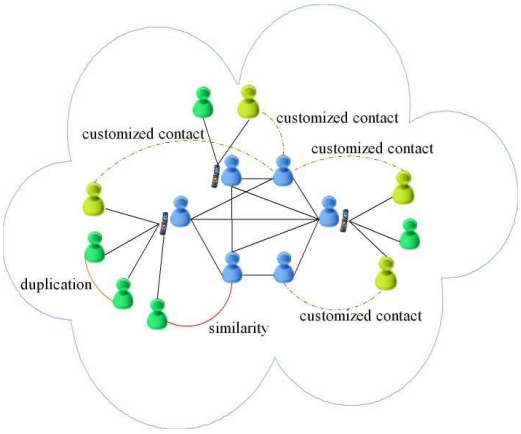


Figure 3. Phonebook-centric social network

with similarity and duplication detection algorithm as *phonebook-centric social network* (Figure 3).

The similarity detection creates the following two 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,

$$E_S = \{(u_M, u_{Pc}) : u_M \in U_M,\ u_{Pc} \in U_{Pc},\\ (u_M, u_{Pc}) \notin E_{MPc},\ \exists u'_M : (u'_M, u_M) \in E_{MM},\ (u'_M, u_{Pc}) \in E_{MPc}\} \quad (8)$$

$$E_D = \{(u_{Pc}, u'_{Pc}) : u_{Pc}, u'_{Pc} \in U_{Pc},\ u_{Pc} \neq u'_{Pc},\\ \exists((u_M, u_{Pc}), (u_M, u'_{Pc})) \in E_{MPc},\ u_M \in U_M\} \quad (9)$$

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. 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 mechanism as customization. The set of customized contacts is denoted by U_C.

After a similarity was resolved between a private contact and a member, a new customization edge will be created, which represents the connection between

the newly created customized contact and the original member. Later if a member changes some of her or his personal details (e.g. nickname), it will be automatically updated via the customization edge in the phonebook, which belongs to the relevant customized contact. However if the owner of the customized contact has already changed this personal detail, the change by the member will not have any affect. This customization mechanism enables that members can edit their phonebook, but their phonebooks can also get updates from the network.

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,

$$E_{MoC} = \{(u_{Mo}, u_C) : u_{Mo} \in U_M, u_C \in U_C, \\ \exists u'_M \in U_M, u'_M \neq u_{Mo}, (u_{Mo}, u'_M) \in E_{MM},\ (u_{Mo}, u_C) \in E_{MC}\} \quad (10)$$

$$E_{MC} = \{(u_M, u_C) : u_M \in U_M, u_C \in U_C, \\ \exists u'_M \in U_M, u'_M \neq u_M, (u_M, u'_M) \in E_{MM},\ (u_M, u_C) \in E_{MoC}\} \quad (11)$$

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

$$G_{PCSN} = (U, E),\ \text{where} \quad (12)$$

$$U = U_M \cup U_{Pc} \cup U_C \quad (13)$$

$$E = E_{MPc} \cup E_{MoC} \cup E_{MC} \cup E_D \cup E_S \quad (14)$$

The number of edges in E_{MC} is a key point phonebook-centric social networks from the scalability point of view because of the synchronization mechanism.

4. Phonebookmark

One of the key features of phonebook-centric social networks is that 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 ones phonebook also get the latest information about one 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 notifies 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. 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.

4.1. Main functions of Phonebookmark. *Phonebookmark* implements all phonebook-centric social network features. The similarity handling algorithm in [10] allows detecting 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.

Phonebookmark provides a semi-automatic similarity detecting and resolving mechanism. First it detects similarities and calculates similarity weight values that indicate how likely the two node of a similarity edge refer to the same person. *Phonebookmark* uses this weights also to determine the proper order of multiple similarities (Figure 4(a)).

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 4(b)).

(a) (b)

Figure 4. (a) Handling multiple similarities (b) Semi-automatic similarity resolution

During the operational period of *Phonebookmark*, the similarity detecting algorithm found about 1200 similarities and users have resolved more than 90 percent of these, which is an encouraging number for analyzing the distribution of similarities and propose a model for it. Following we refer to this rate as $P_R \approx 0.9$.

5. Distributions in phonebook-centric social networks

In this section we present empirical results about the distribution of the in- and out-degree, phonebook sizes, and the similarities in our network.

