The Analysis of Similarities and Registration Delay 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*—Phonebook centric social networks provide a synchronization mechanism between phonebooks of the users and the social network which allows detecting other users listed in the phonebooks. After that, if one of their contacts changes her or his personal detail, it will be propagated automatically into the phonebooks, after considering privacy settings. We participated in the implementation of a phonebook centric social network, called Phonebookmark and investigated the structure of the network. We used the data of this network for building the proposed models. In such social networks two entities may identify the same person if some parameters are similar, e.g.: phone number, address, etc. We call such entity pairs as similarities. Previously it was shown that the distribution of similarities follows a power law. Also a model was proposed by us, which can be used to estimate the total number of similarities, which is very important from scalability point of view in such networks. However the accuracy of the model is another question, because of the infinite variance of the power law distribution, which is used for modeling the number of similarities involving a user. The paper presents interesting and practical problem of analysis of similarities in social networks with application to mobile phonebooks management. The presented contribution includes both theoretical and practical components as well. We show that using the fact that a member of the network can only be involved in a limited number of similarities results in a similarity distribution with a finite variance. By using the central limit theorem we show the accuracy of our estimation. We also highlight that this model can be used in other power law distributions which apply to the requirements. Finally we also propose a performance model which can be used during the resource requirement design of such phonebook centric social networks.**

Keywords-component; social networks, mobile phones, power law distribution, variance, central limit theorem, queue model

I. INTRODUCTION

Nowadays social based websites, like social networks are becoming increasingly popular. These solutions not only available from web browsers but there are several existing mobile clients as well. These mobile applications are mainly simple clients to the network with some additional features. The phonebooks in the mobile devices represent social relationships that can be integrated in the social networks.

The relationship between social networks and mobile phones is noticeable as the popularity of such systems increase. In [1] we analyzed such networks from similarity handling point of view. In this paper we extend those results with important models and performance evaluation. Our new model allows a much more accurate prediction about the scalability of networks where the connections follow power law distributions.

In the last decade the internet related technologies developed rapidly. As reasons of this growth new type of solutions and applications have appeared. One of the most popular solutions are social network sites (SNS). Since their introduction, social network sites such as Facebook, MySpace and LinkedIn have attracted millions of users, many of whom have integrated these sites into their daily practices and they even visit these multiple times per day. These popular online social networks are among the top ten visited websites on the Internet [2]. End of 2010 it was reported that Facebook surpasses Google as number one U.S. site [3]. The basic idea behind such networks is that users can manage personal relationships online on these networks.

According to new statistics [4] Facebook has more than 750 million users, 50% of the active users log on to Facebook in any given day, more than 35 million users update their status each day and an average user spends more than 55 minutes per day on Facebook. Facebook began in early 2004 and the above statistics show that such popular social networks can have a huge growth which has to be considered during the design of any SNS.

Mobile phones and mobile applications are another hot topic nowadays. Facebook statistics also show that there are more than 65 million active users currently accessing Facebook through their mobile devices. People that use Facebook on their mobile devices are almost 50% more active on Facebook than non-mobile users. The increasing capabilities of mobile devices allow them to participate in social network applications as well. Mobile phone support in general social networks are usually limited mainly to photo and video upload capabilities and access to the social network using the mobile web browser.

However we should consider the fact, that 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 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 our 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. In the rest of this paper we refer to this solution as a *phonebook centric social network* (PCSN). Discovering and handling such similarities in phonebook centric social networks is a key issue. If a member changes some of her or his detail, it should be propagated in every phonebook to which she or he is related after considering privacy settings. In addition to that, with the help of detected similarities the system can keep the phonebooks always up-to-date.

Power law distribution is quite common in social networks and similar internet related graphs as measurements and examples show in Section 2. The number of similarities in phonebook centric social networks is very important from performance and scalability point of view.

We show that the distribution of similarities can be modeled with a random variable X with $\Pr[X \geq x] \sim cx^{-\alpha}$, if $x \leq n$ and $\Pr[X \geq x] = 0$ otherwise, where α>1 and n is a relevant upper bound.

As a main contribution of this paper, we show that the distribution of similarities has a finite variance which allows us to use the central limit theorem to prove the accuracy of our estimation of the total number of similarities. This model can be used generally in other similar distributions.

As a practical result, the concept of phonebook centric social networks was applied in the *Phonebookmark* project at Nokia Siemens Networks. *Phonebookmark* is a phonebook centric social network implementation by Nokia Siemens Networks. We took part in the implementation and 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 suitable number for analyzing the behavior of the network. During this period we have collected and measured different type of data related to the social network.

