# Discovery of Social Groups Using Call Detail Records

Huiqi Zhang and Ram Dantu

Department of Computer Science and Engineering
University of North Texas
Denton, TX 76201 USA
{hz0019,rdantu}@unt.edu

**Abstract.** In this paper we propose the affinity model for classifying social groups based on mobile phone call detail records. We use affinity to measure the similarity between probability distributions. Since phone calls are stochastic process, it makes more sense to use probability affinity to classify the social groups. This work is useful for enhancing homeland security, detecting unwanted calls (e.g., spam) and product marketing. For validation of our results, we used actual call logs of 100 users collected at MIT by the Reality Mining Project group for a period of 8 months. The experimental results show that our model achieves good performance with high accuracy.

**Keywords:** Social groups, call detail records, Hellinger distance, affinity.

## 1 Introduction

Social groups can be defined as sets of people who have common interests, share their experience, express similar ways of thinking and reacting, share the same opinions, do similar things and have the same goals.

There are various applications for social groups. For example, in marketing, if someone buys something, his family members and friends are likely to have the same interests to buy the same or a similar thing and have a similar level of income although we do not know how much they earn. So we may find potential buyers by social groups. Another important application for social groups is in the area of national security. For example, if somebody is a terrorist or robber, his intimate friends or socially close communication partners are likely (not necessary) to be terrorists or robbers too, since no law-abiding person wants to have some friends who are terrorists or robbers. One more application is used to quantify the telecommunication presence. On different days and at different times people usually would like to communicate with different groups of people. For example, we prefer to communicate with our colleagues in work time and to communicate with our family members, relatives and friends in non-work time. Further, in our busy hours we only would like to have necessary communications with our socially close members such as family members, bosses and others. Additionally, we may enhance detecting unwanted calls (e.g., spam) by social groups. For example, the spammers are definitely socially far from us. If we are not sure that the incoming calls which come from socially far

R. Meersman, Z. Tari, and P. Herrero (Eds.): OTM 2008 Workshops, LNCS 5333, pp. 489–498, 2008.

members are spam or not, the system may not let the phone ring and forward the calls to the voice box automatically.

Most social network research and social relationship analysis are based on blogs, emails or the World Wide Web [1- 26]. Since mobile phones have become the main communication media for people in recent years, some researchers' interests in social networks concentrate on social relationship analysis based on call detail records [27-37].

In this paper we propose the affinity [38] model to classify social groups. In Section 2 we briefly review the related work. In Section 3 we describe the model and method for social groupings. We performed the experiments with actual call logs and discuss the results in Section 4. We describe the validation of our model, conducted by the actual call logs, in Section 5. Finally, we have the conclusions in Section 6.

## 2 Related Work

A social network is defined as a set of actors (individuals) and the ties (relationships) among them [1]. There are two fundamental interests in social networks: the relational ties and the actors. In [2] the author summarized the social network measures involving the relational ties, the actors, and the overall consequences of the network topology. In [3] the authors proposed a concept describing a composite network that incorporates the multi-dimensionality of interpersonal relations is the meta-matrix.

In [4] the author discusses the strength of social relations between two persons measured with the email conversation. By the Loom system in [5] users can visualize social relationships on the basis of Usenet conversation. In [6] the several text analysis procedures are used to compute a net and visualizations of social relations among authors. Microsoft Netscan [7] is used for searching, visualizing and analyzing newsgroups. The various visualization techniques and extensive relation analysis and content analysis are combined to allow for an improved navigation in the Usenet. The Friend of-a-Friend (FOAF) project [8] explores the application of semantic web technologies to describe personal information: professional and personal lives, their friends, interests and other social relations. In [9, 10] the automatic detection methods of subgroup structure of social networks are investigated based on expected density and edge betweenness. In [11] the authors use block technique, which focuses on the pattern of connectivity to find clusters of relationships that might be hidden. In [12] the generalized block modeling method was proposed to enable partitions of smaller, more precise block types and predefined block structures based on attribute data. In [9, 13, 14, 15, 16] the studies focused on identifying tightly-connected clusters or communities within a given graph and inferring potential communities in a network based on density of linkage. In [17] the authors investigate an extensive characterization of the graph structure of the web with various features and consider the subgraph of the web consisting of all pages containing these features. In [18, 19] the self-identified nature of the online communities is used to study the relationship between different newsgroups on Usenet. In [20, 21, 22, 23, 24] social network evolution is studied as its members' attributes change.