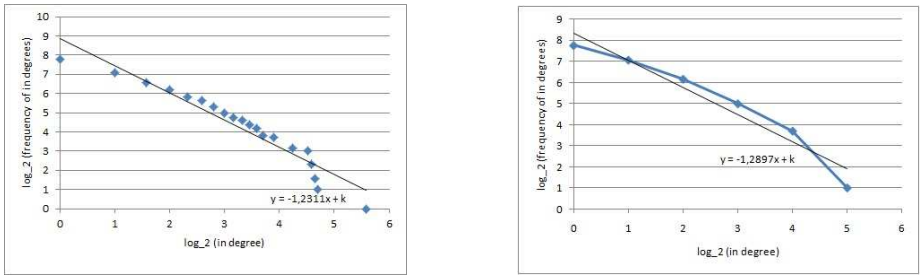
According to [17] a nonnegative random variable X is said to have a power law distribution if

$$\Pr[X \geq x] = cx^{-\alpha}, \tag{15}$$

for constants $c > 0$ and $\alpha > 0$. In a power law distribution asymptotically the tail falls according to the power α. If X has a power law distribution, then on 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 log-log function is the α parameter of the given power law distribution:

$$\ln(\Pr[X \geq x]) = -\alpha \ln(x) + \ln(c) \tag{16}$$

5.1. 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.



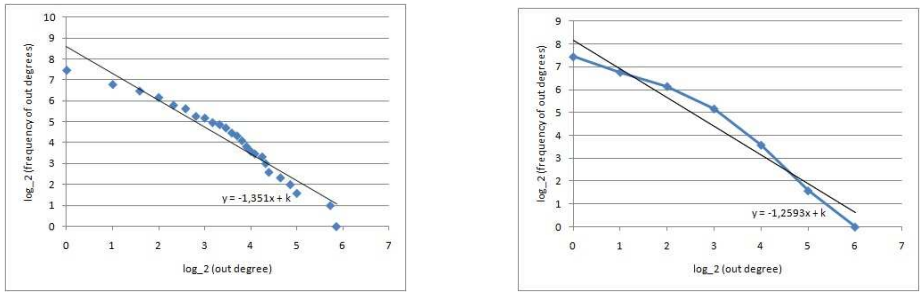
(a) (b)

Figure 5. Distribution of in-degree in *Phonebookmark*. (a) Without logarithmic binning. (b) With logarithmic binning procedure.

Figure 5(a) shows the distribution of in-degrees using logarithmically scaled scaled x- and y-axis. The x-axis represents the in-degree and the y-axis the number of members with at least this in-degree. Figure 5(a) also shows the line obtained by the least squares fitting technique. The slope of this line corresponds to the exponent of the power law distribution $\alpha = 1.2311$.

Figure 5(b) depicts the in-degree distribution using the logarithmic binning procedure (see [2], [1]). With this method we obtain an exponent $\alpha = 1.2897$. The exponents 1.2311 and 1.2897 obtained by the two methods are quite close to each other.

The out-degree – without and with the logarithmic binning procedure – of *Phonebookmark* is depicted on Figure 6.



(a) (b)

Figure 6. Distribution of out-degree in *Phonebookmark*. (a) Without logarithmic binning. (b) With logarithmic binning procedure.

The exponent of the power law we get is $\alpha = 1.351$ without binning and

$\alpha = 1.4034$ with binning. In several papers (e.g. [8]) a power law distribution is referred as:

$$\Pr[X = x] = c' x^{-\beta}. \tag{17}$$

(17) can be obtained by the derivation of one minus the right hand side of (15), where $\beta = \alpha + 1$ and $c' = \alpha \cdot c$, see for example [2], [1].

Similar degree distribution to our phonebook-centric social network was also reported in [6], where the distribution of the collaboration graph of movie actors follows a power law with $\beta = 2.3 \pm 1$ (i.e. $\alpha = 1.3 \pm 1$).

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 $\beta = 2.73$ for out-degree ($\alpha = 1.09$ and $\alpha = 1.73$ respectively).

5.2. Phonebook sizes. Figure 7 shows the tail distribution of the phonebook-sizes such that the x-axis has linear scale and the y-axis logarithmic scale. The points on this figure fit very well to a line, which means that the tail of the phonebook sizes decreases exponentially. This is a big contrast to the inverse polynomial tail of in- and out-degrees.