The rest of the paper is organized as follows. Section 2 describes related work in the field of social networks and power law distributions. Section 3 introduces the structure of phonebook centric social networks. Section 4 summarizes our previously published model related to calculating the total number of similarities in the network. Section 5 states a general theorem related to the variance of power law distribution with relevant upper bound and uses it to prove the accuracy of the model described in Section 4. Section 6 shows that the total number of similarities is close to their expected value. Section 7 shows a performance model for calculating the expected queue length for similarity processing. We also show measurements related to Phonebookmark based on this model. The model can be used during the design of any different phonebook centric social networks. Finally, Section 8 concludes the paper and proposes further research plans.

II. Related Work

In [5] the authors have defined social network sites (SNSs) as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site.

According to this definition, the first recognizable social network site launched in 1997. SixDegrees.com allowed users to create profiles, list their Friends and, beginning in 1998, surf the Friends lists. Each of these features existed in some form before SixDegrees, of course. Profiles existed on most major dating sites and many community sites. AIM and ICQ buddy lists supported lists of Friends, although those Friends were not visible to others. Classmates.com allowed people to affiliate with their high school or college and surf the network for others who were also affiliated, but users could not create profiles or list Friends until years later. SixDegrees was the first to combine these features.

After that social networks have developed rapidly and the number of features increased. Nowadays most sites support the maintenance of pre-existing social networks, but others help strangers connect based on shared interests, political views, or activities. Some sites cater to diverse audiences, while others attract people based on common language or shared racial, sexual, religious, or nationality-based identities. Sites also vary in the extent to which they incorporate new information and communication tools, such as mobile connectivity, blogging, and photo/video-sharing.

As the functions of the SNSs flared, the number of users increased rapidly. Handling the extending number of users efficiently in SNSs is a key issue as it was visible in case of Friendster. Friendster was launched in 2002 as a social complement to Ryze. It was designed to help friends-of-friends meet, based on the assumption that friends-of-friends would make better romantic partners than would strangers. As Friendster's popularity surged, the site encountered technical and social difficulties. Friendster's servers and databases were ill-equipped to handle its rapid growth, and the site faltered regularly, frustrating users who replaced email with Friendster.

Huge amount of papers and popular books, such as Barabási's Linked [6] 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 [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 α>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 [8] (see. [9]), while describing the distribution of income in the population. In

1935 Zipf [10] and Yule [11] 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*.

Mislove et al. [12] studied the graph properties of several online real-world social networks. Their paper presents a large-scale measurement study and analysis of the structure of multiple online social networks. They examined data gathered from four popular online social networks: Flickr, YouTube, LiveJournal, and Orkut. They crawled the publicly accessible user links on each site, obtaining a large portion of each social network's graph. Their data set contains over 11.3 million users and 328 million links. Their measurements show that high link symmetry implies indegree equals outdegree; users tend to receive as many links as the give, the observed networks are power law with high symmetry.

In [13], the graph structure of the Web has been investigated which also can be considered as a special variant of social network [14] 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. [15] showed that the in-and out-degree distribution of the interaction graph of the studied MySpace applications also follow such distributions.

There has been a great deal of theoretical work on designing random graph models that result in a Web-like graph. Barabási and Albert [16] 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. [17] 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. [18]. Mitzenmacher [19] gives an excellent survey on the history and generative models for power law distributions. Aiello et al. [20] studies random graphs with power law degree distribution and derives interesting structural properties in such graphs.

Detecting similar or matching parameters of users is an important part in our phonebook centric social network. There is a huge amount of work for general similarity and match detection algorithms. According to [21] typical systems have an effectiveness (accuracy) of, at best, forty percent. A new measurement [22] showed that 55 percent of the first 20 records retrieved by Google Scholar are relevant. As shown in 3.3, the precision of Google Scholar remains relatively high even after the first 50 hits. Within the first 100 search results, 39 percent of GS records are relevant. Figure 3.3 also reveals that the utility of GS could be improved if relevant results were concentrated more heavily within the first 20 or 30 hits rather than the first 50 or 100.

The key difference between other works on online social networks and our work is that we extended social networks with mobile phone support and we discovered that the distribution of similarities follows power law. We proposed a model to estimate the number of similarities and despite the infinite variance of power law distribution we proved the accuracy of our model.

III. Structure of Phonebook centric Social Networks

Phonebook centric social networks are extending the well-known social network sites, they have a similar web user interface, but they add several major mobile phone related functions to the system. Following consider social networks as graphs. In case of general social networks, nodes are representing registered members and edges between them represent social relationships (e.g. friendship). After this we should notice that each member has a private mobile phone with a phonebook (Figure 1).

On Figure 1 we can see that phonebook contacts results new type of nodes in the graph representation and the edges between these private phonebook contacts and members represent which member "owns" those private contacts.

One of the key advantages of phonebook centric social networks is that they allow real synchronization between private phonebook contacts and the social network.

