Predicting Human Contacts in Mobile Social Networks using Supervised Learning *

Kazem Jahanbakhsh
Computer Science Department
University of Victoria
B.C., Canada
jahan@cs.uvic.ca

Valerie King
Computer Science Department
University of Victoria
B.C., Canada
val@cs.uvic.ca

Gholamali C. Shoja
Computer Science Department
University of Victoria
B.C., Canada
gshoja@cs.uvic.ca

ABSTRACT

Having access to human contact traces has allowed researchers to study and understand how people contact each other in different social settings. However, most of the existing human contact traces are limited in the number of deployed Bluetooth sensors. In most experiments, there are two types of participants, the ordinary ones who carry cellphones and a specially selected group who additionally carry sensors. Although the contacts between any pair of participants are known when at least one of them carry a sensor, the contacts between any pair of participants are "hidden" when both of them carry their cellphones. In this paper, we employ two well-known supervised classifiers for predicting hidden contacts among participants who carry their cellphones. The performance results of our supervised classifiers show the applicability of using machine learning algorithms for contact prediction task. The results also show that a small subset of features such as number of common neighbors and total overlap time play essential roles in forming human contacts. Finally, we show that contacts of nodes with high centralities are more predictable than nodes with low centralities.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology —*Classifier design and evaluation, Feature evaluation and selection*

General Terms

Algorithms; Experimentation.

Keywords

Human Mobility, Contact Graph, Machine Learning, Supervised Classifier, Prediction, Degree Centrality.

*This work was supported by grants from Natural Sciences and Engineering Research Council (NSERC) of Canada.

Simplex '12, April 17 2012, Lyon, France.

1. INTRODUCTION

In this paper, our main goal is to explore the possibility and benefits of using supervised learning algorithms for predicting missing contacts in existing contact traces. We say two people are in *contact* if they happen to be in close proximity of each other ($< 10m$). Proposing accurate models which can explain how people contact each other in different social environments may not be feasible unless we have access to large size of human contact traces collected from different social settings. During the last few years, researchers have started collecting human contact traces by distributing a number of Bluetooth sensors among a set of social groups [3, 5, 11]. Although these datasets have helped researchers understand human mobility better, most of them suffer from limitation of practical number of sensors which were distributed among experiment participants.

In most experiments, participants are of two types, the ordinary ones who carry their cellphones and a specially selected group who additionally carry wireless sensors. Contact data can only be collected by the sensor-carrying participants, and thus only contacts involving at least one sensor-carrying participant can be detected. The other contacts are "hidden." Furthermore, previous experiments contained a large number of recorded contacts with people who carried their own Bluetooth-enabled devices such as cellphones [7]. Such collected contact datasets can be potentially expanded if one can predict the missing contacts among people who did not carry any sensors. The main motivation for this work is to address the problem of predicting social behaviour of a population where we have only direct observable information about a subset of the population.

For predicting the hidden contacts, we employ a supervised learning approach in which we use training data to devise a classifier function for predicting the classes of unseen data. First, we extract several features by using information from the underlying structure of the *contact graph* (i.e. the graph in which nodes are people and edges are contact events between them), social profiles of people, and static sensors. We use two supervised learning classifiers namely *Logistic Regression* and *K-Nearest Neighbor* for predicting the hidden contacts. We validate our classifiers by taking two different approaches as described in sections 4.4.1 and 4.4.2. Finally, we examine the effect of nodes' *degree centralities* i.e. the total number of contacts a node had during a social event on their predictability levels. The contributions of this paper are as follows:

1. We show the applicability of using supervised learn-

ing algorithms by taking two different approaches for validating our classifiers.

2. We demonstrate that the *number of common neighbors* and the *total overlap time* are the most significant features in contact prediction.

3. Finally, we show that contacts of nodes with high centralities are more predictable than nodes with low centralities.

2. RELATED WORK

In [6], we introduced a weighted *contact graph* in which nodes represented people and there was an edge between two nodes if they had at least one contact with one another during the event. We also assigned a weight to each edge showing the overall time that the corresponding end-nodes spent together during the social event. Having a partial weighted contact graph, we devised several methods for predicting the missing parts of contact graphs. In [7], we extracted several features for inferring the missing contacts in different social environments. We evaluated the performance of each feature in predicting missing contacts. We showed that combining the number of common neighbors feature with social data of people provides the best prediction results [7].

