Social Relationship Classification based on Interaction Data from Smartphones

Deyi Sun **Wing Cheong Lau**
Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong

***Abstract*—Wireless Communications and Mobile Computing have fundamentally changed the way people interact and communicate with each other. As the command-center of the user's communications with the outside world, smartphones hold the key to understand the user's social relationship with other people of interest. In this paper, we propose to use the unique multi-model interaction data from smartphone to classify social relationships. We firstly carry out a social interaction data collection campaign with a group of smartphone users to obtain real-life multi-modal communication data and model the data as a *social interaction matrix*. Then we perform a statistical analysis on the social interaction matrix to identify the interesting interaction patterns in the data. After applying different classification algorithms on social interaction matrix, we find that SVM outperforms KNN and decision tree algorithms, with a classification accuracy of 82.4% (the accuracies of KNN and decision tree are 79.9% and 77.6% respectively). Additionally, with dimensionality reduction algorithms, we embed the social interaction data containing 65 features into a 9-dimensional space while preserving the high classification accuracy. We also demonstrate the viability of applying CUR decomposition to identify important features so as to conserve energy during interaction data collection. In particular, based on the 13 out of 65 features selected by the CUR approach, we can still achieve classification accuracy of 77.7% while substantially cut down the amount of raw interaction data to be collected, stored and processed.**

***Index Terms*—social relationship; smartphone;**

I. Introduction

Social relationship is a significant part of our life. It is obvious that the type of relationship between a pair or group of people does have strong impact on how they would respond to different types of events. For example, the promotion/spreading of a new Gaming CD can be more effective via group of high-school classmates rather than among the members of a family [9]; members of the same football team will have much stronger influence on each other regarding their choice of sports gear. We therefore believe that social relationship identification and classification are critical to the characterization of social networks. This, in turn, has wide range of applications in sociology, psychology, public health as well as product-marketing.

There are also immediate benefits/ applications of automated social relationships classification in smartphones. Smartphone users are not willing to divide contacts into different groups because there are usually hundreds of contacts in smartphone. Automatic relationship classification will help users to easily manage their contacts. Other potential applications include automated secretary service which instructs the smartphone to take different actions according to the relationship types between the smartphone user and the correspondent. For example, when a smartphone user is on vacation, the secretary service can forward ordinary calls to the voicemail while still be able to receive urgent calls from important/personal contacts. Based on the social-relationship between the sender and the recipient, the smartphone can also automatically prioritize incoming messages for reading/processing. The main contributions of our social relationship classification research can be summarized into three parts:

1) We conduct a social interaction data collection campaign to collect social interaction data from multiple communication channels and model the data with social interaction matrix.
2) We perform statistical analysis of the data and find interesting interaction patterns.
3) We apply machine learning based classification and dimension reduction algorithms on the social interaction matrix to predict relationship and save smartphone resources in data collection and classification.

We organize this paper as follows. In Section II, we review some related work about social network analysis. The problem formulation and social interaction data acquisition are described in Section III and Section IV. We pursuit a statistical data analysis for the data collected from our campaign in Section V and then describe and evaluate the various automated social relationship classification algorithms we proposed in Section VI. Section VII is the conclusion our work.

II. Related Work

There has been much research on the study of social networks. Topics relevant to this paper include community detection [2], [3], [4], [5], the modeling of social influence [6], [7], [8] and relationship strength or type classification. However, most of these works only assume a single type of relationship among the people in the social network. More recently, work on community detection in heterogeneous graphs starts to appear: Tang et al. [10] use Block Value Decomposition to model the interaction between different types of nodes in a multi-mode network. Their algorithm can classify the nodes in each mode into different groups. Tang et al. [12] study the problem of community detection in multi-dimensional networks. Sachan et al. [16] propose a generative model that can identify community in large social networks based on the topics they discuss, their interaction types and the types of connections between people. However, all of

978-1-4244-9529-0/13/$31.00 ©2013 IEEE

these works emphasize on node classification of a network. In contrast, our work on social relationship identification and classification focuses instead on classifying edges based on node-interaction patterns and the attributes of the terminating nodes of the edge.

