Approximating Social Ties Based on Call Logs: Whom Should We Prioritize?

Mohammad Erfan, Alim Ul Gias, Sheikh Muhammad Sarwar, and Kazi Sakib

Institute of Information Technology
University of Dhaka, Dhaka–1000, Bangladesh
`{bit0326,alim,smsarwar,sakib}@iit.du.ac.bd`

Abstract. Telecommunication service providers could analyze the call logs of individual users to provide personalized services like recommending Friends and Family (FnF) numbers, that could be registered for having lower tariff during communication. However, suggesting such numbers is not simple as the decision making process can often be misled by factors like overall talk-time or call frequency. This paper proposes a technique to assist mobile phone users for effectively identifying FnF numbers using the underlying social network of a call graph. It resembles to a star network, having the user at the center, where each of the edges are weighted using three call log attributes: call duration, call location and call period. The weighted call graph is then used to calculate user's social closeness based on which FnF numbers are suggested. The experiment was conducted on a set of real life data and it is seen that the proposed technique can effectively suggest FnF numbers with an F-measure value of 0.67.

Keywords: Social network, mobile phone call graph, call log attributes, social closeness

1 Introduction

Personalized services can be defined as context-specific services to each individual client [1]. Since mobile phone can always be carried by the client, it has the potential to be an ideal medium of such personalization. This potential allows telecommunication service providers to offer multiple personalized services like customized ring tone suggestions or different location based information. Providing such services can increase the popularity of the service providers [2] and help them to sustain in the competitive market. Thus, it is high time telecommunication companies focused more on providing personalized services to each user for eventually increasing their overall revenue.

Recommending Friends and Family (FnF) numbers [3], that can be registered by the user for having lower call rate, could be an example of customized personal services. To provide such service, it is essential to retrieve each user's calling pattern by analyzing their respective call logs. However, analysis and interpretation of these call logs is particularly difficult and introduces major research issues [4].

A possible solution could be, utilizing the underlying social network [5] of those call logs. Nevertheless, constructing such network involves multiple challenges like relative prioritization of in and out degrees, and assigning weight to edges by using different call log attributes.

Although multiple research has been conducted based on mobile phone call logs, none has focused directly to provide personalized services to each individual user. There have been some work that focus on investigating social ties based on different call attributes [6], [7]. Some of the researchers have focused on predicting user behavior [8] and community dynamics [9]. Moreover, there are works which forecast incoming calls [10] and correlate international calls with the export-import business of the country [11]. However, it is arguable that whether these state of the arts can be used to provide effective personalized services that can satisfy each individual user.

This paper proposes a methodology to create a weighted call graph based on individual call logs and use the graph's underlying social network for predicting FnF numbers. To weight the edges, three call log attributes are used which are - call duration, call period and call location. The technique empirically measures the effect of these attributes towards social closeness and calculates their relative significance. The summation of these relative significances, that prevents any of the attributes being predominated, is considered as the weight. A closeness score is then calculated by aggregating the product of the weights with values respect to their call being either incoming or outgoing. This score is used by the method to rank and suggest numbers for registering as FnF.

The method was implemented and its performance was evaluated based on a real life dataset. The dataset includes all required call log attributes along with a user provided score which indicates the strength of social tie with the person of that particular call. These scores were considered as the ground truth when weights for edges of the graph are being learned. The social closeness was then measured, based on which numbers were ranked, to predict FnF numbers. The performance of the proposed technique was compared along with traditional approaches based on overall talk-time and call frequency. It seen that the method outperforms the call frequency and overall talk-time based approaches with 38% and 10% higher F-measure value.

Rest of the paper is organized as follows: Section 2 highlights state of the art techniques for analyzing call logs to derive different social contexts. The proposed technique to predict FnF numbers is presented in Section 3. Section 4 includes the experimental details and performance of the proposed technique. Finally, this paper is concluded in Section 5 with a discussion about our work and future research directions.

2 Literature Review

Although log analysis has been a major research concern for a considerable amount of time [12], its application in mobile phone call pattern analysis has started comparatively later. To the best of authors' knowledge, no work directly

focus on utilizing the underlying social network obtained from a single user's call logs to provide customized services. The work that have been done mostly focus on constructing social networks from call logs, predicting the user behavior and identifying multiple social contexts like social affinity or community dynamics. This section highlights these contributions and discusses how they contrast to our problem domain.