In this paper, on the other hand, we extract new features by using nodes degrees, temporal information of contact graphs (e.g. the total overlap time), and information of static sensors. More importantly, we employ two supervised learning algorithms for predicting missing contacts. The supervised algorithms enable us to combine all of our features in a systematic way. The machine learning area also provides a framework for evaluating the significances of the extracted features in different social events.

Nowell et al. studied the link prediction problem in a citation network where they extracted several features for predicting future collaborations among researchers [9]. Hasan et al. extended the Nowell's work by employing several supervised learning classifiers [4]. They showed that supervised learning is an efficient approach for addressing the link prediction problem. Leskovec et al. also used the logistic regression to predict the sign of links in social networks [8].

Song et al. studied the limits of predictability in human mobility by studying the mobility patterns of cellphone users [10]. They discovered a high degree of regularity in human mobility resulting in a potential 93% predictability in user mobility. Vu et al. exploited the high level of regularity in human mobility in order to predict locations where a person will go and the people she will contact in a given time of a day [12]. Wang et al. used mobile phone data and found a strong correlation between individuals' movements and their connectedness in their social network [13].

3. PROBLEM DEFINITION

To present the problem formally, let us denote the set of sensors with V_{int} where sensors are considered as *internal nodes*. We also designate the set of *external nodes* that are nodes which do not carry any sensors with V_{ext}. We assume that two nodes are in *contact* at time t if they happen to be in close proximity of each other at t. As a result, we can model the human mobility by translating each contact between two nodes such as u and v into an undirected edge between them. It is clear that these edges are dynamic as they appear and disappear over time. Thus, we divide the experiment time into equal intervals of τ seconds called *time intervals*. We choose $\tau = c \times T$ where c is a constant integer, and T is the *inquiry interval* of wireless sensors, namely the time gap between two consecutive sensings. The coefficient c is usually chosen to be one or two.

Let $\Lambda_k = [t_0 + k\tau, t_0 + (k+1)\tau]$ denote the k^{th} time interval where $0 \leq k < k_{max}$ and t_0 is the starting time of the experiment. We show people's interactions during the k^{th} time interval with an undirected contact graph G_k that contains contacts between people in Λ_k. In $G_k = (V_k, E_k)$, edges in $V_{int} \times (V_{int} \cup V_{ext})$ are known while edges in $V_{ext} \times V_{ext}$ are missing. Our objective is to predict these missing edges.

4. CONTACT PREDICTION USING CLASSIFICATION ALGORITHMS

In this section, we review two supervised classification algorithms by which we formulate the relationship between a dependent variable (i.e. output variable) and one or more independent variables (i.e. features). In our case, we use these classifiers to estimate the probability of a contact between a pair of external nodes as a function of their feature vector.

4.1 Logistic Regression Overview

There are many problems where we want to find the class which an item belongs to. In our problem, each pair of external nodes such as (u, v) can belong either to an *edge* or to a *non-edge* class. Thus, we want to find the probability that a given pair of external nodes belongs to *edge* class that is the probability that a contact actually happened between them. This can be formulated as a binary classification problem where the output variable $y \in \{0, 1\}$. In particular, Logistic regression can be used to formulate our problem where the hypothesis function satisfies the $h_\Theta(x) \in [0, 1]$ condition. In particular, we choose the hypothesis function as below [2]:

$$h_\Theta(x) = g(\Theta^T X) = \frac{1}{1 + e^{-\Theta^T X}}, \tag{1}$$

where $g(.)$ is the logistic function, X is the feature vector, and Θ is the parameter of the model that we want to learn. If we show the *edge* class with one and the *non-edge* class with zero, we can compute the contact probability between u and v using the logistic regression as follows:

$$p(y = 1|X; \Theta) = h_\Theta(x) = 1 - p(y = 0|X; \Theta) \tag{2}$$

We compute the likelihood of the parameter Θ as follows:

$$L(\Theta) = p(\vec{y}|X; \Theta) = \prod_{i=1}^{m} p(y^{(i)}|X^{(i)}; \Theta), \tag{3}$$

where m is the size of training set. We find the parameter Θ such that it maximizes the likelihood $L(\Theta)$.