Hangal et al. [7] use the weights of social ties to improve search in an online social network. Xiang et al. [13] build a model of relationship strength for online social networks by assuming that online relationship strength is determined by the profile similarity and the relationship strength itself can, in turn, influence the online interactions. In our work, we will take a further step by classifying the different types of relationships within a social network and then investigate how heterogeneous relationships within a network would impact its responses to different internal/external events.

Recently, several relationship classification algorithms and their applications have surfaced for analyzing online as well as physical-world social networks. MacLean et al. [14] report a system which can classify "friends" in an Email address book into different groups by assuming that people who co-exist in the same recipient-list should belong to the same group. Tang et al. [9] combine relationship classification with information propagation in online social networks. Their semi-supervised learning algorithm is based on the assumption that edges belong the same community should be of the same type. Eagle et al. [1] has classified relationship based on Bluetooth proximity data collected by mobile phones. They built a Gaussian Mixture Model to learn different relationships, namely, colleagues, outside friends, people within a user's circle of friends, within the community under study. Tang et al. [15] give a semi-supervised learning algorithm based the Partially-labeled Pairwise Factor Graph Model, to make relationship inference. However, all the above works try to predict one or two types of relationships based on the interaction data from one communication channel. None of them make use of multi-modal interaction data which are as rich as those available from the smartphones of the subjects. In our work, we will leverage the rich information derived from smartphone logs to classify multiple types of social relationships.

III. Problem Formulation for Relationship Classification

We propose to classify relationship with interaction data from different communication channels, namely phone call, Email, online SNS and physical location/proximity. We use social interaction matrix to model this multi-modal interaction data and firstly generate artificial data for motivation and illustration. Then we conduct a data collection campaign to obtain real-life, multi-modal interaction data. In particular, 25 participants, namely undergraduate, postgraduate students and staffs in our university, take part in the campaign. We achieve the interaction data of 7178 pairwise people, out of which 777 pairs have relationship labels. Their relationship types are manually labelled by the participants involved.

To model the interaction data from different communication channels, we use the social interaction matrix defined as follows:

$$M = [M_1, M_2, \cdots, M_n]. \quad (1)$$

Rows of M denote pairs of relationship and columns are the interaction features. Each submatrix M_i denotes interaction features extracted from one communication channel, which are the features we defined in the previous section. In our data set, we have interaction data from 4 communication channels. Then, M is formed by concatenating 4 submatrices, $M = [M_1, M_2, M_3, M_4]$, denoting the interaction features from phone call, Email, online social network and physical location/proximity. For example, suppose the i-th row of M denotes a pair of people A and B and the j-th column denotes the number of phone calls made on weekday. Then M_{ij} will be the number of phone calls between A and B that are made on weekday. The social interaction matrix can model the complicated, multi-modal interaction data into a simple matrix format. Then we can easily apply our classification and dimensionality reduction algorithms for further analysis. After formatting the interaction data into matrix, we apply *K-Nearest-Neighbor* (KNN) [21], Decision Tree [22] and *Support Vector Machine* (SVM) [19] for social-relationship classification. The corresponding performance of these classification algorithms will be discussed in Section VI.

Besides classification, we also use dimension reduction algorithms to embed the interaction data into lower dimensions so as to saving smartphone resources. In particular, we have applied *Principal Component Analysis* (PCA) [18] as well as CUR decomposition [25] on the interaction matrix M for this purpose. While PCA represents a set of data samples with correlated features using the linear combination of a set of k orthogonal principal components, CUR directly selects part of the rows and/or columns from the original interaction matrix for dimension reduction. In particular, CUR decomposes the social interaction matrix $M(m \times n)$ into three submatrices:

$$M = C \times U \times R. \quad (2)$$

The matrix C is composed by part of the columns from M. M. W. Mahoney et al. proposes a method to implement CUR in [25]. To select columns form matrix M, they set an importance score for each row by calculating the co-relation between each column and principal components. Then columns are selected according to their score randomly. Since CUR tries to pick "important" columns from the interaction matrix, it can help us to identify important features systematically, thus reduction data collection storage and battery power on the smartphone.