The social network analysis and social clusters of the above work are mainly based on blogs, emails or the World Wide Web [23, 25, 26]. In [27, 28] the structure and tie strength of mobile telephone call graphs was investigated. In [29] the authors applied

the spectral clustering method to telephone call graph partition. In [30] the authors discovered the communities of mobile users on call detail records using a triangle approach. In [31] the algorithm based on clique percolation was developed to investigate the time dependence of overlapping social groups so as to discover relationships characterizing social group evolution and capturing the collaboration between colleagues and the calls between mobile phone users. In [32] the authors performed a new method for measuring human behavior, based on contextualized proximity and mobile phone data, to study the dyadic data using the nonparametric multiple regression quadratic assignment procedure (MRQAP), a standard technique to analyze social network data [33, 34], discover behavioral characteristics of friendship using factor analysis and predict satisfaction based on behavioral data. In [35] the authors studied the stability of social ties by defining and measuring the persistence of the ties. In [36] the authors propose a spreading activation-based technique to predict potential churners by examining the current set of churners and their underlying social network. In [37] the spatiotemporal anomalies of calls and patterns of calling activity are investigated using standard percolation theory tools. Almost all above research focused on large social networks and social groups. We focus on individual social groups using the probability model, *affinity*, which is different from the previous work on the measurements based on mobile phone call detail records.

## 3 Model

### 3.1 Formulation

Groups correspond to clusters of data. Cluster analysis concerns a set of multivariate methods for grouping data variables into clusters of similar elements.

In first step we apply affinity, that is measured in a probability scale, instead of simple/basic similarity coefficients. In the second step we define an aggregation criterion for merging similar clusters of elements. In the third step we use some way to assess the validity of the clustering results.

Affinity measures the similarity between probability measures. A related notion is the Hellinger distance. Since our problem belongs to discrete events, we only consider finite event spaces. Let

$$S_N = \{P = (p_1, p_2, ...p_N) \mid p_i \geq 0, \sum_{i=1}^{N} p_i = 1\}$$

be the set of all complete finite discrete probability distributions and $P,Q \in S_N$. The Hellinger distance between P and Q is defined as [4]

$$d_H^2(P,Q) = \frac{1}{2}\sum_{i=1}^{N}(\sqrt{p_i} - \sqrt{q_i})^2 \tag{1}$$

$d_H^2(P,Q) \in [0,1]$, $d_H^2(P,Q) = 0$ if P = Q and $d_H^2(P,Q) = 1$ if P and Q are disjoint. The affinity between probability measures P and Q is defined as [4]

$$A(P,Q) = 1 - d_H^2(P,Q) = \sum_{i=1}^{N}\sqrt{p_i q_i} \tag{2}$$

$A(P,Q) \in [0,1]$, $A(P,Q) = 1$ if P = Q and $A(P,Q) = 0$ if P and Q are disjoint.

Proof.

$$A(P,Q) = 1 - d_H^2(P,Q) = 1 - \frac{1}{2}\sum_{i=1}^{N}(\sqrt{p_i} - \sqrt{q_i})^2 = 1 - \frac{1}{2}\sum_{i=1}^{N}(p_i - 2\sqrt{p_i}\sqrt{q_i} + q_i)$$

$$= 1 - \frac{1}{2}(\sum_{i=1}^{N} p_i - 2\sum_{i=1}^{N}\sqrt{p_i q_i} + \sum_{i=1}^{N} q_i) = \sum_{i=1}^{N}\sqrt{p_i q_i}$$

For finite and discrete data, let M(X,Y) be a L×N matrix, where X represents the set of data units and Y is a set of N categorical variables. In this paper $Y_j$ (j=1,...,N) is a vector of frequencies. Thus $Y_j$ may be the represented by the L coordinates $n_{ij}$ (i=1, 2, …, L) which is a frequency. We will refer to the j-th column profile as the corresponding conditional vector with $n_{ij} \Big/ \sum_{i=1}^{L} n_{ij}$ . This profile vector may be a discrete conditional probability distribution law. It is often a profile or probability vector of the population, where the set X of L data unit represents a partition of some random sample of subjects in L classes. In this paper $p_i = n_{ij} \Big/ \sum_{i=1}^{L} n_{ij}$ . The column profiles have a major role since the similarity between pairs of variables will be measured using an appropriate function, the affinity in this paper, of their profiles.