There have been some significant work which are concerned with predicting the user behavior. For an example, Dong et. al. proposed a technique to create an undirected binary graph from the call log to find the human behavior of different age and gender [8]. The researchers in [13] proposed a method to derive the user relation using Bayesian network. Phithakkitnukoon et. al. analyzed multiple call log attributes such as call location, talk-time, calling time and inter-arrival time to infer behavioral dependencies [10]. The researcher in [14] inferred friendship network by utilizing the behavioral and survey data. Zhang et. al. have focused on information retrieval techniques to find human behavior patterns and community dynamics [9]. There have also been work to assess the strength of social tie based on call duration [6].

Identifying different social contexts has been another field of major research interest. Phithakkitnukoon et. al. proposed a method to identify social network based on the user phone call log [15]. They have used correlation coefficient to select call log features and then extracted user's social behavioral pattern from that network. Another work of Phithakkitnukoon et. al. has been presented in [16] where they created a network graph to derive socially closest and distant members. The researchers in [11] analyzed international phone calls from Rwanda, between January 2005 to December 2008, to derive the interpersonal connections between the people of different countries. They have used those findings to measure and evaluate the social tie between nations.

Researchers have also focused on the construction of call graph by utilizing different call log attributes. The researchers in [17] created a social network graph using call duration, source number, destination number and physical location of the user. They have extracted these information from Call Detail Record (CDR) [18] files and and generated the adjacency matrix. The edges were weighted using the call duration value. Motahari et. al. have also used the CDR files and created multiple affinity networks such as family members, utility network, friends and co-workers [7]. They have used CDR files containing 4.3 million phone call data which was collected from 360,000 subscribers.

Review of the state of the art shows that these works are generally focused on predicting users' behavioral pattern rather than utilizing those patterns for providing better user experience. It is arguable that whether these methods can be used for determining user's social tie along with utilizing the findings in a separate strategy to suggest FnF numbers. From different research, it is also seen that the edges of the constructed call graph is weighted using call duration. Nevertheless, it may not be the correct representation of social affinity. The reason behind this is though frequent or long communication likely indicates a strong tie, little or no communication does not necessarily indicate a weak tie [6].

Thus it is required to incorporate other call log attributes, to assign weights on graph edges, for actual representation of individual's social tie. This weighted graph can be easily used to infer the socially closest members and their numbers can be suggested as FnF.

3   Proposed Method

This research proposes a methodology to create a social network graph using the phone call log. The network will resemble to a star network where the phone user will be the center node and individuals with whom communication has occurred are the peripheral nodes. There will be weighted edges between the central and peripheral nodes which will indicate the strength of their social ties.The graph will have two type of edges, one for the incoming calls and other for the outgoing ones. The frequency of calls will not effect the number of edges since it will be considered during the weight calculation. A sample social network is presented in Figure 1. From the figure it can be seen that communications are represented by weighted edges where the outgoing call to individual p_n has a weight w_{n1} and the incoming call has a weight w_{n2}.
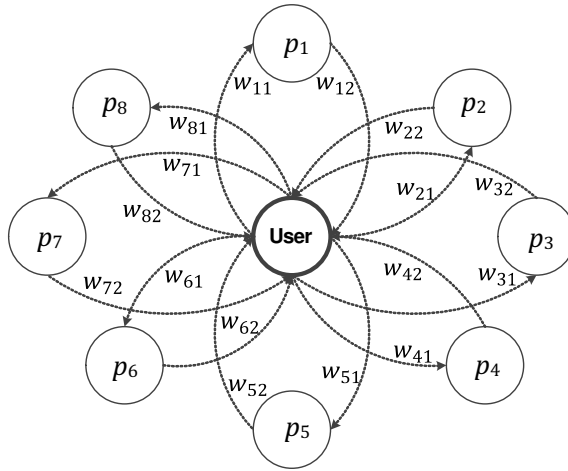
Fig. 1. Call graph generated from a single user's call logs