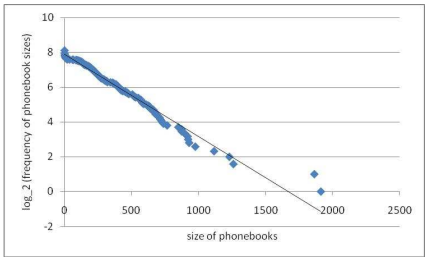


Figure 7. Size of phonebooks in *Phonebookmark*

5.3. Similarities. Based on the database and database logs of *Phonebookmark* we mamaged to measure the distribution of similarities raised by a member during registration and first phonebook synchronization. Figure 8 shows the distribution of similarities with the logarithmic binning procedure, where the x-axis represents the number of similarities and the y-axis means how many people causes at least that amount of similarities when registers and synchronizes. Again we use logarithmically scaled x- and y-axis. Figure 8 shows that the points are close to a line, thus the distribution of similarities can be well approximated by a power law. The exponent of the power law distribution is $\alpha = 1.276$ ($\beta = 2.276$).

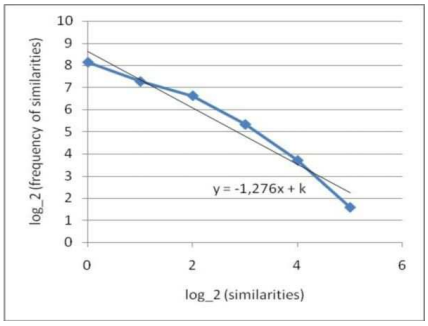


Figure 8. Distribution of similarities with logarithmic binning procedure

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

$$\Pr[X \geq x] = x^{-1.276}. \tag{18}$$

Let N_S be 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 $N_M \Pr[X \geq x] = N_M x^{-1.276}$, where N_M is the number of members in the network.

6. Estimating the total number of similarities

According to the empirical results of the previous section we model the number of similarities generated during a member registration by a probability variable X, which follows a power law. More preciselly, X models the number of similarities proposed by the automatic similarity detection algorithm. The expected number of (proposed) similarities is

$$E[X] = \sum_{x=1}^{\infty} x \Pr[X = x]. \tag{19}$$

Note that x starts from one, because a new member registration involves at least one similarity, as 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 resolved similarities N_S in a phonebook-centric social network can be calculated with the following formula:

$$N_S = N_M E[X] P_R, \text{ where} \tag{20}$$

$N_M = |U_M|$ is the number of registered members and $P_R \approx 0.9$ is the rate of the similarities resolved by the users, as described in Section 4. In order to calculate $E[X]$, we need the probabilities $\Pr[X = x]$, which can be obtained from (18) by derivation:

$$\Pr[X = x] = c' \frac{1}{x^\beta}, \tag{21}$$

where $\beta = \alpha + 1$. In order to be a probability distribution, $\sum_{x=1}^{\infty} c' x^{-\beta} = 1$. Thus,

$$c' = \frac{1}{\sum_{x=1}^{\infty} \frac{1}{x^\beta}} = \frac{1}{\varsigma(\beta)}, \tag{22}$$

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

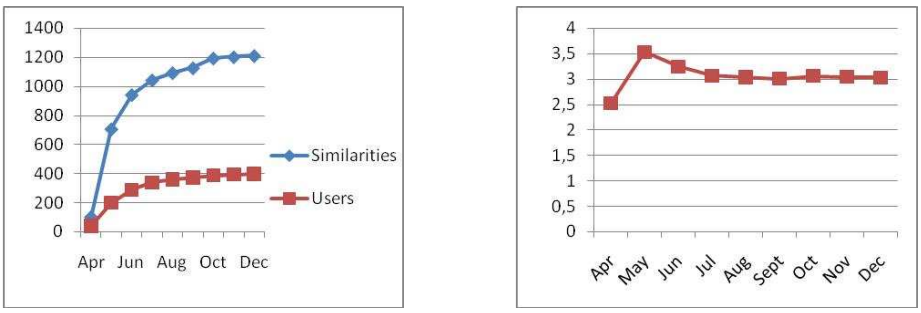
$$\begin{aligned} E[X] &= \sum_{x=1}^{\infty} x \Pr[X = x] = \sum_{x=1}^{\infty} x \frac{1}{\varsigma(\beta)} \frac{1}{x^\beta} \\ &= \frac{1}{\varsigma(\beta)} \sum_{x=1}^{\infty} \frac{1}{x^{\beta-1}} = \frac{\varsigma(\beta-1)}{\varsigma(\beta)}. \end{aligned} \tag{23}$$