### 3.2 Real-Life Data Sets and Parameters

***Real-life traffic profile*:** In this paper, the actual call logs are used for analysis. These actual call logs are collected at MIT [39] by the Reality Mining Project group for a period of 8 months. This group collected mobile phone usage of 100 users, including their user IDs (unique number representing a mobile phone user), time of calls, call direction (incoming and outgoing), incoming call description (missed, accepted), talk time, and tower IDs (location of phone users). These 100 phone users are students, professors and staff members. The collection of the call logs is followed by a survey of feedback from participating phone users for behavior patterns such as favorite hangout places; service providers; talk time minutes and phone users' friends, relatives and parents. We used this extensive dataset for our social group analysis in this paper. More information about the Reality Mining Project can be found in [39].

In our lives we have relationships with a small group of individuals in our social network such as family members, relatives, friends, neighbors and colleagues. Based on these social relationships, we divide the time of a day into working time (8am-5pm) and nonworking time (5:01pm-7:59am). Further, in the two time periods we divide our social network members into three categories: socially close members, socially near members and socially far members.

- *Socially Close Members:* These are the people with whom we maintain the strongest social relationship. Quantifying by phone calls we receive more calls from them and we tend to talk to them for longer periods of time. Family members, intimate friends and colleagues in the same team belong to this category.
- *Socially Near Members:* These relationships are not as strong as those of family members, intimate friends and colleagues in the same team. Sometimes, not always, we connect each other and talk for a considerably longer

periods. We mostly observe intermittent frequency of calls from these people. Distant relatives, general friends, colleagues in a different team and neighbors are in this category.

- *Socially Far Members:* These people have weaker relationships with each other in social life. They call each other with less frequency. We seldom receive calls from them and talk each other in short time.

We use the affinity formula (2) to classify social groups based on the time of the day, call frequencies, reciprocity and call duration.

*Time of the day:* Everyone has his/her own schedule for working, studying, entertainment, sleeping, traveling and so on. The schedule is mainly based on the time of the day and day of the week. We divide the time of the day into two parts: working time (8am-5pm) and nonworking time (5:01pm-7:59am).

*Call frequencies:* The call frequency is the number of incoming or outgoing calls in a period of time. The more the number of incoming or outgoing calls in a period of time, the more socially close the caller and callee relationship.

*Call duration:* The call duration is how long both caller and callee want to talk to each other. The longer the call duration is in a period of time, the more socially close the caller and callee relationship.

*Reciprocity:* Reciprocity represents the response by one party to calls from another party.

### 3.3 Computing the Affinity

In this paper, we use three attributes incoming (*in*), outgoing (*out*) and reciprocity (*reci*) of calls.

Let $m_i, n_i$ be the number of calls, where $i \in \{in, out, reci\}$. $P = (p_{in}, p_{out}, p_{reci})$ is a vector of normalized frequencies over the training period and $Q = (q_{in}, q_{out}, q_{reci})$ is a vector of normalized frequencies of the same attributes observed over the testing period. Then

$p_i = m_i / \sum_i m_i$ where $i \in \{in, out, reci\}$ and $q_i = n_i / \sum_i n_i$ where $i \in \{in, out, reci\}$.

The affinity between P and Q is computed as follows:

$$A(P,Q) = \sum_i \sqrt{p_i q_i} \text{ where } i \in \{in, out, reci\} \quad (3)$$

We used the data from the data set of four months, the Fall semester since the communication members were relatively less changed in a semester for students. We compute the affinity values using formula (3) for four months of the call log data. We define:

- Socially close members if $0.7 < A(P, Q) \leq 1$
- Socially near members if $0.3 < A(P, Q) \leq 0.7$
- Socially far members if $0 \leq A(P, Q) \leq 0.3$

## 4 Experiment Results and Discussion

In Figure 1, the x-axis indicates the phone numbers that are used to communicate with user29 for four months, and the y-axis indicates the affinity values based on both number of calls and call duration respectively. From figure 1 user29 has seven socially close members, eight socially near members and twenty four socially far members in this four-month period. The details of group members are listed in table 1. In Table 1 we divide the social group members into work time members and non-work time members. In general, during work time we prefer to talk to colleagues, bosses, secretaries, clients and customers, occasionally speak to family members and friends for special cases, and during non-work time we usually talk to family members and friends, and we occasionally speak to colleagues, clients and customers for special cases. Note that some people may be both our work time colleagues and non-work time friends. Thus the set of work time members and the set of non-work time members may overlap. In Table 1 user29, who was a student, had one socially close member, two socially near members and one socially far member in work time and seven socially close members, eight socially near members and twenty-four socially far members in non-work time.