IV. Data Collection Campaign for Social Relationship Identification

To achieve real life interaction data, we carry out a social interaction data collection campaign, extracting communication data from 4 channels. The phone call data we collect from data collection participants includes phone number country code, call time, duration and direction. In Email interaction data, we extract sender address, recipient list, data and size of each Email. For online social network data, we require the data

collection participants to be active Facebook or Renren users. We use the standard protocol of these two site to access both the interaction data (comments or replies) and personal profiles with the authorization of the participants. We get the physical location/proximity interaction data by asking the participant to manually label where do they meet their contacts and the meeting frequencies. After data collection, we highlight the frequent contacts in each communication channel for the participant to give relationship labels, which will be used as ground truth in classification. We define 6 types of social relationship label, which are listed in Table I. These social relationships cover most of the social group of the participants. We ask them to assign one type of relationship label to each contact.

Relationship category	Relationship label	Description
Families	Fam	Family members, significant other, relatives.
Study/work	Pro	Professors, supervisors, teachers of the subject.
	Col	Colleagues, classmates, labmates of the subject.
	Stu	Students, subordinates of the subject.
Friends	Acq	Acquaintances, ordinary friends.
	Cfr	Close friends of the subjects.

TABLE I
DEFINITIONS OF 6 TYPES OF SOCIAL RELATIONSHIPS. WE USE (FAM, PRO, COL, STU, ACQ, CFR) TO DENOTE THEM FOR SHORT.

After collecting real-life data from data collection participants, we extract interaction features for the further relationship classification. In the phone call logs interaction data, the interaction features we define include "temporal" features which will identify the temporal patterns in phone call interactions. For example, we count the number of phone calls that are made on weekdays, weekends, daytime and night between the subjects and their contacts. We also extract "directional" features, i.e. originator vs. recipients for the interactions. We also extract personal profile information from SNS to establish "profile" features which are used to determine the similarity between a subject and a given contact. Lastly, we define features for face-to-face proximity interaction, including the meeting places (home, work and others) and frequency.

V. STATISTICAL ANALYSIS OF SOCIAL INTERACTION DATA

A. Coverage of social interaction data

The interaction data we have collected spreads in four communication channels, namely phone call, physical location/proximity, Email and online social network. We get this data from 25 student helpers in our university, including Hong Kong local/non-local, postgraduate/undergraduate students. The data covers their recent 3 to 6 months interaction. In our data collection, we totally get 7178 contacts from 25 data collection participants' recent interaction data (about 287 contacts for each participant). For each participant, we ask him/her to label 30 to 50 most frequent contacts in each communication channel. Finally, we get 777 labeled pairwise relationships. Dunbar have shown that the size of one's social group is about 150, which can be called *Dunbar Number* [23]. In this group, we usually have a more close social circle of size 30 to 50 people [24], which is similar to the number of labeled relationships in our data collection. So, the social relationships we extract can cover most of one's 150 social group. And most of the relationships in the close circle should be labeled.

B. Social relationships statistics

In our data collection campaign, we collect the recent interaction data of participants in four communication channels, which includes all kinds of social relationships. We divide social relationships into 6 types, in Table I, and ask the participants to manually label all frequent contacts. These labels serve as classification ground truth. Figure 1 shows the distribution of social relationship types among all of the labeled data. From this figure, nearly half of social relationships in our data collection are close friends. It reflects that college student like to spend more time interacting with their friends.