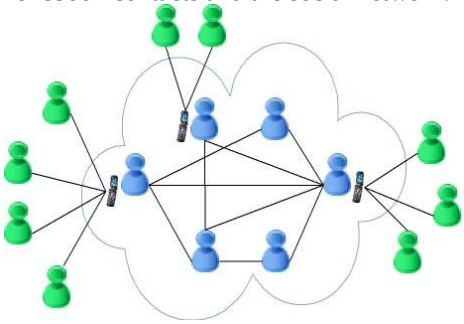

Figure 1. Phonebook-enabled social network

In order to enable such mechanism we need a similarity detecting algorithm. Such an algorithm is able to compare two person entries (members and private contacts, too) and determine how likely they represent the same person and propose a corresponding weight for the detected similarity.

Figure 2 represents the graph structure if the similarity detecting algorithm has finished comparing the relevant person entries.

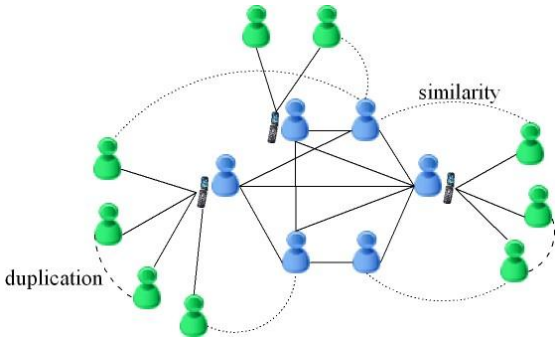

Figure 2. Detected similarities and duplications

On Figure 2 the dotted edges between member 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 similarities and duplications are detected there is a semi-automatic step, the members having private contacts in their phonebook, which are detected as similar to other members, have to decide whether the detected similarities are relevant ones, i.e.: accept or reject them. We call this step similarity resolution. In addition to that, members can also decide about the relevancy of detected duplications in their phonebooks. Figure 3 represents the graph structure after some of the members have resolved the detected similarities and duplication.

Besides that we can see on Figure 3 that four from the five similarities were accepted and there is still one in the system (the member has not checked it yet). Accepting a similarity means that a customized link edge is being formed between the private contact(s) in one's phonebook and the relevant member who represent the same person in the system. The private contacts that are linked to members via this type of customized links are called *customized contacts*.

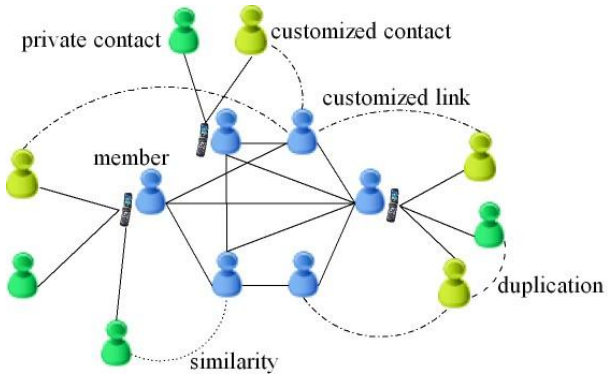

Figure 3. Resolving similarities and duplications

One of the key advantages of phonebook centric social networks are these customized links, because if a member changes 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. Additional important advantages of phonebook centric social networks are:

- Private contacts can be managed (list, view, edit, call, etc.) from a browser.
- Similarity detecting algorithm realizes the user if duplicate contacts are detected in its phonebook 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 detailed structure and edge rule definition was described in [23].

In [24] we have introduced a phonebook-centric social network implementation, called Phonebookmark. Phonebookmark provides a semi-automatic similarity detecting and resolving mechanism. First it detects similarities and calculates a similarity weight for them, which indicates, how likely the entries identify the same person. (Figure 4).

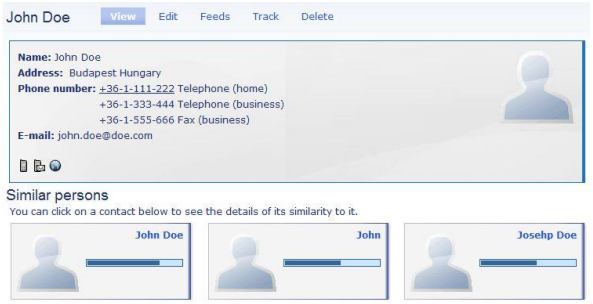

Figure 4. Dealing with multiple similarities

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 accept or reject the similarity, which is the base of the semi-automatic behavior (Figure 5).

Figure 5. Semi-automatic similarity resolution

IV. NUMBER OF SIMILARITIES

We model the number of similarities generated during a member registration by a random variable X. More precisely, X models the number of similarities proposed by the automatic similarity detection algorithm. In [22] we showed that X can be well approximated by a power law distribution. Using this model we gave estimation on the total number of similarities in the system. Now we summarize this model.