4.2 K-Nearest Neighbor Overview

Another method that we use for classification is the K-Nearest Neighbor method (i.e. KNN). We employ this method to estimate the probability distribution of edge existence between external node pairs given their feature vectors. KNN

is a non-parametric estimator where it does not make any assumptions about the probability distribution function. However, logistic regression makes specific assumption about the form of the logistic function. In KNN, we have all of our training points in a d-dimensional space where d is the number of features. When we want to find out the label of a given external pair such as (u, v), we first find the K nearest neighbors of (u, v) in the feature space using the Euclidean distance. Then, we classify (u, v) by returning the class that majority of its K nearest neighbors belong to [2].

4.3 Features Extraction

While configuring a supervised learning classifier, we must explore all important features that might have influence on the output variable that we wish to predict. Here, we need to explore all possible independent variables that have impact on the probability of contacts between two external nodes in the time interval Λ_k.

4.3.1 Contact Graph-based Features

First, we focus on features that are based on local properties of contact graphs. In particular, for a given time interval Λ_k we construct the partial contact graph G_k. Next, we extract several features from the structure of G_k in order to predict the probability of contacts. For $G_k = (V_k, E_k)$, let $N^k(u)$ denote the neighborhood set of node u that contains all nodes which had at least one contact with node u in Λ_k:

$$N^k(u) = \{v | (u, v) \in E_k\} \tag{4}$$

Using the neighborhood set of node u, we devise several degree based features. First, we see that degree of a node $u \in V_k$ (i.e. $|N^k(u)|$) represents the number of contacts that node u has had during Λ_k. We assume that if node u has a high number of contacts, it is more likely to contact a randomly chosen node than a node v with low number of contacts. As a result, for a given pair of external nodes such as (u, v) we use $|N^k(u)|$ and $|N^k(v)|$ as the first two degree-based features. Moreover, for a given pair of nodes such as (u, v) we assume that the contact probability between them not only depends on their individual degrees, but it also depends on the product of their degrees in Λ_k. Therefore, we choose $|N^k(u)| \times |N^k(v)|$ as the third feature.

If two nodes u and v are in close proximity of each others, they are likely to contact each other in the near future. It has been previously shown that the number of common neighbors between a pair of nodes can be employed to estimate the geographical distance between them [7]. We similarly use the number of common neighbors between two nodes as the fourth feature that is $ncn(u, v) = |N^k(u) \cap N^k(v)|$.

4.3.2 Contact Duration

Contact duration is another important feature that is useful for prediction task in addition to number of contacts. As it has been shown in Figure 1, not only the fact that two nodes such as u and v have seen the same node w influences the probability of a contact between u and v, but the overlap time that node w has been in contact with u and v also matters (i.e. $\Delta_{ov}(u, v; w) = t_{12} - t_{21}$). As a result, we define the total overlap time feature as shown in Equation 5 where for each pair of external nodes we compute the total overlap time that these two nodes have spent with the same third nodes in Λ_k.

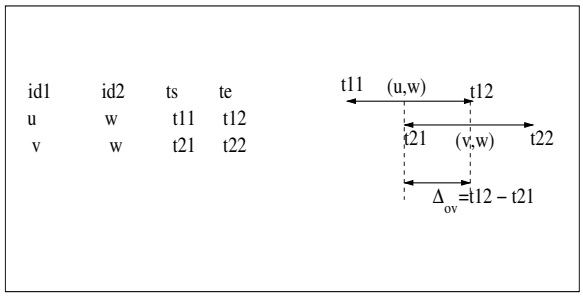

Figure 1: Contact duration as a feature.

$$T^k_{ov}(u, v) = \sum_{w \in (N^k(u) \cap N^k(v))} \Delta^k_{ov}(u, v; w) \tag{5}$$