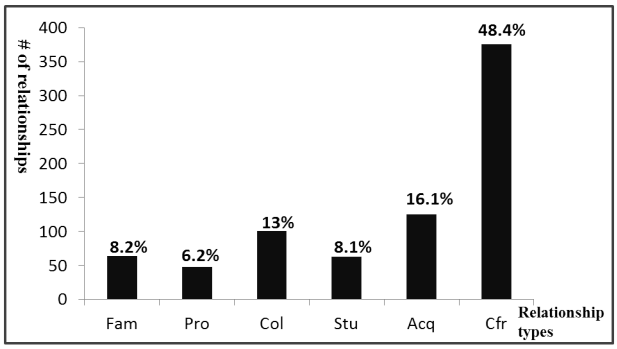

Fig. 1. Number of relationships identified for the 25 campaign participants. Meanings of relationship labels are defined in Table I

Intuitively, different relationships prefer different communication channels. For example, the participants usually choose to send Emails to professors or course tutors, while call their families and friends. The choice of communication channels sometimes can give indication about the type of relationships. In Figure 2, we show the average number of monthly conversations in different channels per subject. From Figure 2, postgraduate students have much more Email interactions than undergraduate students. These Emails are mainly exchanged with their work/study relationships. Undergraduate students typically have more physical meeting with friends. Phone call and physical meeting are the mainly communication channels between families. A lot of people that appear in one's online social network are acquaintances. So Online SNS seems to be a place to know new people.

C. Interaction patterns for different types of Relations

Among all of our contacts, we usually only stay in touch with just a few of them. Others are less frequent contacts. For example, in Figure 3, we show the cumulative distribution of

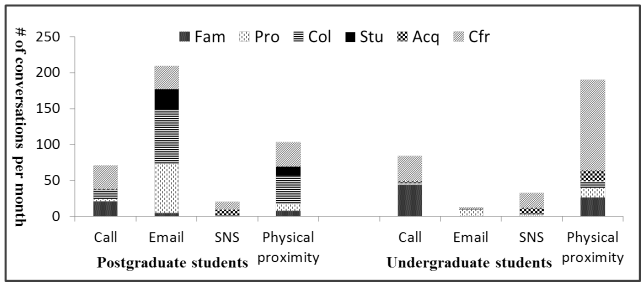

Fig. 2. Number of monthly conversations in each communication channel per subject. The data collection subjects are divided into postgraduate/undergraduate groups. The meanings of relationship labels are defined in Table I.

monthly conversation frequency in different communication channels. The horizontal axis is in logarithm scale. From this figure, nearly 50% relationships have more than 4 conversations of phone calls and Email every month. But only about 10% of them have more than 16 conversations.

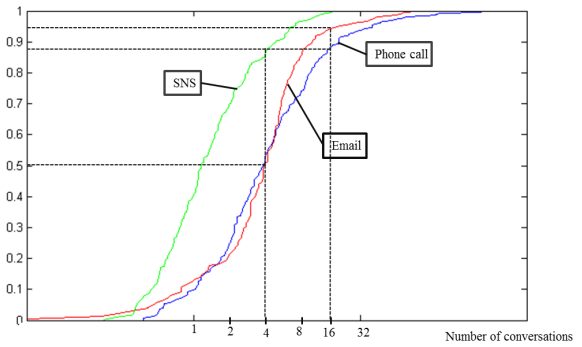

Fig. 3. Cumulative distribution of interaction frequency in one month, including three communication channels phone calls, Email and SNS. The horizontal axis is in logarithm scale.