The total number of accepted similarities N_S in a phonebook centric social network can be estimated with the following formula:

$$N_S = NE[X]P_R, \quad (3)$$

where N is the number of registered members and P_R is the rate of the similarities accepted by the users. Measurements in [22] showed that P_R can be approximated with 0.9. In order to estimate $E[X]$, we need the probabilities $Pr[X=x]$, which can be obtained from the complementary cumulative distribution function $\Pr[X \geq x] \sim cx^{-\alpha}$ by derivation:

$$\Pr[X = x] \sim c' x^{-(\alpha+1)}. \quad (4)$$

In order to be a probability distribution, $\sum_{x=1}^{\infty} c' x^{-(\alpha+1)} = 1$. 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. Thus, $c' = 1/\zeta(\alpha+1)$, where $\zeta(.)$ denotes the Riemann Zeta function. Then the expected value is:

$$\begin{aligned} E[X] &= \sum_{x=1}^{\infty} x \Pr[X = x] \\ &= \sum_{x=1}^{\infty} x \frac{1}{\zeta(\alpha+1)} x^{-(\alpha+1)} \\ &= \frac{1}{\zeta(\alpha+1)} \sum_{x=1}^{\infty} x^{-\alpha} = \frac{\zeta(\alpha)}{\zeta(\alpha+1)}. \end{aligned} \quad (5)$$

The expected total number of accepted similarities N_S in a phonebook centric social network can be estimated with the following formula:

$$N_S = N_M \frac{\zeta(\alpha)}{\zeta(\alpha+1)} P_R. \quad (6)$$

For $\alpha>1$, $\zeta(\alpha)/\zeta(\alpha+1)$ is a finite constant. In our case, for $\alpha=1.276$, we obtain that the expected total number of similarities is

$$N_S = 2.9196 * 420 * 0.9 = 1103. \quad (7)$$

However in this model the X random variable has power law distribution which has infinite variance thus the accuracy of this model is an issue. In the next section we show how to prove the accuracy of this model by stating and a general theorem related to the variance of power law distributions with relevant upper bound.

V. Variance Model for Power Law Distribution with Upper Bound

For $\alpha \leq 2$, a power law distribution has infinite variance, which prevents to apply the central limit theorem in order to obtain that the total number of similarities will be close to their expected value. However we can use the following fact

Fact: If the phonebooks do not contain duplicates then the number of similarities caused by a member is at most $2(N-1)$ [23].

With other words, in the interval $[0,2(N-1)]$ 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 $N-1$ members and, for each private contact of u, there is at most one similar member in the network.

We show that the distribution of similarities resulting from this fact has a finite variance. This allows us to use the central limit theorem to prove the accuracy of our estimation of the total number of similarities in Section 4.

Theorem 1: Let X be a random variable with $\Pr[X = x] = cx^{-\beta}$ if $x \leq n$ and $\Pr[X = x] = 0$ otherwise, where $\beta = \alpha+1$, $2<\beta<3$. In this case the variance can be estimated with $\sigma^2 X = \Theta\left(n^{3-\beta}\right)$.

For the proof we used two lemmata.

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

Proof: From the Steiner formula, the variance is estimated as $\sigma^2 X = E[X^2] - (E[X])^2$. $E[X]$ was defined previously, thus we need to estimate only the $E[X^2]$. By definition:

$$\begin{aligned} E[X^2] &= \sum_{x=1}^{\infty} x^2 \Pr[X = x] \\ &= \sum_{x=1}^{n} x^2 c \frac{1}{x^{\beta}}. \end{aligned} \quad (8)$$

Now we can apply that n is an upper bound on the value of X. This way (1) can be followed as:

$$\begin{aligned} &E[X^2] = c \sum_{x=1}^{n} x^{2-\beta}. \\ &\text{Let } y = \frac{1}{c} E[X^2]. \end{aligned} \quad (9)$$

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

$$z^{2i} \Pr[X = z^i] = (z^i)^{2-\beta} = \left(2^{\frac{1}{\beta-2} i}\right)^{2-\beta} = 2^{-i} \quad (10)$$

Figure 6 illustrates how we performed the estimation, with the $f1$ function.

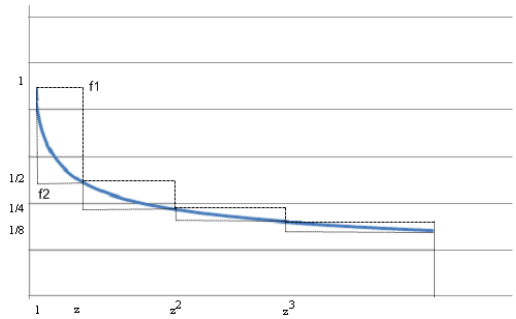

Figure 6. Staged estimation function

Now we are able to approximate *y* from top:

$$
\begin{aligned}
y &\le \sum_{i=0}^{\log_z n} \left(z^{i+1} - z^i\right)\left(z^i\right)^{2-\beta} \\
&= \sum_{i=0}^{\log_z n} \left(z^{i+1} - z^i\right)2^{-i} \\
&= \sum_{i=0}^{\log_z n} (z-1)z^i 2^{-i} \\
&= (z-1)\sum_{i=0}^{\log_z n} \left(\frac{z}{2}\right)^i \\
&= (z-1)\left(\frac{\left(\frac{z}{2}\right)^{\log_z n+1} - 1}{\frac{z}{2}-1}\right) \\
&= (z-1)\left(\frac{\frac{z}{2}n^{\frac{1}{\log_{z/2} z}} - 1}{\frac{z}{2}-1}\right) \\
&= (z-1)\left(\frac{\frac{z}{2}n^{\frac{1}{1+\log_{z/2} 2}} - 1}{\frac{z}{2}-1}\right).
\end{aligned}
\tag{12}
$$

The explanation to the last step:

$$\log_{z/2} z = \log_{z/2} 2\frac{z}{2} = 1 + \log_{z/2} 2 \tag{13}$$

To continue, first we have to check the following calculation. Remember that *z* was described with β and $\beta=\alpha+1$. This way:

$$\log_{z/2} 2 = \frac{\log_2 2}{\log_2 z/2} = \frac{1}{\log_2\left(\frac{2^{\frac{1}{\beta-2}}}{2}\right)} = \frac{1}{\frac{1}{\beta-2}-1} = \frac{\beta-2}{3-\beta}. \tag{14}$$

Therefore:

$$n^{\frac{1}{1+\log_{z/2} 2}} = n^{\frac{1}{1+\frac{\beta-2}{3-\beta}}} = n^{3-\beta}. \tag{15}$$

This way (2) looks as follows:

$$y \le (z-1)\left(\frac{\frac{z}{2}n^{\beta-3}-1}{\frac{z}{2}-1}\right). \tag{16}$$

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

$$
\begin{aligned}
\sigma^2 X &= E[X^2] - (E[X])^2 = cy - \Theta(1) \\
&= c(z-1)\left(\frac{\frac{z}{2}n^{\beta-3}-1}{\frac{z}{2}-1}\right) - \Theta(1) \\
&\le c\left(\frac{(z-1)z}{z-2}n^{3-\beta}\right) - \Theta(1) \\
&= O\left(n^{3-\beta}\right).
\end{aligned}
\tag{17}
$$

□

Lemma 2: Let *X* be a random variable with $\Pr[X=x] = cx^{-\beta}$ if $x \le n$ and $\Pr[X=x] = 0$ otherwise, where $\beta=\alpha+1$, $2<\beta<3$. In this case the variance is $\sigma^2 X = \Omega\left(n^{3-\beta}\right)$.

Proof: Similarly to *Lemma 1* if we give a lower bound on *y* using function *f2* is shown on Figure 6:

$$y \ge \sum_{i=0}^{\log_z n} \left(z^{i+1} - z^i\right)2^{-(i+1)}, \tag{18}$$

we obtain that $\sigma^2 X = \Omega\left(n^{3-\beta}\right)$. □

Proof (of Theorem 1) The proof is straightforward by applying Lemma 1 and 2:

$\sigma^2 X = \Theta\left(n^{3-\beta}\right)$, because $\sigma^2 X = \Omega\left(n^{3-\beta}\right)$ and $\sigma^2 X = O\left(n^{3-\beta}\right)$

□

In our case the upper bound *n* to the total number of similarities is *2(N-1)*.

VI. Applying Central Limit Theorem for the Distribution of Similarities

Following we show that the total number of similarities are close to their expected value.

Theorem 2: Let N be the number of members in a phonebook centric social network and $S_N=X_1+X_2+...+X_N$ where X_i , $i=1,...,N$, is a random variable representing the number of similarities raised by member i, i.e. $\Pr[X_i = x] = \kappa x^{-\beta}$ if $x \le n$ and $\Pr[X_i = x] = 0$ otherwise, where $n=\Theta(N)$, $\beta=\alpha+1$, $2<\beta<3$, and κ is a constant. Let $\mu = E[X_i]$. Then

$$\Pr[S_N \ge c\mu N] \approx 1-\Phi(m),$$

where $m = a(c-1)\mu N^{\frac{\beta}{2}-1}$ and a is an appropriate constant.