Based on (20) the expected total number of similarities N_S in a phonebook-centric social network can be calculated with the following formula:

$$N_S = N_M \frac{\varsigma(\beta-1)}{\varsigma(\beta)} P_R. \tag{24}$$

For $\beta > 2$, $\varsigma(\beta-1)/\varsigma(\beta)$ is finite. In our case, for $\beta = 2.276$, we obtain that $\varsigma(\beta-1)/\varsigma(\beta) \approx 2.9196$. Therefore, the expected total number of similarities is $N_S = 2.9196 \cdot 420 \cdot 0.9 = 1103$.

6.1. Measured similarities. During the operation of *Phonebookmark* the trends of users and similarities was measured in [10]. These measurements empirically verify the previous model. Figure 9(a) and Figure 9(b) illustrate that although the number of users increases, the average number of resolved similarities per user remains on the same level.



(a) (b)

Figure 9. Trend of similarities and users in *Phonebookmark*. (a) Number of users and similarities from April to December 2008; (b) the average number of similarities per user

According to the previous model and the measurements, the total number of similarities increases linearly with the number of members, which is good news from the aspect of scalability. But how good is this estimation? Note that we used the data of the same phonebook-centric social network *Phonebookmark* for the measurements and to create the model. A power law distribution has an infinite variance, if $\alpha \leq 2$. How far could be the total number of similarities from the expected value? In the next section we use the fact that the number of similarities caused by a member cannot be arbitrarily high and prove an $O(N_M)$ upper bound on the total number of similarities, with high probability.

7. Upper bound on the total number of similarities

Now we prove that, if $\alpha > 1$ ($\beta > 2$), then the total number of similarities in a phonebook-centric social network is $O(N_M)$, with high probability. Let A be an event which depends on N_M. We say that A occurs with high probability (w.h.p.) if $\Pr[A] \geq 1 - N_M^{-\gamma}$ for an arbitrarily chosen constant $\gamma \geq 1$.

In our proof we assume, that the phonebooks do not contain duplicates (if a duplicate is detected, the users delete them immediately). Then the following holds:

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 higher similarities is zero. In order to

see this, note that a member u can be similar to at most one private contact of each of the other $N_M - 1$ members and, for each private contact of u, there is at most one similar member in the network.

Theorem 1. *Consider a phonebook-centric social network of N_M members, where, for each member u, the number of similarities caused by u is modeled by a random variable X_u with $\Pr[X_u = x] = c \cdot x^{-\beta}, 2 < \beta < 3$, if $x \in [1, aN_M]$, for a constant $a > 0$, and $\Pr[X_u = x] = 0$, otherwise. The random variables $\{X_u : u \in U\}$ are assumed to be independent. Let $S = \sum_{u \in U_M} X_u$. Then $S = O(N_M)$, w.h.p.*

For proving Theorem 1 we need the following three Lemmata.

Lemma 2. *Let $U_1 \subseteq U_M$ be the set of members causing at most $\left(\frac{N_M}{\ln N_M}\right)^{\frac{1}{\beta-1}}$ similarities, i.e. $U_1 = \left\{u \in U_M : 1 \leq X_u < \left(\frac{N_M}{\ln N_M}\right)^{\frac{1}{\beta-1}}\right\}$. Let $S_1 = \sum_{u \in U_1} X_u$. Then, w.h.p., $S_1 \leq c_1 \cdot N_M$, for an appropriate constant c_1.*

Proof. We divide the members of U_1 into $O(\ln N_M)$ groups according to the number similarities caused by them and we give an upper bound on the number of users in each group, w.h.p. Then we derive an upper bound on the number of similarities in the groups, w.h.p.

Let B_i be the set of members causing at least 2^i and less than 2^{i+1} similarities, $0 \leq i < \frac{1}{\beta-1} \log_2 \left(\frac{N_M}{\ln N_M}\right)$. Let p_i be the probability that a member u is in B_i, i.e. $p_i = \Pr[2^i \leq X_u < 2^{i+1}]$. Then $p_i = \sum_{2^i \leq j < 2^{i+1}} cj^{-\beta}$ and it can be upper bounded as $p_i \leq \sum_{2^i \leq j < 2^{i+1}} c2^{-i\beta} = 2^i c2^{-i\beta} = c2^{-i(\beta-1)}$ and lower bounded as $p_i > \sum_{2^i \leq j < 2^{i+1}} c2^{-(i+1)\beta} = 2^{-\beta} c2^{-i(\beta-1)}$. Thus $p_i = c_i 2^{-i(\beta-1)}$, for an appropriate constant c_i, $2^{-\beta}c < c_i \leq c$.