4.3.3 Social Information

The third class of features that we use in our feature vector is the social similarity between nodes. Our main intuition is that mobile nodes are more likely to contact other nodes that are socially similar to them. In Infocom 2006's data, participants reported a brief version of their social profiles including their affiliations, research interests, country of birth and so on. Let us define the *social focus* as a set of people who share the same research interest, speak the same language, or were born in the same country. Foci are a way of summarizing many possible reasons that two people contact each other: because they are from the same country, have the same affiliation, or share the same interest. We define the *Foci distance* between two given nodes as the cardinality of the smallest social focus that both of them belong to. We derive the Foci distance between two given nodes u and v as below [7]:

$$d_{foc}(u, v) = min\,|\{F | u, v \in F\}|\,, \tag{6}$$

where F is the social focus that both u and v belong to. Considering Equation 6, we define the *Foci Similarity* between two nodes as follows:

$$sim^{soc}_{foc}(u, v) = \frac{1}{d_{foc}(u, v)} \tag{7}$$

4.3.4 Static Sensors

Infocom 2006 dataset includes information from static nodes which were deployed in different conference rooms to detect mobile nodes in that room. These static nodes have longer radio ranges than mobile sensors. We use the data from static nodes as another feature for our learning algorithm. We add a new feature which just counts the number of common static nodes which have been seen by a pair of external nodes in Λ_k. This feature basically tells us if the corresponding mobile nodes are in the same room or not.

4.4 Training/Validating the Classifiers

Since in human contact traces we do not have any information about the contacts between external nodes, there is not any way for us to validate the predicted contacts between them. To get around this issue, we choose a random subset of internal nodes and label them as *external*

Table 1: Real Data Description

Dataset	Inf 05	Inf 06	Roller
No. of Sensors	41	79	62
Length	3 days	4 days	3 hours
Scanning period	120 sec	120 sec	15 sec
No. of Ext. Nodes	206	4321	1050

surrogates (i.e. V_{surext}). These external surrogates play the same role as external nodes. We remove all contacts observed by external surrogates that are all edges such as $(u, v) \in V_{surext} \times V_{surext}$ [7]. We generate the partial contact graphs by removing edges among external surrogates. We use these partial contact graphs to train and test our classifiers. Next, we describe two approaches for training and testing our classifiers.

4.4.1 Approach I

In the first approach, we test the possibility of using a contact dataset such as A as the training data in order to predict the missing contacts for another contact dataset B. Using Infocom 2005, Infocom 2006, and Rollernet datasets we examine how accurately we can do the prediction if for instance we use the Infocom 2005 as the training data while using the Infocom 2006 as the test data.

4.4.2 Approach II

In the second approach, we use the well known *k-fold cross validation* technique in which we use part of the dataset as the test data while the rest of data is used as the training data. This way we train our predictor using the training data and then we use the test data to evaluate the learned algorithm.

5. PREDICTION RESULTS

In this section, we present our prediction results using logistic regression and KNN classifiers. We use the Weka software for testing the chosen classifiers [1]. For training and validating our classifiers, we use the two approaches described in the previous section. For our contact datasets, we use Info 05 and Info 06 datasets that were collected from Infocom conferences in 2005 and 2006, respectively [5]. We also use Rollernet dataset (i.e. Roller) containing the contacts from a set of people who participated in a rollerblading tour in Paris [11]. In all of these datasets, they distributed a limited number of sensors among a subset of people who attended the event. All datasets include the recorded contacts by sensor devices (iMotes). Each recorded contact includes the ID of the sensor, the ID of the device which was seen by the sensor, and the start and the end time when the two devices where in the close proximity of each other. Table 1 describes the properties of the datasets.

In both Infocom 2005 and 2006, we use the data collected on the first day of the conference. We choose the time interval (τ) to be 240, 240, and 30 seconds for Infocom 2005, 2006, and Rollernet datasets respectively. For logistic regression, we use 0.5 as our threshold where we classify each pair of external surrogates with a predicted probability greater than the threshold as an *edge*. For KNN, we choose $K = 3$. In all of our experiments, we label 60% of nodes as external surrogates. We repeat each experiment ten times with different subsets of external surrogates and show the averages.