Initially call log records will be taken from the phone. The call logs should have the following information: call duration, call period (a period of the day like 3:00 AM to 6:00 AM) and call location (the place from which the call was made). It is essential to utilize all these information for weighting graph edges since using only call duration can often misled regarding the approximation of social closeness. For an example, a user may talk with a person at his office for a long time. However, that person could be a client instead of being a friend

or relative. So it will be wrong to consider that person as a socially close one based on the call duration. Moreover, talking with a someone during the office hours (like 9:00 AM to 5:00 PM) for a long time might misled in estimating the nature of social affinity. Thus it would be more reasonable to assign a weight by combining all these three information. The proposed method utilizes these call log attributes to assign a weight which is defined in Equation (1).

$$\omega_{ij} = \sum_{k=1}^{N} (\alpha \cdot \delta_{ij_k} + \beta_{q(ij_k)} + \gamma_{r(ij_k)}) \tag{1}$$

In Equation (1), ω_{ij} refers to the weight of the edge between the person p_i and phone's user for all calls of type j (like incoming, outgoing or missed). The set of all j type calls, with respect to person p_i is represented by $\mathcal{C}_{ij}$. If $|\mathcal{C}_{ij}| = N$, the weight ω_{ij} will be the summation of N values. The values will depend on call duration (δ), call period and call location. The value of call duration is used by multiplying it with a weight α that reflects the significance of call duration towards social ties. To represent the significance of call period and call location, two sets of values, $\mathcal{S}$ and $\mathcal{L}$ has been introduced. The elements of those two sets are represented by β and γ respectively. To define their relation with the set of calls $\mathcal{C}_{ij}$, two mapping functions $q : \mathcal{C}_{ij} \to \mathcal{S}$ and $r : \mathcal{C}_{ij} \to \mathcal{L}$ are defined.

In a nutshell, the weight will be calculated by considering each of the calls of type j from person p_i. For each particular call k, its duration will be multiplied by a significance factor α. This multiplied value will be added with the corresponding β and γ value of that particular call. However, the system must at first calculate those significance values. The social closeness should then have to be calculated for suggesting the FnF numbers. The procedure for these tasks are discussed in details in the following subsections.

3.1 Determining the significance values

Since longer communications likely indicate a strong tie [6], the proposed method uses the correlation coefficient between the call duration and social closeness as the value of α. However, information regarding the social closeness of people from each call is a pre-requisite to calculate the coefficient. To determine the values of β and γ, one has to first fix the number of call periods and locations that should be considered. The set of these periods and locations are defined as T and L respectively. For each of the elements in set T and L, the method generates a separate value of β and γ respectively. These values, combined together, are then represented by two sets $\mathcal{S}$ and $\mathcal{L}$ respectively.

The first step in calculating the values of β and γ requires to define a threshold value σ. This value is used to identify the people who are socially closest to the phone user. If the social closeness value of a person is above the threshold, that individual is considered as having a strong social tie. The second steps involves calculating the percentage of people over σ for each of the call periods and locations. These percentages provide an insight regarding the significance of each period and location. However, to consider those as weights, their relative

32

significance is needed to be considered which is calculated at the final step. This whole procedure is illustrated in Algorithm 1.

Algorithm 1 Determining the significance values

Input: The set of all selected call periods T and locations L
Output: α, $\mathcal{S}$ and $\mathcal{L}$
 1: **Begin**
 2: Set the value of α to the correlation coefficient between call duration and social closeness
 3: Define a threshold σ for social closeness to determine socially closest persons
 4: Calculate the percentage of calls from people whose social closeness value is over σ on each call period $t_i \in T$ and location $l_i \in L$
 5: Initialize $\mathcal{S} \leftarrow \emptyset$ and $\mathcal{L} \leftarrow \emptyset$
 6: **for each** $t_i \in T$ **do**
 7: $\quad \beta_i = \dfrac{\text{percentage of people over } \sigma \text{ for period } t_i}{\text{summation of percentage values } \forall_{t_i \in T}}$
 8: $\quad \mathcal{S} \cup \{\beta_i\}$
 9: **end for**
10: **for each** $l_i \in L$ **do**
11: $\quad \gamma_i = \dfrac{\text{percentage of people over } \sigma \text{ for location } l_i}{\text{summation of percentage values } \forall_{l_i \in L}}$
12: $\quad \mathcal{L} \cup \{\gamma_i\}$
13: **end for**
14: **End**