Proof: $X_1, X_2, \ldots, X_N$ are N independent and identically distributed random variables, each having finite values of expectation μ and variance $\sigma^2>0$. The central limit theorem states that the distribution of the sample average of these random variables approaches the normal distribution with a mean μ and variance σ^2/n. Let

$$Z_N = \frac{S_N - N\mu}{\sigma\sqrt{N}} \tag{19}$$

By the central limit theorem, the distribution of Z_N approaches the standard normal distribution:

$$Z_N \to \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt \tag{20}$$

The density function looks as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \tag{21}$$

Now we determine the probability that S_N is greater or equal than c times of its expected value, for a constant $c>1$. For $S_N = cE[S_N] = c\mu N$ then

$$\begin{aligned} Z_N &= \frac{c\mu N - N\mu}{\sigma\sqrt{N}} \\ &= \frac{(c-1)N\mu}{\Theta(\sqrt{N^{3-\beta}})\sqrt{N}} \\ &= \frac{(c-1)N\mu}{\Theta(\sqrt{N^{4-\beta}})} \\ &= \frac{(c-1)\mu}{\Theta\left(N^{1-\frac{\beta}{2}}\right)} \\ &\ge a(c-1)\mu N^{\frac{\beta}{2}-1}, \end{aligned} \tag{22}$$

where a is an appropriate constant. Therefore:

$$\Pr[S_N \ge c\mu N] \le \Pr\left[Z_N \ge a(c-1)\mu N^{\frac{\beta}{2}-1}\right]. \tag{23}$$

Let $m = a(c-1)\mu N^{\frac{\beta}{2}-1}$. Since, by the central limit theorem, the distribution of Z_N can be approximated by the standard normal distribution, we have

$$\begin{aligned} &\Pr[S_N \ge c\mu N] \le \Pr[Z_N \ge m] \\ &\approx 1-\Phi(m). \end{aligned} \tag{24}$$

□

Theorem 3: For $m = a(c-1)\mu N^{\frac{\beta}{2}-1}$:

$$1-\Phi(m) \le \frac{1}{2\sqrt{\pi}} e^{-\gamma N^{\beta-2}},$$

where $\gamma = (a(c-1)\mu)^2/2$ is a constant.

Proof:

$$\begin{aligned} 1-\Phi(m) &= 1-\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{m} e^{\frac{-t^2}{2}} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{m}^{\infty} e^{\frac{-t^2}{2}} dt \\ &= \frac{1}{2} erfc\left(\frac{m}{\sqrt{2}}\right) \end{aligned} \tag{25}$$

$$= \frac{1}{2} \frac{\Gamma\left(\frac{1}{2}, \left(\frac{m}{\sqrt{2}}\right)^2\right)}{\sqrt{\pi}},$$

where $\Gamma(a,x)$ is the incomplete gamma function.

$$\Gamma(a,x) = \int_x^\infty t^{a-1} e^{-t} dt. \quad (26)$$

For an integer r:

$$\Gamma(r,x) = (r-1)! e^{-x} \sum_{k=0}^{r-1} \frac{x^k}{k!}. \quad (27)$$

Because, for $x \geq 1/2$, $\Gamma\left(\frac{1}{2}, x\right) \leq \Gamma(1,x)$, for $m \geq 1$:

$$\begin{aligned} 1-\Phi(m) &= \frac{1}{2} \frac{\Gamma\left(\frac{1}{2}, \left(\frac{m}{\sqrt{2}}\right)^2\right)}{\sqrt{\pi}} \\ &\leq \frac{1}{2} \frac{\Gamma\left(1, \left(\frac{m}{\sqrt{2}}\right)^2\right)}{\sqrt{\pi}} \\ &= \frac{1}{2\sqrt{\pi}} e^{-\frac{m^2}{2}} \\ &= \frac{1}{2\sqrt{\pi}} e^{-\gamma N^{\beta-2}}, \end{aligned} \quad (28)$$

where $\gamma = (a(c-1)\mu)^2/2$ is a constant. □

VII. Modeling Processing Time for Similarities

As we have highlighted, similarity detecting and handling is a key issue in phonebook centric social networks. First the similarity algorithm has to find similar persons then handle them properly. Phonebookmark uses a semi-automatic similarity resolving mechanism. First it detects similarities and calculates a probability for them, which indicates how likely the corresponding phonebook contact and the member of the network identify the same person. This detecting algorithm runs in the background continuously on server side and it has to scan the members of the network at registration or synchronization events. In case of multiple similarities, Phonebookmark uses the similarity probability values to determine the proper order. The details of the algorithm are discussed in [25].

The behavior of the similarity detecting algorithm is similar to a queuing system where the processing unit is the algorithm and the entities in the queue are the person pairs which are waiting for comparison. The responsiveness of the algorithm is critical as similarity handling is a key feature of phonebook centric social networks compared to other solutions.

In the following model we consider only the registration operation, since it can bring the most similarity, because a totally new phonebook is being uploaded in the system. This operation can be divided for two main tasks. Firstly, when a member registers, she or he should be compared to every phonebook contact in all phonebooks present in the network. If we consider the number of private contacts in a phonebook as a random variable X_{Pc}, this means $E[X_{Pc}]*N$ comparisons, where $E[X_{Pc}]$ is the expected value of the phonebook sizes, N is the number of members in the network before the registration and PC denotes to private contacts. After the initial state of the social network, when the number members N is high, it can be considered as a relative constant value in one processing step of the queue model.