Figure 2 shows the call network of subset of call detail records in one month in which there are 326 vertices labeled by phone number ids which denote the communication members and the corresponding arcs representing the incoming or outgoing calls by the arrows. There are about 3200 communication members in the four month call detail records. Since the space is limited, we only use the call network of subset of call detail records to show the relationships among the communication members. The phone number id of user29 is 264 and the part of his communication members is shown in Figure 2.

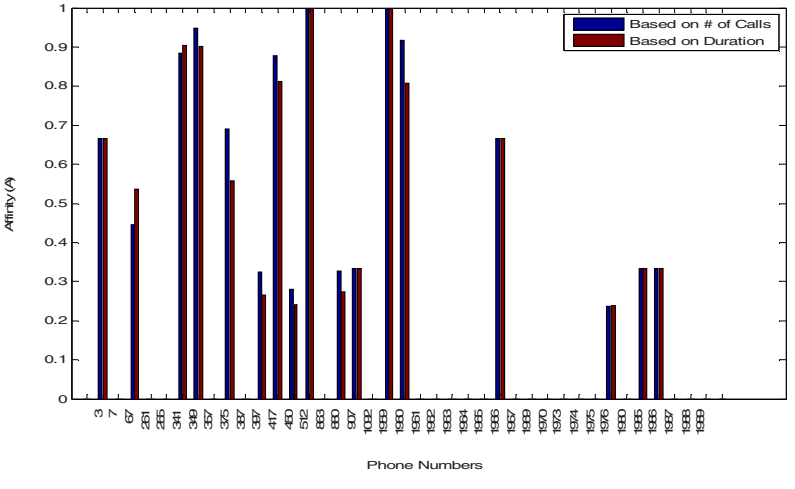


**Fig. 1.** The affinity values for user29

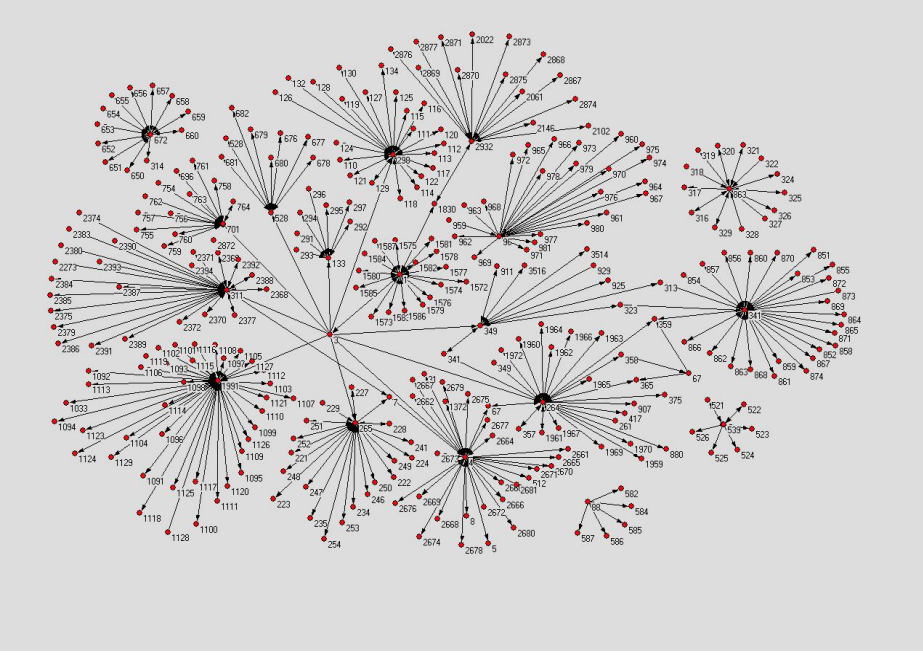

**Fig. 2.** The call network of subset of call detail records

**Table 1.** Social groups for phone user29 (The results in the Table 1 are computed by the equation 3)

| User ID (Total # of members) | Work Time Member ID (# of members) | | | Non-work Time Member ID (# of members) | | |
|---|---|---|---|---|---|---|
| | Close | Near | Far | Close | Near | Far |
| User29 (39) | 1959 (1) | 67, 1968 (2) | 1967 (1) | 265, 349, 375, 417, 512, 1959, 1960 (7) | 3, 67, 397, 863, 907, 1966, 1985, 1986 (8) | 7, 261,341, 357, 387, 417, 450, 863, 1092 1961,1962, 1963,1964, 1965,1969, 1970,1973, 1974,1975, 1976,1980, 1987,1988, 1989 (24) |