In the social interaction matrix, we define temporal interaction features, such as the ratio of the number of phone calls that are made in weekday/weekend and daytime/night. We think that people tend to communicate with different relationships in different times of day and week. In addition to temporal interaction features, we also define "directional" features like incoming/outgoing phone calls, receiving/ sending Emails, etc. We know that social relationships are not symmetric such as seniors/youngers, boss/subordinates. As a result, we need to distinguish who is the communication "initiator" and who is "passive" recipient when classifying social relationships. For example, in our data set, Figure 4 shows the patterns of Email directional features. We define three directions in Email interaction, which include "sent", "received", "co-recipient". The Email conversations usually happen between work/study relationships (*pro, col, stu*). And compared to the colleague relationships, the participants receive more Emails from the supervisors. But they seldom co-exist in the same recipient list with the students they teach.

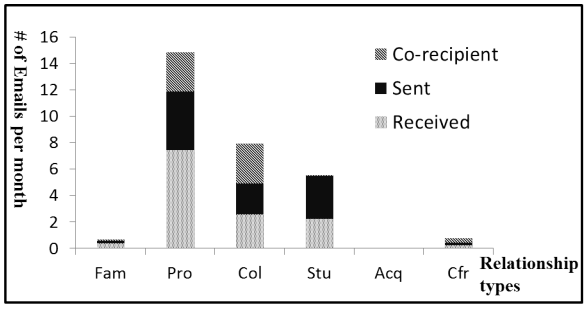

Fig. 4. No. of Emails sent/received/as co-recipients.

VI. Automatic Social Relationship Classification Based on Smartphone Interaction Data

In Section V, we have shown the statistical analysis of social interaction data, which motivate us to conduct social relationship classification on the social interaction matrix. In this section, we will test different algorithms in classifying social relationship types automatically based on the refinement of the social interaction matrix defined in Equation 1. We also apply dimensionality reduction techniques like PCA [18] and CUR [25] to reduce data complexity, refine feature definition and also help to save smartphone resources.

A. Comparison of different classification algorithms

We have applied three classification algorithms (KNN, decision tree and SVM) to predict the types of the pairwise social relationship between the smartphone user and each of his/her frequent correspondents. Figure 5 shows the *confusion matrix*, *accuracy*, *precision*, *recall* and *F-measure* of these algorithms. The meaning of relationship type labels are defined in Table I. For KNN, we use brute force method to find the best $k \in [1, 10]$ and we find that $k = 7$ results in the highest classification accuracy, 79.9%. We use C4.5 algorithm to build our decision tree and get the best *confidence factor* $c = 0.2$ by step search in the interval $[0.1, 0.5]$ with step size 0.1. The *minimum number objects* is set to 2. Then we get a classification accuracy 77.6% with decision tree. To use SVM, we follow the instructions in [19]. With the grid search method in [19], we have the SVM parameter $C = 512.0$ and $\gamma = 0.00195$. The classification accuracy of SVM is 82.4%.

Because of the large group of *Cfr* relationship (in Figure 2), other relationships can be easily misclassified to label *Cfr*, thus affecting the *recall* of classification (e.g. labels *Fam, Col, Acq*). Label *Col* (colleagues, classmates, etc.) suffers the most because it is hard to define a boundry between colleagues/classmates and close friends for college students. Sometimes, one pair of people has multiple relationship labels. Among the three algorithms, SVM outperforms KNN and decision tree in overall accuracy, average precision, recall and F-measure. For the remaining analysis in this section, we will focus on using SVM as the classification algorithm.

k-Nearest-Neighbor				PREDICTED CLASS					
ACTUAL CLASS	PRECISION	RECALL	F-MEASURE	*Fam*	*Pro*	*Col*	*Stu*	*Acq*	*Cfr*
Fam	0.875	0.656	0.75	42	1	0	0	3	18
Pro	0.667	0.75	0.706	0	36	4	0	0	8
Col	0.8	0.594	0.682	0	7	60	12	1	21
Stu	0.833	0.952	0.889	0	0	2	60	0	1
Acq	0.768	0.704	0.743	3	1	0	0	88	33
Cfr	0.805	0.891	0.846	3	9	9	0	20	335
AVERAGE	0.801	0.799	0.795	ACCURACY = 79.9%					