3.2 Calculating Social Closeness and Suggesting FnF numbers

After determining the values of α, β and γ, the weight ω_{ij} of call type j from person p_i can be calculated using Equation (1) . The weights ω_{ij} for each of the individuals are included in the set $\mathcal{W}_{ij}$ which is used by the proposed method to represent a social network as shown in Figure 1. To calculate the social closeness of the person p_i, the method uses a mapping function $\phi : \mathcal{J} \rightarrow \mathcal{I}$, where $\mathcal{J}$ and $\mathcal{I}$ represent the set of call types and their significances respectively. This mapping function ϕ is used to determine the weight of different call types. Assigning different weights to different call types is important since the significance of different call types (incoming or outgoing) could be different. During the calculation of social closeness for person p_i, as shown in Equation (2), the weights of that person ω_{ij} for call type j is multiplied by its respective weight provided by the mapping function ϕ.

$$\theta_i = \sum_{j=1}^{M} \phi(j) \cdot \omega_{ij} \tag{2}$$

The process of determining the social closeness and ranking the individuals based on that score is presented in Algorithm 2. Initially, as used in Algorithm 1,

a threshold σ is defined to identify socially closest persons. This threshold is used to calculate the percentage of socially closest persons for each call of type j. The percentage values are then used to determine the relative significance of different call types. Using these relative significance values, the method calculates the social closeness as defined in Equation (2). Each of the individuals, with whom communication has occurred, are sorted in descending order of social closeness. The method then suggests the numbers of certain individuals, at the top of the list, as FnF numbers.

Algorithm 2 Calculating Social Closeness

Input: The set $\mathcal{W}_{ij}$, $\mathcal{C}_{ij}$, $\mathcal{J}$ and the mapping function ϕ
Output: List of individuals sorted in descending order of social closeness
 1: **Begin**
 2: Define a threshold σ for social closeness to determine socially closest persons
 3: Calculate the percentage of calls from people whose social closeness value is over σ for each j type of calls
 4: Initialize $\mathcal{I} \leftarrow \emptyset$
 5: **for each** $j \in \mathcal{J}$ **do**
 6:     $\lambda_j = \dfrac{\text{percentage of people over } \sigma \text{ for call type } j}{\text{summation of percentage values } \forall_{j \in \mathcal{J}}}$
 7:     $\mathcal{I} \cup \{\lambda_j\}$
 8: **end for**
 9: Use Equation (2) to calculate social closeness for each of individual
10: Sort the list of individuals in descending order of social closeness
11: **End**


4 Experimental Setup and Results

This section discusses about different experimental parameters that were used during the evaluation of our proposed approach. It also includes the performance of the proposed technique with a comparison along with two baseline approaches to suggest FnF numbers.

4.1 Experimental Parameters

For experimental purpose the call logs were collected from 18 individuals. The call logs included the person's mobile phone number, call duration and call time. The call location was manually collected from them. The participants had also provided a social closeness score, to each of the persons from the call log, in a scale of 1-10. Moreover, each of them were also asked to select 3 numbers from these call logs which they could register as FnF numbers. 5 of the participants provided those numbers and thus their data were used for the testing purpose. Rest of the data were used to train the proposed method that involves learning the values of α, β, γ and λ. During the calculation of these weights, the value of

σ was set to 8 which is considerably higher enough in the social closeness scale for recognizing socially closest persons.