We introduce a Bernoulli variable $Y_{u,i}$ for the event that u is in B_i, i.e. $\Pr[Y_{u,i} = 1] = p_i$ and $\Pr[Y_{u,i} = 0] = 1 - p_i$. Let $Y_i = \sum_{u \in U_M} Y_{u,i}$ and $m_i = E[Y_i] = N_M p_i$. Then by applying the Chernoff inequality (see e.g. [15]), we obtain that, for $\epsilon > 0$,

$$\Pr\left[Y_i \geq (1+\varepsilon)m_i\right] \leq e^{-\frac{\min(\varepsilon, \varepsilon^2)}{3} m_i}. \tag{25}$$

From (25) we obtain a probability of at most $N_M^{-\gamma'}$, for an arbitrary constant $\gamma' > 0$, by setting $\epsilon = \max\left(3\gamma' \frac{\ln N_M}{m_i}, \sqrt{3\gamma' \frac{\ln N_M}{m_i}}\right)$. Therefore, w.h.p., the number of similarities caused by users in $B(i)$ is at most

$$\begin{aligned} 2^{i+1}(1+\epsilon)m_i &= 2^{i+1}(1+\epsilon)N_M p_i = 2^{i+1}(1+\epsilon)N_M c_i 2^{-i(\beta-1)} \\ &= (1+\epsilon)2c_i 2^{-i(\beta-2)} N_M. \end{aligned} \tag{26}$$

Since $m_i = N_M p_i = N_M c_i 2^{-i(\beta-1)} > N_M c_i \frac{\ln N_M}{N_M} = c_i \ln N_M$, the value of ϵ is at most $\max(1, 3\frac{\gamma'}{c_i})$, which is a constant. Thus, w.h.p., the number of similarities caused by users in $B(i)$ is at most $c_i' 2^{-i(\beta-2)} N_M$, where $c_i' = (1+\max(1, 3\frac{\gamma'}{c_i}))2c_i$ is a constant. Let $c' = \max_i c_i'$. Then the number of similarities caused by members of U_1 is at most

$$N_M c' \sum_{0 \le i < \frac{1}{\beta-1} \log_2 \frac{N_M}{\ln N_M}} 2^{-i(\beta-2)} \le N_M c' \frac{1}{1-2^{2-\beta}} = O(N_M) \tag{27}$$

with probability at least $1 - O(N_M^{-\gamma'} \ln N_M)$. By setting $\gamma = \gamma' - \epsilon$, for an arbitrarily small constant $\epsilon > 0$, this probability is at least $1 - O(N_M^{-\gamma})$. □

Lemma 3. *Let $U_2 \subseteq U_M$ be the set of members causing at least $\left(\frac{N_M}{\ln N_M}\right)^{\frac{1}{\beta-1}}$ and less than $N_M^{\frac{1}{\beta-1}+\delta}$ similarities, where $0 < \delta < 1 - \frac{1}{\beta-1}$ is a constant, i.e. $U_2 = \left\{u \in U_M : \left(\frac{N_M}{\ln N_M}\right)^{\frac{1}{\beta-1}} \le X_u < N_M^{\frac{1}{\beta-1}+\delta}\right\}$. Let $S_2 = \sum_{u \in U_2} X_u$. Then, w.h.p., $S_2 < c_2 \cdot N_M$ for an appropriate constant c_2.*

Proof. We use the fact that the probability that a particular member causes at least j and less than k similarities, $j, k \in \mathbb{N}$, $1 < j < k \le aN_M + 1$, can be well approximated by $\int_j^k cx^{-\beta} dx$, i.e. $\Pr[j \le X_u < k] = \sum_{j \le i < k} ci^{-\beta}$ is at least $c\int_j^k x^{-\beta} dx$ and at most $c\int_{j-1}^{k-1} x^{-\beta} dx \le c((j-1)^{-\beta} + \int_j^k x^{-\beta} dx) = c' j^{-(\beta-1)}$, for an appropriate constant c'. Therefore, the probability that a particular member causes at least $\left(\frac{N_M}{\ln N_M}\right)^{\frac{1}{\beta-1}}$ similarities is $p = c' \frac{\ln N_M}{N_M}$.