C4.5 Decision Tree				PREDICTED CLASS					
ACTUAL CLASS	PRECISION	RECALL	F-Measure	*Fam*	*Pro*	*Col*	*Stu*	*Acq*	*Cfr*
Fam	0.754	0.719	0.736	46	2	0	1	4	11
Pro	0.714	0.625	0.667	3	30	3	0	0	12
Col	0.734	0.574	0.644	0	6	58	8	0	29
Stu	0.851	0.905	0.877	1	0	2	57	0	3
Acq	0.75	0.744	0.747	1	0	1	0	93	30
Cfr	0.79	0.848	0.818	10	4	15	1	27	335
AVERAGE	0.773	0.776	0.773	ACCURACY = 77.6%					

Support Vector Machine				PREDICTED CLASS					
ACTUAL CLASS	PRECISION	RECALL	F-Measure	*Fam*	*Pro*	*Col*	*Stu*	*Acq*	*Cfr*
Fam	0.882	0.703	0.783	45	1	1	0	1	16
Pro	0.85	0.708	0.773	0	34	6	0	1	7
Col	0.775	0.614	0.685	0	4	62	10	2	23
Stu	0.861	0.984	0.919	0	0	0	62	0	1
Acq	0.805	0.76	0.782	2	1	0	0	95	27
Cfr	0.822	0.91	0.864	4	0	11	0	19	342
AVERAGE	0.823	0.824	0.819	ACCURACY = 82.4%					

Fig. 5. Classification results of KNN, decision tree and SVM.

B. Comparison of interaction data in different communication channels

We also compare the data from different communication channels. To achieve this, we discard one communication feature set at a time and compare the drop of classification performance. The greater the performance drops when a feature is left out, the more informative the feature is. Table II shows that location/proximity data plays a very important role in relationship classification: the accuracy drops more than 10% after eliminating location/proximity features. It indicates that the meeting location and frequency of a pair/group of subjects are key features for identifying their relationship. Besides, online social network interaction features also show its significance. The last row in Table II depicts the classification results after we eliminating all the "profile-related" features, i.e. the node-specific information, from the social interaction matrix. The low performance caused by their elimination means that background information of the subject (the node here) plays an important role in the relationship classification problem.

C. Dimensionality reduction on social interaction data

Besides relationship classification, we also apply dimension reduction algorithms on the data, which have some practical usage in our problem. For example, the classification on high dimensional data in smartphone need more precious smartphone resources, such as CPU, memory and storage. Low dimensional data will help to save smartphone resources in interaction data collection, storage and classification.

We observe a long tail distribution of eigenvalues for the social interaction matrix when *Principal Component Analysis* (PCA) is performed. This indicates that we can use a small number of eigenvectors to capture majority of the variance in the data. The classification accuracy of low dimensional data with SVM is shown in Figure 6. The accuracy increases as we use more dimensions for classification. But there is clear diminishing of return beyond the 9th principal components. In Figure 6, the dash line shows the accuracy of classifying full feature data (without dimensionality reduction), which is 82.4%.

Data sets	Accuracy	Recall	Precision	F-measure
All features	82.4%	82.3%	82.4%	81.9%
Without calls features	80.1%	80.5%	80.1%	79.4%
Without L/P features	71.8%	71.8%	71.8%	70.8%
Without Email features	79.0%	79.0%	79.0%	78.3%
Without SNS features	75.4%	75.0%	75.4%	74.7%
Without "profile" features in SNS	76.3%	76.3%	76.3%	75.2%

TABLE II
COMPARE THE DROP OF PERFORMANCES BY ELIMINATING ONE CHANNEL FEATURE SET AT A TIME. (L/P = LOCATION/PROXIMITY).