The value of α was calculated by determining the correlation coefficient between call duration and social closeness. It has been found that the call duration and social closeness has a positive relationship, though not that much strong. This has been illustrated in Figure 2 by a scatter diagram where social closeness is plotted against call duration. The value of correlation coefficient was found to be 0.14 and it was set as the value of α.
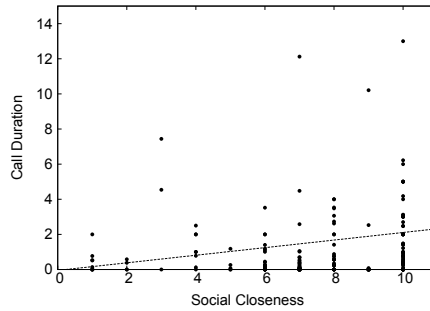
**Fig. 2.** Scatter diagram representing the correlation coefficient between social closeness and call duration

For the calculation of β, we have divided a day into 8 time periods. The percentage of socially closest calls and their relative weight is presented in Table 1. A scatter diagram is presented in Figure 3 to illustrate the percentage of socially closest people in two different time periods.

Table 1. Relative significance of different time periods used in the experiment

| Period	Percentage of calls above σ	Relative weight
12:00 AM to 3:00 AM	60%	0.163
3:00 AM to 6:00 AM	67%	0.18
6:00 AM to 9:00 AM	25%	0.07
9:00 AM to 12:00 PM	55%	0.15
12:00 PM to 3:00 PM	26%	0.071
3:00 PM to 6:00 PM	41%	0.112
6:00 PM to 9:00 PM	56%	0.153
9:00 PM to 12:00 AM	37%	0.101

For calculating the value of γ, the locations were categorized before collecting the data. For each of the call logs, the location was selected as either home or work place. Although there were instances where users were neither in home nor

(a) Time period 3:00 PM to 6:00 PM (b) Time period 9:00 PM to 12:00 AM
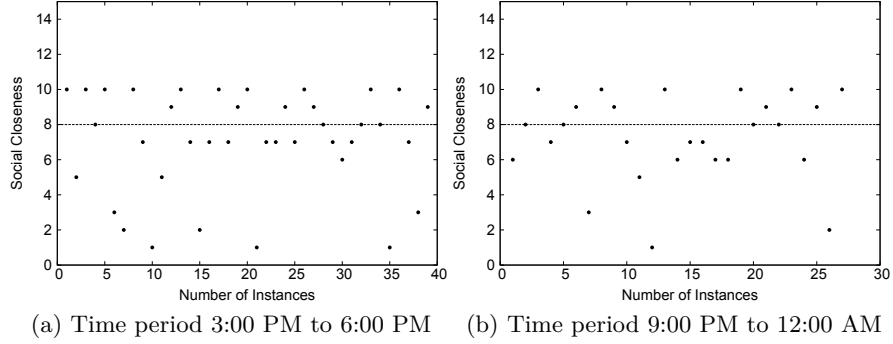
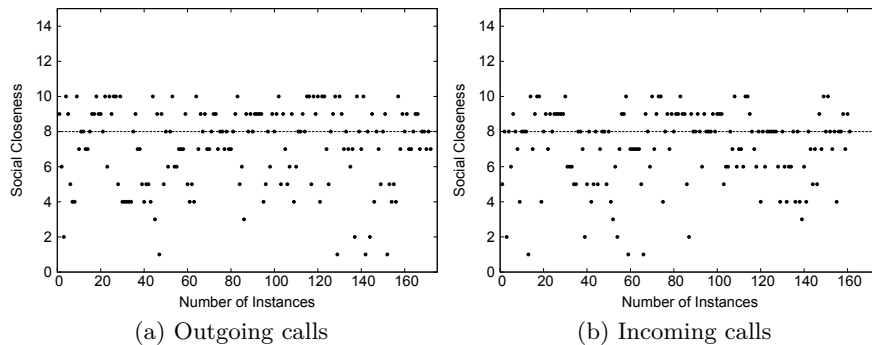**Fig. 3.** Visualization of calls from socially closest persons in two different time periods

in work place, still they had used various measures two provide the location information. Examples include selecting the place nearest to their current position or based on the nature of call and its period. These data, as presented in Figure 4, was used to calculate the value of γ. For the home and workplace, the value turned out to be 0.44 and 0.56 respectively.