Table 2: Approach I's performance results (Logistic Regression/KNN)

Training Data	Info 05	Roller	Roller
Test Data	Info 06	Info 05	Info 06
TPR	0.29/0.31	0.39/0.38	0.35/0.41
FPR	0.05/0.11	0.08/0.11	0.07/0.09
Correctly classified	79%/75%	75%/72%	80%/77%
RMSE	0.44/0.42	0.48/0.45	0.44/0.40

Table 3: Approach II's performance results (Logistic Regression/KNN)

Session Type	Keynote	Lunch	Coffee
TPR	0.18/0.24	0.37/0.40	0.41/0.43
FPR	0.03/0.08	0.04/0.07	0.02/0.02
Correctly classified	81%/78%	84%/81%	92%/92%
RMSE	0.42/0.40	0.39/0.36	0.26/0.24

5.1 Approach I's Results

In approach I, we use the Infocom 2005, 2006, and Rollernet datasets. We choose $k = 81$, $k = 15$, and $k = 32$ for Infocom 2005, 2006, and Rollernet respectively. Table 2 shows the performance results for the logistic regression and KNN classifiers. As we can see, both classifiers outperform a random predictor. Thus, we can conclude that our three datasets have similar structures such that we can train our classifiers by using one of them while using the other one as the test data. Moreover, we observe that using the Rollernet as the training data provides better prediction results than using the Infocom 2005.

5.2 Approach II's Results

In a conference setting, different events happen during each day such as keynote talks, panels, coffee/lunch breaks, and regular sessions. Here, we just focus on three different types of sessions including keynote, lunch, and the last coffee break. In our second approach, we use the *k-fold cross validation* method to evaluate our predictors. We use the Infocom 2006 dataset. We partition the V_{surext}-induced subgraph of CG_k into five different subsets. We repeat our experiment five times where each time we use one subset of the induced subgraph as the test data and the rest as the training data. Finally, we compute the overall performance by computing the average of all trials. Our 5-fold cross validation results for logistic regression and KNN classifiers are shown in Table 3.

As we can see, the true positive rate (i.e. TPR) is low. This is because there is not enough information in the extracted features. Moreover, we have found several instances of external pairs where there are not any internal nodes around them. In this case, there is not much information for our classifiers to use for prediction. Even if the TPR is not high, we see that the false positive rate (i.e. FPR) is low.

5.3 Features Significance

In this part we compare the significance of different features described in subsection 4.3. This is important because we can find out what features play important role in contact prediction. We use the Infocom 2006 as our dataset and choose 60% of nodes as external surrogates. We extract the

Table 4: The average rank for different features (Infocom 2006)

Session Type	Keynote	Lunch	Coffee
degree	4	5	5
degree	7	7	7
degree product	3	3	6
ncn	1	1	2
total overlap	2	2	1
social	5	6	4
ncsn	6	4	3

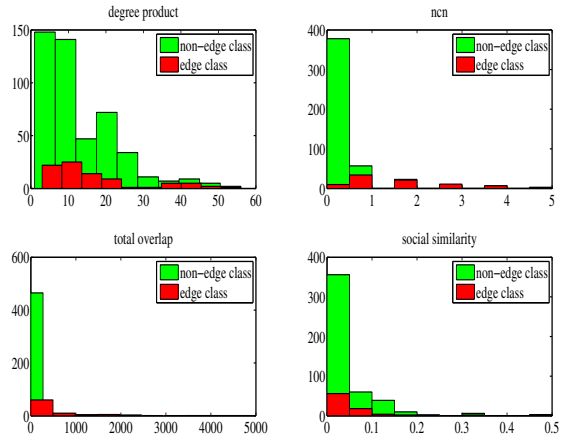

Figure 2: Class density distributions for lunch session (Infocom 2006)

feature vectors for all possible pairs of external surrogates in our test data. We use this test data to rank different features. It is important to note that for each type of session we repeat our experiment ten times where in each trial we randomly pick 60% of internal nodes and label them as external surrogates. For ranking the features, we use three different algorithms including information gain, gain ratio, and Chi-Square. We have shown the average ranks of different features for the three types of sessions in Table 4. As we can see, the number of common neighbors and the total overlap time are the most significant features in all sessions. This is because both of these features contain the geographical proximity data. The product of degrees and the number of common static nodes also appear as the next important features.

We can also evaluate the importance of our features for different types of sessions by using their class density distributions. Because of the space limitation, we have only shown the class density distributions of the four most important features for lunch break in Figure 2. One interesting pattern that we have found in the class density distribution is that *edge* and *non-edge* classes become distinguishable when the number of common neighbors and the total overlap time increase. We observe almost the same pattern for degree product and social similarity but these later features are not as significant as the former ones.