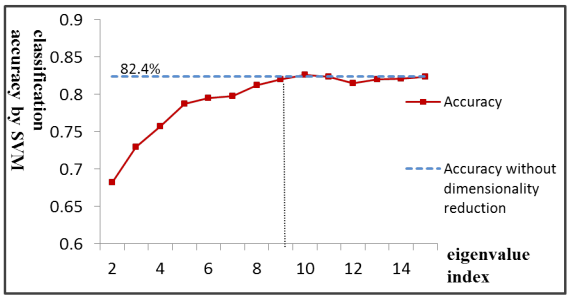

Fig. 6. The classification accuracy of social interaction data after dimensionality reduction.

We also apply CUR decomposition on our social interaction matrix to see how feature selection by CUR impacts the classification results. We use the CUR method in [25], set the rank parameter of CUR to $k = 9$ and the number of rows to be maximum, $r = 777$. The relationship classification results after CUR are shown in Table III, with a comparison to PCA. The classification performance of CUR drops as we reduce the number of features (i.e. columns) included in the social interaction matrix. Because CUR discards some original features, while PCA will not. Then it is therefore reasonable that PCA outperforms CUR when reduced to low dimension. However, the advantage of CUR is that it helps us refine feature definition and conserve smartphone battery in interaction data collection. CUR decomposition can help to determine a proper time resolution for activating smartphone for interaction data collection.

D. Considerations of user privacy

Privacy is one of the most important concerns in designing smartphone applications. In our work, user privacy can be

# of columns in the matrix	65	30	20	13
PCA	82.4%	83.0%	81.9%	82.0%
CUR	82.4%	79.7%	78.8%	77.7%

TABLE III
COMPARISON OF CLASSIFICATION ACCURACY AFTER PCA AND CUR DECOMPOSITION (ORIGINAL DATA DIMENSION IS 65).

protected when deploying the social relationship classification application into real world smartphones. In this application, there will be a model training stage, where we need to collect some interaction data to train our classifier. This stage needs to be done in the backend server because it needs a lot of computation. Data collected from a representative set of friendly users, with explicit manual labels, can be used to train model which is applicable for the general users. Once the model has been trained, the training data will be deleted and no more data needs to be shipped to the server. The smartphones will receive the classification model and perform relationship classification locally. The relationship classification results are only used by the smartphone user locally. So, for the normal users of this application, their interaction data will only be used in their own smartphone, thus ensuring the security of personal interaction data.

E. Conserving smartphone resources

Comparing to personal computers (PC), smartphones usually have relatively limited resource, such as CPU, memory, storage, battery power and even the Internet data service. In our application, smartphones need to perform interaction data collection, storage and classification. The resources consumed in each stage are directly related to the dimensions of data. High dimensional data brings burdens to the smartphone. Suppose the original data dimension is n and the reduced dimension is $k(k < n)$. PCA embeds the interaction data from $O(n)$ into $O(k)$ dimensions for storage and classification, with little accuracy penalty. But the smartphones still need to collect all the interaction features for dimension reduction. CUR algorithm helps us to refine feature definition by removing some of the columns out of the social interaction matrix, thus reducing the dimension of original data. In fact, one can use the measurement data of different time resolution (i.e. columns picked by CUR). Then we can turn on the sensors and collect interaction data in a proper time resolution. In CUR, the interaction features can be reduced from $O(n)$ to $O(k)$, but it suffers accuracy penalty from 82.4% to 77.7% when $k = 13$. In real life application, we should also carefully select k for PCA and CUR to find a best trade off between saving resources and classification accuracy.

VII. CONCLUSION

In this paper, we propose to classify social relationship based on the interaction data from multiple communication channels in smartphone. We carried out a social interaction data collection campaign to collect real life interaction data and model it with social interaction matrix. In the statistical analysis, we found that the interactions between people show temporal, directional pattern, etc. In our relationship classification problem, SVM outperforms KNN and decision tree (accuracies are 82.4%, 79.9% and 77.6% respectively). The interaction features from online social network and physical location/proximity contribute more to the classification results. At last, with PCA, we embed the data from 65 to 9 dimensions while preserving high classification accuracy. We also use CUR decomposition to help us refine feature definition and save smartphone energy in data collection.