(a) Location - Home (b) Location - Workplace
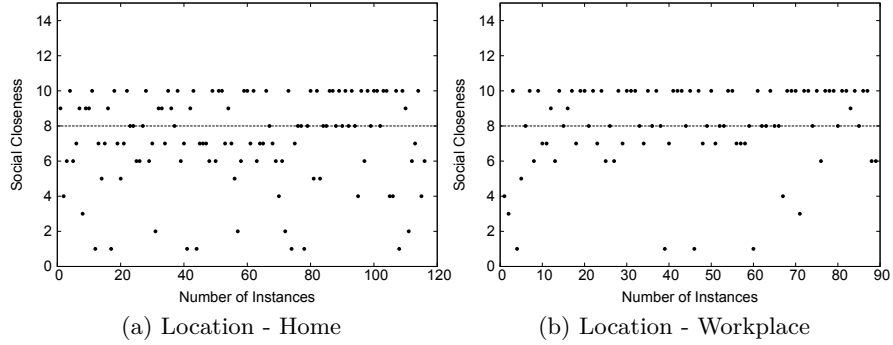
Fig. 4. Visualization of calls from socially closest persons in two separate locations

In this work, ignoring the missed calls, incoming and outgoing calls has been considered to measure the value of λ. The percentage of calls from socially closest persons for outgoing and incoming calls can be visualized from Figure 5. Since there were only two call types, after calculating the weight of work place, the other weight was calculated by Equation 3. The weights for the incoming and outgoing calls has been calculated as 0.44 and 0.56 respectively.

$$\phi(j) = \begin{cases} \lambda & \text{if the value of j is 1} \\ 1 - \lambda & \text{otherwise} \end{cases} \tag{3}$$

| (a) Outgoing calls | (b) Incoming calls |

Fig. 5. Visualization of incoming and outgoing calls from socially closest persons

4.2 Performance Evaluation

The proposed technique tried to match 2 out of 3 FnF numbers provided by each individuals. The system suggested 5 numbers, one after another, and stopped at the point where two numbers had being matched. Thus for 5 individuals, the system tried to match 10 FnF numbers with at most 25 attempts. We have compared the performance of our system with two baseline approaches which are based on overall call duration and call frequency. The evaluation technique for these two approaches remained the same. It is seen that the proposed method has identified 7 FnF numbers with 22 attempts. On the other hand, the duration-based and frequency based approaches has identified 6 and 3 numbers with 22 and 25 attempts respectively.

Table 2. Performance comparison of the proposed technique with respect to duration-based and frequency based approaches

Approaches	Precision	Recall	F-measure
Duration-based	0.27	0.6	0.57
Frequency-based	0.8	0.2	0.19
Proposed technique	0.32	0.7	0.67

The performance of the technique, as shown in Table 2, is interpreted in terms of precision, recall and F-measure. It is seen that proposed method outperforms the other two approaches though the value of the precision is low. However, in this case, recall is more important than precision. This is because for each of the individuals, we are suggesting at most 5 numbers to match with 2 numbers. Due to this, the precision will always be lower due to having more false positives. Nevertheless, the user will always be satisfied if the system chooses more from the set of relevant numbers. So during the calculation of F-measure, as done in [19], the method needs to vary the weights of precision and recall. Thus in our

case, recall has been weighted five times higher than the precision during the calculation of F-measure using Equation (4).

$$F_B = (1 + B^2) \cdot \frac{precision \cdot recall}{B^2 \cdot precision + recall} \qquad (4)$$

For the individuals with whom the performance was evaluated, the pattern of score provided by the system and the social closeness score provided by the phone users is illustrated in Figure 6. Although the scale of those two scores are different, it is seen that their curves are very much similar. As the system suggested FnF numbers based on calculated closeness scores, the performance is better than the other two approaches.
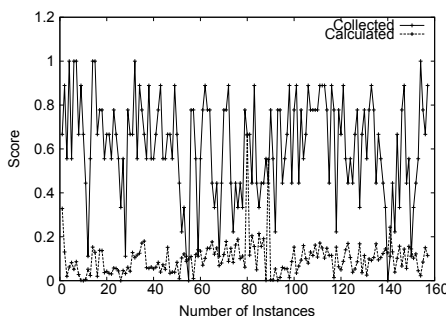
Fig. 6. The pattern of the scores collected from the individuals and calculated by the proposed method

5 Conclusion and Future Work

This paper introduces a technique to prioritize individuals based on call logs and suggest FnF numbers using that prioritize list. The method utilizes the underlying social network of the call graph for a single phone user. The edges of the call graph has been weighted during the prioritization process to incorporate different call log attributes like duration, period and location. The proposed methodology was implemented and tested against a set of real life data. The system has illustrated its effectiveness in identifying FnF numbers with an F-measure value of 0.67. This research also introduces a set of new research ideas for the research communities. This includes assessing the existence of a nonlinear relationship, between duration and social closeness, to improve the performance of the system. Moreover, one could utilize other call log attributes and optimize their total number to get better performance.