For each member $u \in U_M$, we introduce a Bernoulli variable Y_u, such that $Y_u = 1$, if $X_u \ge \left(\frac{N_M}{\ln N_M}\right)^{\frac{1}{\beta-1}}$ and $Y_u = 0$, otherwise. Thus, $\Pr[Y_u = 1] = p$ and $\Pr[Y_u = 0] = 1 - p$. Let $Y = \sum_{u \in U_M} Y_u$ and $m = E[Y] = c' \ln N_M$. By setting $\epsilon = \max(1, \frac{3\gamma}{c'})$ and applying the Chernoff inequality,

$$\Pr\left[Y \ge (1+\epsilon)m\right] \le e^{-\frac{\min(\epsilon, \epsilon^2)}{3} m}, \tag{28}$$

we obtain that $\Pr[Y \ge (1+\epsilon)c' \ln N_M] \le N_M^{-\gamma}$. Since each member u in U_2 causes at most $N_M^{\frac{1}{\beta-1}+\delta}$ similarities, the total number of similarities caused by members in U_2 is at most $N_M^{\frac{1}{\beta-1}+\delta}(1+\epsilon)c \ln N_M = O(N_M)$, w.h.p. □

Lemma 4. *Let $U_3 \subseteq U_M$ be the set of members causing at least $N_M^{\frac{1}{\beta-1}+\delta}$ similarities, where $0 < \delta < 1 - \frac{1}{\beta-1}$ is a constant, i.e. $U_3 = \left\{u \in U_M : N_M^{\frac{1}{\beta-1}+\delta} \le X_u\right\}$. Let $S_3 = \sum_{u \in U_3} X_u$. Then, w.h.p., $S_3 < c_3 \cdot N_M$ for an appropriate constant c_3.*

PROOF. Again we use the fact that probability that a particular member causes at least j similarities, $1 < j \leq aN_M$, is $c'j^{-(\beta-1)}$, for an appropriate constant c'. The probability that a particular member causes at least $N_M^{\frac{1}{\beta-1}+\delta}$ similarities is $c'\frac{1}{N_M^{1+\delta(\beta-1)}}$. Let p_k be the probability that at least k members cause at least $N_M^{\frac{1}{\beta-1}+\delta}$ similarities. Then

$$p_k \leq \binom{N_M}{k}\left(c'\frac{1}{N_M^{1+\delta(\beta-1)}}\right)^k \leq \frac{\left(c'\frac{N_M}{N_M^{1+\delta(\beta-1)}}\right)^k}{k!} = \frac{\left(c'N_M^{-\delta(\beta-1)}\right)^k}{k!}. \tag{29}$$

For $k = \frac{\gamma}{\delta(\beta-1)}$, the probability p_k is at most $c'^k N_M^{-\gamma}$. Since each of these k users causes at most $a \cdot N_M$ similarities (in our phonebook-centric social network at most $2(N_M - 1)$ similarities), we obtain that $S_3 \leq a \cdot k \cdot N_M = O(N_M)$, w.h.p. □

PROOF. (Theorem 1) Let S_1, S_2, S_3 and c_1, c_2, c_3 be defined as in the previous three lemmata. Then by these lemmata we obtain that the probability that the total the number of similarities $S = S_1 + S_2 + S_3$ is greater than $\max(c_1, c_2, c_3) \cdot N_M$ is at most $3 \cdot O(N_M^{-\gamma}) = O(N_M^{-\gamma})$. □

Remarks 1. This theorem can be applied for several dynamically evolving self-organizing networks, where a power law distribution with an exponent $1 < \alpha < 2$ has been observed. For example, in our social network, each member has another member at most once as contact. Then the maximum degree is bounded by the number N_M of members. Thus, the overall number of contacts is $O(N_M)$, w.h.p.

Another example is the Web graph [8]. The node degree follows a power law distribution with an exponent $1 < \alpha < 2$ and the maximum node degree is bounded by the number N of web pages. Thus, the overall number of edges in the Web graph is bounded by $O(N)$, w.h.p.

8. Conclusion

In this paper we have investigated the structure of phonebook-centric social networks. We have analyzed the data of *Phonebookmark*, which is a phonebook-centric social network implementation. Based on measurements we have experienced that the distribution of in- and out-degrees as well as the distribution of the similarities follow a power law. In contrast to this, we have found that the tail of the distribution of phonebook sizes decreases exponentially.