Based on the database of Phonebookmark we were able to estimate the distribution of phonebook sizes. Figure 6 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 provides a simple empirical test for whether a random variable has an exponential distribution. In this case the gradient of the function gives the parameter of the exponential distribution (Figure 6).

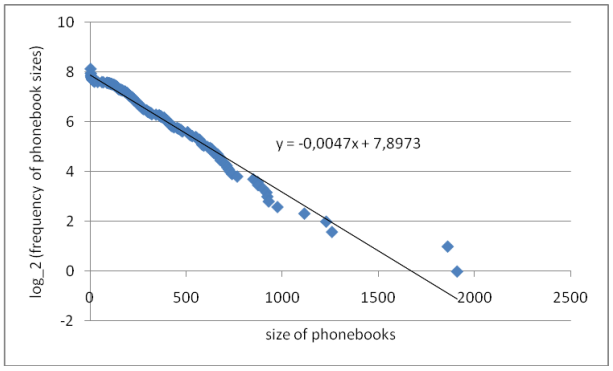

Figure 6. Size of phonebooks in Phonebookmark

In this measurement this parameter is *0.0047*, the expected value of the exponential distribution can be calculated as the reciprocal of this parameter, thus the expected value of phonebook sizes according to this measurement is *212*. Following we refer to $E[X_{Pc}]$ as C. This shows that the phonebook sizes can be modeled very well with an exponential distribution.

The other task during the member registration is to check, which members of the network are in the phonebook of the new member. This task requires $N*X_{Pc}$ comparisons, where the size of the new phonebook is modeled also with exponential distribution.

This way the amount of comparisons required by a member registration is modeled with the random variable X^*_{Pc}:

$$X^*_{Pc} = C * N + X_{Pc} * N = N * (C + X_{Pc}) \quad (29)$$

Following we show that X^*_{Pc} has exponential distribution.

Lemma 4: X^*_{Pc} random variable has exponential distribution.

Proof:

Because of the linear transformation, the distribution function of X^*_{Pc} looks as follows:

$$F_{X^*_{Pc}}(x^*) = F_{X_{Pc}}(\frac{x^* - N * C}{N}), \quad (30)$$

when $N>0$, which is always true in our case. This way since the distribution of X_{Pc} and X^*_{Pc} looks the same, X^*_{Pc} has also exponential distribution. □

We model the registration rate of members as a Poisson process with λ parameter and we assume that a person pair comparison is the time unit.

Theorem 4: In order to keep the stability of the similarity detecting the following is required for the rate of member arrival:

$$\lambda < \frac{1}{2CN}.$$

Proof: According to Kleinrock's model for queuing systems (Section 3.2 in [26]), when the arrival rate is modeled with a λ parameter Poisson distribution and the processing with exponential distribution with ν parameter then the requirement for stability:

$$\frac{\lambda}{\nu} < 1. \quad (31)$$

This means that the expected value of serving time ($1/\nu$) is smaller than the expected value of time between arrivals ($1/\lambda$). In our case the expected value of the serving time is $E[X^*_{Pc}]$, since we considered a person pair comparison as the time unit. By applying *Lemma 4* we can see that X^*_{Pc} has an exponential distribution and the expected value of it is calculated by:

$$E[X^*_{Pc}] = E[CN + X_{Pc}N] = CN + NE[X_{Pc}] = 2CN. \quad (32)$$

In case of exponential distributions, the reciprocal of the expected value is the λ parameter of the distribution. This way the requirement of the stability looks as follows:

$$\frac{\lambda}{\frac{1}{2CN}} < 1, \quad \lambda < \frac{1}{2CN} \quad (33)$$

□

This way the average number of person pairs Q waiting for comparison can be calculated (Section 3.2 in [26]) with:

$$Q = \frac{\lambda}{\frac{1}{2CN} - \lambda} \quad (34)$$

Based on this model, the resource requirement of the similarity detecting can be calculated in real environment, considering the speed of the processing unit(s). In order to demonstrate the behavior of this queue, we have made measurements regarding to the registration of the members in Phonebookmark.

Figure 7 illustrates the queue length considering $2C*N$ and $2.5C*N$ processed person-pair comparison in one step.

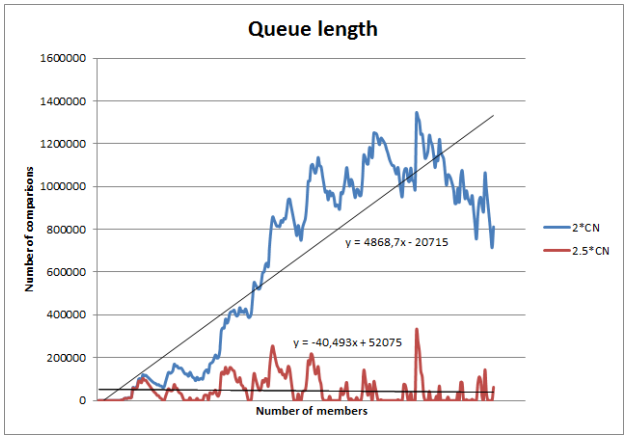

Figure 7. Queue length for similarity calculation

The x-axis shows as the number of members in the system increases, while the y-axis represents the number of comparison steps when a new member registers (sum of the remaining comparison and the new ones). It can be seen that the average queue length can be decreased significantly, when the processing speed increases.