Acknowledgement

References

1. Ho, S.Y., Kwok, S.H.: The attraction of personalized service for users in mobile commerce: an empirical study. ACM SIGecom Exchanges 3(4), 10–18 (2002)
2. Witteman, M.: Efficient proximity detection among mobile clients using the GSM network. Master's thesis, University of Twente (2007)
3. Dey, B., Newman, D., Prendergast, R.: Analysing appropriation and usability in social and occupational lives: An investigation of bangladeshi farmers' use of mobile telephony. Information Technology & People 24(1), 46–63 (2011)
4. De Melo, P.O.V., Akoglu, L., Faloutsos, C., Loureiro, A.A.: Surprising patterns for the call duration distribution of mobile phone users. In: Machine learning and knowledge discovery in databases, pp. 354–369. Springer (2010)
5. Chin, A., Zhang, D.: Mobile Social Networking. Springer (2014)
6. Wiese, J., Min, J.K., Hong, J.I., Zimmerman, J.: Assessing Call and SMS Logs as an Indication of Tie Strength. Tech. rep., Human-Computer Interaction Institute, School of Computer Science, Carnegie Mellon University (2014)
7. Motahari, S., Mengshoel, O.J., Reuther, P., Appala, S., Zoia, L., Shah, J.: The impact of social affinity on phone calling patterns: Categorizing social ties from call data records. In: Proceedings of the Sixth Workshop on Social Network Mining and Analysis. pp. 1–9. ACM (2012)
8. Dong, Z.B., Song, G.J., Xie, K.Q., Wang, J.Y.: An experimental study of large-scale mobile social network. In: Proceedings of the 18th International Conference on World Wide Web. pp. 1175–1176. ACM (2009)
9. Zhang, D., Guo, B., Yu, Z.: The emergence of social and community intelligence. Computer 44(7), 21–28 (2011)
10. Phithakkitnukoon, S., Dantu, R., Claxton, R., Eagle, N.: Behavior-based adaptive call predictor. ACM Transactions on Autonomous and Adaptive Systems (TAAS) 6(3), 21 (2011)
11. Blumenstock, J.E.: Using mobile phone data to measure the ties between nations. In: Proceedings of the 2011 iConference. pp. 195–202. ACM (2011)
12. Oliner, A., Ganapathi, A., Xu, W.: Advances and challenges in log analysis. Communications of the ACM 55(2), 55–61 (2012)
13. Park, H.S., Cho, S.B.: Building mobile social network with semantic relation using Bayesian network-based life-log mining. In: Second International Conference on Social Computing (SocialCom). pp. 401–406. IEEE (2010)
14. Eagle, N., Pentland, A.S., Lazer, D.: Inferring friendship network structure by using mobile phone data. Proceedings of the National Academy of Sciences 106(36), 15274–15278 (2009)
15. Phithakkitnukoon, S., Dantu, R.: Inferring social groups using call logs. In: On the Move to Meaningful Internet Systems: OTM 2008 Workshops. pp. 200–210. Springer (2008)
16. Phithakkitnukoon, S., Dantu, R.: Mobile social group sizes and scaling ratio. AI & society 26(1), 71–85 (2011)
17. Zargari, H.: Social network modelling: Retrieval correlated graphs by mobile phone's chronological billing files. In: The Second International conference "Problems of Cybernetics and Informatics". pp. 67–70 (2008)
18. Horak, R.: Telecommunications and Data Communications Handbook. John Wiley & Sons (2007)
19. Li, X., Wang, Y.Y., Acero, A.: Learning query intent from regularized click graphs. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 339–346. ACM (2008)