## 5 Validation

To evaluate the accuracy of our model, we used actual call logs of 100 phone users and randomly choose 10 phone users. These users include students, professors and

staff members. The best way to validate the results is to contact the phone users to get feedback. But because of the privacy issues it is almost impossible to use this way. Thus we use quantitatively hand-labeling methods to validate our model. We have used the data of the four months to classify the social groups. In order to validate our model, we hand labeled the communication members based on the number of calls, duration of calls in the period, history of call logs, location, time of arrivals, and other humanly intelligible factors.

Table 2 describes the experimental results for 10 phone users. Our model achieves good performance with high accuracy of 94.19%.

**Table 2.** Social groups for phone users

| User ID | Total # of members | Close | Near | Far | Hit | Fail | Unsure |
|---|---|---|---|---|---|---|---|
| 29(student) | 39 | 7 | 8 | 24 | 38 | 0 | 1 |
| 41(professor) | 39 | 6 | 6 | 27 | 23 | 0 | 2 |
| 21(student) | 20 | 5 | 2 | 13 | 18 | 1 | 1 |
| 74(student) | 13 | 2 | 4 | 7 | 12 | 0 | 1 |
| 88(staff) | 66 | 5 | 9 | 42 | 63 | 0 | 3 |
| 33(staff) | 31 | 4 | 2 | 25 | 31 | 0 | 0 |
| 15(student) | 29 | 10 | 4 | 15 | 25 | 2 | 2 |
| 49(student) | 18 | 6 | 2 | 10 | 16 | 1 | 1 |
| 50(student ) | 63 | 6 | 14 | 43 | 61 | 0 | 2 |
| 95(professor) | 8 | 1 | 4 | 3 | 8 | 0 | 0 |

## 6 Conclusion

In this paper we proposed the affinity model for classifying the social groups based on mobile phone call detail records. We use affinity to measure the similarity between probability distributions.

We may find the short-term friends, say a month, or long-term friends, say a year or more years using our model by adjusting the parameters.

This work is useful for enhancing homeland security, detecting unwanted calls (e.g., spam), communication presence, marketing etc. The experimental results show that our model achieves good performance with high accuracy.

In our future work we plan to detail the social group classification, analyze the social group evolution and study the social group dynamics.

## Acknowledgement

We would like to thank Nathan Eagle and Massachusetts Institute of Technology for providing us the call logs of Reality Mining dataset.

This work is supported by the National Science Foundation under grants CNS-0627754, CNS-0516807, CNS-0619871 and CNS-0551694. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References


1. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
2. Brass, D.J.: A social network perspective on human resources management. Research in Personnel and Human Resources Management 13, 39–79 (1995)
3. Carley, K.M., Lee, J., Krackhardt, D.: Destabilizing networks. Connections 24(3), 79–92 (2002)
4. Ogatha, H.: Computer Supported Social Networking for Augmenting Cooperation. In: Computer Supported Cooperative Work, vol. 10, pp. 189–209. Kluwer Academic Publishers, Dordrecht (2001)
5. Donath, J., Karahalios, K., Viegas, F.: Visualizing conversation. In: Proceeding of Hawaii International Conference on System Sciences, vol. 32 (1999)
6. Sack, W.: Conversation Map: A text-based Usenet Newsgroup Browser. In: Proceeding of ACM Conference on Intelligent User Interfaces, pp. 233–240 (2000)
7. N.N. Microsoft Netscan, `http://netscan.research.microsoft.com`
8. Brickley, D., Miller, L.: FOAF Vocabulary Specification, Namespace Document (2004), `http://xmlns.com/foaf/0.1`
9. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. In: Proceedings of the National Academy of Sciences of United State of America, vol. 99(12), pp. 7821–7826 (2002)
10. Newman, M.E.J.: Modularity and community structure in networks. In: Proceedings of the National Academy of Sciences, vol. 103, pp. 8577–8583 (2006)
11. Broder, Kumar, A.R., Maghoul, F., Raghavan, Rajagopalan, P.S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. Computer Networks 33(2), 309–320 (2000)
12. Doreian, P., Batageli, V., Ferligoj, A.: Generalized Blockmodeling. In: Granovetter, M. (ed.), Cambridge University Press, Cambridge (2005)
13. Flake, G., Lawrence, S., Giles, C.L., Coetzee, F.: Self-Organization and Identification of Web Communities. IEEE Computer 35(3) (2002)
14. Flake, G.W., Tarjan, R.E., Tsioutsiouliklis, K.: Graph Clustering and Minimum Cut Trees. Internet Math. vol. 1 (2004)
15. Hopcroft, J., Khan, O., Kulis, B., Selman, B.: Natural communities in large linked networks. In: Proceeding of 9th SIGKDD (2003)
16. Newman, M.E.J.: Detecting community structure in networks. Eur. Phys. J. B 38, 321–330 (2004)
17. Dill, S., Kumar, R., McCurley, K., Rajagopalan, S., Sivakumar, D., Tomkins, A.: Self-similarity in the Web. In: 27th International Conference on Very Large Data Bases (2001)
18. Borgs, C., Chayes, J., Mahdian, M., Saberi, A.: Exploring the community structure of newsgroups. In: Proceeding of 10th ACM SIGKDD (2004)
19. Viegas, F., Smith, M.: Newsgroup Crowds and AuthorLines. In: Hawaii International Conference on System Science (2004)
20. Holme, P., Newman, M.: Nonequilibrium phase transition in the coevolution of networks and opinions. arXiv physics/0603023 (2006)
21. Sarkar, P., Moore, A.: Dynamic Social Network Analysis using Latent Space Models. SIGKDD Explorations: Special Edition on Link Mining (2005)
22. Backstrom, L., Huttenlocher, D., Kleinberg, J.: Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD, pp. 544–554 (2006)