As a main contribution, we have shown that the total number of similarities in phonebook-centric social networks depends linearly on the number of registered members N_M. We validated this result also with measurements. Finally, we have proved an $O(N_M)$ upper bound on the total number of similarities in the network, with high probability. This is good news, since the scalability of the system mainly depend on the total number of similarities. We have shown, that this upper bound can also be applied for several dynamically evolving self-organizing networks, where a power law distribution with an exponent $1 < \alpha < 2$ has been observed.

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 and B. A. Huberman, Zipf's law and the Internet, *Glottometrics* **3** (2002), 143–150.

[3] W. Aiello, F. R. K. Chung and L. Lu, Heavy–Tailed Probability Distributions in the World Wide Web, Proc. of STOC, 2000, 171–180.

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

[5] A.-L. Barabási, Linked: How Everithing is Connectet to Everything Else, *Perseus Publishing*, 2002.

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

[7] B. Bollobás, O. Riordan, J. Spencer and G. Tusnády, The degree sequence of a scale-free random graph process, *Random Structures and Algorithms* **18**, no. 3 (2001), 279–290.

[8] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, Graph structure in the web, *Computer Networks* **33**, no. 1–6 (2000), 309–320.

[9] M. E. Crovella, M. S. Taqqu and A. Bestavros, Heavy–Tailed probability distributions in the World Wide Web, *A Practical Guide To Heavy Tails* **1** (1998), 3–25.

[10] P. Ekler, Z. Iványi and K. Aczél, Similarity Management in Phonebook-centric Social Networks, Proc. 4th International Conference on Internet and Web Applications and Services (ICIW 2009), 2009, 273–279.

[11] P. Ekler, I. Kelényi, I. Dévai, B. Bakos and A. J. Kiss, Efficient mobile content sharing with local cooperation support, *Journal of Networks* **2** (2009), 119–132.

[12] P. Ekler, J. K. Nurminen and A. J. Kiss, Experiences of implementing BitTorrent on Java ME platform, Proc. 1st IEEE International Peer-to-Peer for Handheld Devices Workshop (CCNC), 2008, 1154–1158.

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

[14] Facebook, Social networking application, `http://www.facebook.com`, May, 2009.

[15] T. Hagerup and C. Rüb, Guided tour of Chernoff bounds, *Inf. Process. Lett.* **33**(6) (1990), 305–308.

[16] I. Kelényi, G. Csúcs, B. Forstner and H. Charaf, Peer-to-Peer File Sharing for Mobile Devices, In Mobile Phone Programming: Application to Wireless Networks, volume of 2 Intelligent Systems at the Service of Mankind, (F. Fitzek, F. Reichert, eds.), 2007, 311–324.

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

[18] MySpace, Social networking application, `http://www.myspace.com`, May 2009.

[19] V. Pareto, Course d'economie politique professé a l'université de Lausanne, 3 volumes, 1897.

[20] Nazir S. Raza and C.-N. Chuah, Unveiling Facebook: A Measurement Study of Social Network Based Applications, Proc. 8th ACM SIGCOMM Conference on Internet Measurement (IMC), 2008, 43–56.

[21] M. Ripeanu, I. Foster and A. Iamnitch, Mapping the Gnutella network: properties of large-scale peer-to-peer systems and implications for system design, Special issue on peer-to-peer networking, *IEEE Internet Computing Journal* **6**(1) (2002), 50–57.

[22] G. U. Yule, Statistical Study of Literary Vocabulary, *Cambridge University Press*, 1944.

[23] G. K. Zipf, The Psycho-Biology of Language, *Houghton Mifflin, Boston, MA*, 1935.

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

PÉTER EKLER
DEPARTMENT OF AUTOMATION
AND APPLIED INFORMATICS
BUDAPEST UNIVERSITY
OF TECHNOLOGY AND ECONOMICS
MAGYAR TUDÓSOK KÖRÚTJA 2
1113 BUDAPEST
HUNGARY

E-mail: peter.ekler@aut.bme.hu

TAMÁS LUKOVSZKI
FACULTY OF INFORMATICS
EÖTVÖS LORÁND UNIVERSITY
PÁZMÁNY PÉTER SÉTÁNY 1/C
1117 BUDAPEST
HUNGARY

E-mail: lukovszki@inf.elte.hu

(Received June 18, 2009; revised January 17, 2012)