REFERENCES

[1] N. Eagle, Alex S. Pentland, "Reality Mining: Sensing Complex Social Systems", *Pers Ubiquit Comput* 2006,10: 255-268.
[2] J. Taylor, D. M. Wilkinson, B. A. Huberman, "Email as Spectroscopy: Automated Discovery of Community Structure within Organizations," *The Information Society*, 21: 133-141, 2005.
[3] G. Palla, I. Dernyi, I. Farkas, T. Vicsek, "Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society," *Nature*, 435: 814-818, 2005.
[4] A. Lancichinetti, S. Fortunato, "Community Detection Algorithms: Acomparative Analysis", *PHYSICAL REVIEW E* 80, 056117 2009.
[5] J. Leskovec, Kevin J. Lang, Michael W. Mahoney, "Empirical Comparison of Algorithms for Network Community Detection", *WWW* 2010.
[6] W. Pan et al, "Modeling Dynamical Influence in Human Interaction Patterns", *IEEE Signal Processing Magazine*, March 2012.
[7] S. Hangal et al, "All Friends are Not Equal: Using Weights in Social Graphs to Improve Search", *SNA-KDD* 2010.
[8] A. Madan, A. S. Pentland, "Modeling Social Diffusion Phenomena Using Reality Mining", *AAAI Spring Symp.*, March 2009.
[9] S. Tang et al, "Relationship Classification in Large Scale Online Social Networks and Its Impacts on Information Propagation", IEEE infocom 2011.
[10] L. Tang, H. Liu, J. Zhang, Z. Nazeri, "Community Evolution in Dynamic Multi-mode Networks", *KDD* 2008.
[11] Y. R. Lin et al, "MetaFac: Community Discovery via Relational Hypergraph Factorization", *KDD* 2009.
[12] L. Tang, X. Wang, H. Liu, "Community Detection in Multi-dimensional Networks", *Technical report*, Arizona State Univ., 2010.
[13] R. Xiang, J. Neville, M. Rogati, "Modeling Relationship Strength in Online Social Networks", *WWW* 2010.
[14] D. MacLean et al, 'Groups Without Tears: Mining Social Topologies from Email", *IUI* 2011.
[15] W. Tang, H. Zhuang, J. Tang, "Learning to Infer Social Ties in Large Networks", *Procs. of ECML and KDD* 2011.
[16] M. Sachan et al, "Using Content and Interactions for Discovering Communities in Social Networks", *WWW* 2012.
[17] N. Eagle, A. S. Pentland, "Eigenbehaviors: Identifying Structure in Routine", *Behav Ecol Sciobiol* 2009, 63:1057-1066.
[18] L. I. Smith, "A Tutorial on Principal Components Analysis",http://www.sccg.sk/haladova/principal_components.pdf, 2002.
[19] C. W. Hsu, C. C. Chang, and C. J. Lin, "A Practical Guide to Support Vector Classification", *Technical Report, National Taiwan University*, April 15, 2010.
[20] http://www.csie.ntu.edu.tw/ cjlin/libsvm/
[21] http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm
[22] J. R. Quinlan, "C4.5: Programs for machine learning", *Morgan Kaufmann Publisher*, 1993.
[23] R. Dunbar, "You've Got to have 150 Friends", *The New York Times*, The Opinion Pages, 2010.
[24] R. Dunbar, "How Many Friends Does one Person Need? Dunbar's Number and Other Evolutionary Quirks", *Harvard Univ. Press,* 2010.
[25] M. W. Mahoney, P. Drineas, "CUR Matrix Decompositions for Improved Data Analysis", *The National Academy of Sciences of the USA*, 2009.