23. Kossinets, G., Watts, D.: Empirical analysis of an evolving social network. Science 311, 88–90 (2006)
24. Wang, X., McCallum, A.: Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. In: Proceeding of 12th ACM SIGKDD (2006)
25. Kumar, R., Novak, J., Raghavan, O., Tomkins, A.: Structure and Evolution of blogspace. Communications of ACM 47(12), 35–39 (2004)
26. Kumar, R., Novak, J., Tomkins, A.: Structure and Evolution of on line social networks. In: Proceedings of the 12th ACM SIGKDD (2006)
27. Nanavati, A.A., Gurumurthy, S., Das, G., Chakraborty, D., Dasgupta, K., Mukherjea, S., Joshi, A.: On the structural properties of massive telecom graphs: Findings and implications. In: Proceedings of the Fifteenth ACM CIKM Conference (2006)
28. Onnela, J.P., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., Kertesz, J., Barabasi, A.L.: Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences of United State of America 104(18), 7332–7336 (2007)
29. Kurucz, M., Benczur, A., Csalogany, K., Lukacs, L.: Spectral Clustering in Telephone Call Graphs. In: Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop (2007)
30. Teng, W., Chou, M.: Mining communities of acquainted mobile users on call detail records. In: Proceedings of the 22nd Annual ACM Symposium on Applied Computing (2007)
31. Palla, G., Barabasi, A., Vicsek, T.: Quantifying social group evolution. Nature 446, 664–667 (2007)
32. Eagle, N., Pentland, A., Lazer, D.: Inferring Social Network Structure using Mobile Phone Data. Science (in submission), `http://reality.media.mit.edu/pdfs/network_structure.pdf`
33. Baker, F.B., Hubert, L.J.: The analysis of social interaction data. Social Methods Res. 9, 339–361 (1981)
34. Krackhardt, D.: Predicting With Networks - Nonparametric Multiple-Regression Analysis of Dyadic Data. Social Networks 10(4), 359–381 (1988)
35. Hidalgo, A.C., Rodriguez-Sickert, C.: The Dynamics of a Mobile Phone Network. Physica A 387, 3017–3024 (2008)
36. Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A.: Social Ties and their Relevance to Churn in Mobile Telecom Networks. In: Proceedings of the 11th ACM international conference on Extending database technology: Advances in database technology (2008)
37. Candia, J., Gonzalezm, M.C., Wang, P., Schoenharl, T., Madey, G., Barabasi, A.: Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical 41, 224015 (2008)
38. Fannes, M., Spincemaile, P.: The mutual affinity of random measures. Periodica Mathematica Hungarica 47(1-2), 51–71 (2003)
39. Massachusetts Institute of Technology: Reality Mining (2008), `http://reality.media.mit.edu/`
40. Zhang, H., Dantu, R.: Quantifying the presence for phone users. In: Proceeding of Fifth IEEE Consumer Communications & Networking Conference (2008)