Figure 8 illustrates the queue length normalized with the number of members.

VIII. 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.

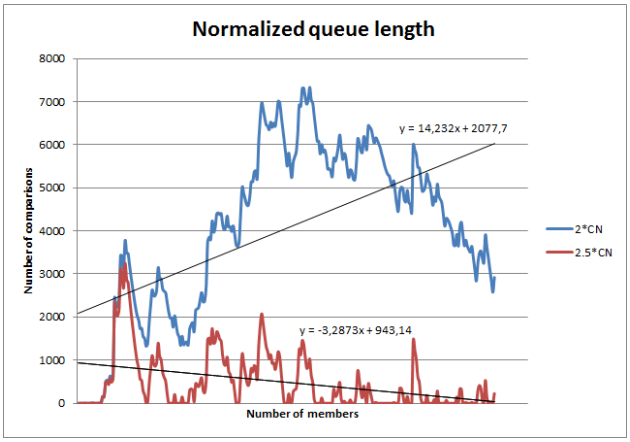

Figure 8. Normalized queue length for similarity calculation

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 similarities is a key parameter from scalability point of view. In our previous research we have shown how to estimate the expected number of similarities [23]. In order to show the accuracy of this model, in this paper we proved that, the distribution of similarities has a finite variance (Section V). This model can be used generally in other similar distributions.

After that, as the variance is finite, we applied central limit theorem to examine the accuracy of our estimation of the total number of similarities. We showed that the total number of similarities is close to their expected value. As a future work, the estimation, stated in Theorem 2, can be refined by taking the speed of the convergence to the limit distribution into account.

Finally we showed that in order to ensure the responsiveness of the network the similarity detecting should work quickly. We have given a queue based model for similarity detecting and we have shown how to calculate the expected queue length, assuming Poisson arrival for member registration. The results can be applied also for the resource requirement in different social networks providing synchronization with an external contact list.

ACKNOWLEDGMENTS

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).

IX. REFERENCES

[1] P. Ekler, T. Lukovszki. *The Accuracy of Power Law based Similarity Model in Phonebook-centric Social Networks*. In: 6th International Conference on Wireless and Mobile Communications (ICWMC). 2010.

[2] Alexa. http://www.alexa.com/topsites. February 2010.

[3] Comcast. http://www.comcast.net/articles/finance/20101230/BUSINESS-US-FACEBOOK-GOOGLE/, December 2010.

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

[5] D. M. Boyd, N. B. Ellison, *Social network sites: Definition, history, and scholarship*, Journal of Computer-Mediated Communication, Volume 13, Issue 1 (2007)

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

[7] 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, 2002.

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

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

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

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

[12] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. *Measurement and analysis of online social networks*. In ACM/USENIX IMC, 2007.

[13] 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.

[14] F. Chierichetti, R. Kumar, S. Lattanzi, M. Mitzenmacher, A. Panconesi, P. Raghavan, On Compressing Social Networks, In: Proc. of the 15th ACM SIGKDD International Conference Knowledge Discovery and Data Mining (KDD'09), 2009.

[15] 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.

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

[17] 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), 279-290, 2001.

[18] A. Fabrikant, E. Koutsoupias, and C. H. Papadimitriou. Heuristically Optimized Trade-offs: A New Paradigm for Power Laws in the Internet. In: *Proc. 29th International Colloquium on Automata, Languages and Programming (ICALP)*, 110-122, 2002.

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

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

[21] Nist special publication 500-255. In The Twelfth Text rEtrieval Conference (TREC 2003).

[22] Walters and H. W., Google scholar search performance: Comparative recall and precision libraries and the academy, volume 9/1, January, 2009.

[23] P. Ekler, T. Lukovszki. *Similarity Distribution in Phonebook-centric Social Networks*. In: 5th International Conference on Wireless and Mobile Communications (ICWMC). 2009.

[24] P. Ekler, T. Lukovszki. *Experiences with phonebook-centric social networks*. In: CCNC'10, Las Vegas. 2010.

[25] Péter Ekler, Tamás Lukovszki, *Learning Methods for Similarity Handling in Phonebook-centric Social Networks*, 10th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics (CINTI 2009), 2009.

[26] L. Kleinrock, *Queueing Systems. Volume 1: Theory*, Wiley-Interscience, pages 94-101, 1975.