We have seen that the number of common neighbors is the most significant feature in contact prediction task. Now, we would like to compare the performance results of our

Table 5: Performance results of NCN linear classifier (Infocom 2006)

Session Type	Keynote	Lunch	Coffee
TPR	0.22	0.39	0.29
FPR	0.05	0.04	0.01
Correctly classified	80%	84%	92%

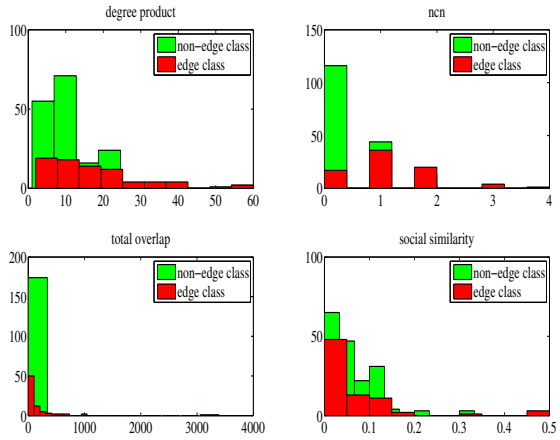

Figure 3: Class density distribution of external nodes with highest centrality (Infocom 2006: keynote)

previous supervised classifiers with a linear classifier that only uses the number of common neighbors for prediction task. We designate a threshold for *ncn* and classify all pairs that have an *ncn* value greater than the threshold in the *edge* class and the rest of pairs in the *non-edge* class. We choose the threshold to be two. The results are shown in Table 5. Comparing Table 5 with 3, we observe that using all features do not give us a significant improvement over the linear classifier that uses only the ncn feature. This again shows the significant role of the ncn feature in predicting the missing contacts.

5.4 Centrality Effect on Predictability

One interesting question that can be asked is what nodes are more predictable. We would like to select external nodes such that their contacts become more predictable. In particular, we want to study the effect of degree centrality of nodes on their predictability. Let us define the degree centrality of node u in time interval Λ to be the total number of contacts that node u has had during Λ. We use the dataset of Infocom 2006 and focus on the first day of the main conference. Assuming that we have the full information of all nodes, we compute the centralities of all internal nodes using the entire data of the 9-hour period. We sort all nodes according to their centralities in a descending order.

For our experiment, we first choose the top 30% of nodes from the sorted list as the external nodes with the highest centralities. Secondly, we choose the bottom 30% of nodes from the sorted list as the least central external nodes. For prediction purpose, we just focus on the keynote session. Figures 3 and 4 show the class density distributions when external surrogates have been chosen to be the most and

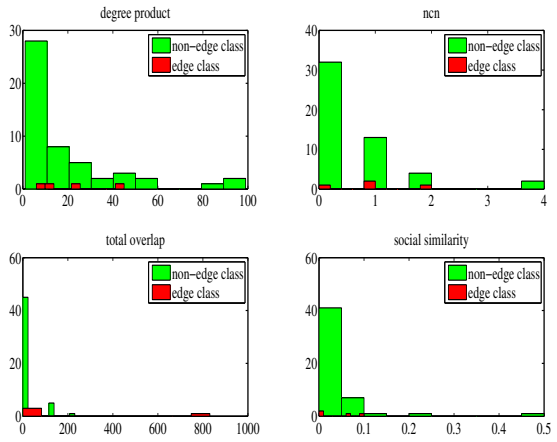

Figure 4: Class density distribution of external nodes with lowest centrality (Infocom 2006: keynote)

Table 6: The effect of centrality on predictability

External Type	Most Cent.	Least Cent.
TPR	37%	0%
FPR	15%	0%
Accuracy	70%	92%

the least central nodes, respectively. We can see that choosing external surrogates from the most central nodes makes the class density distribution of *edges* to be more distinctive form *non-edges* than when we choose the least central nodes as external surrogates.

Thus, we expect that nodes with high centrality to be more predictable than nodes with less centralities. The reader should note that if one chooses the most central nodes as external nodes, then she would have a higher number of pairs that fall in the *edge* class which in return helps the classifier for better prediction results. We have used KNN with $K = 3$ with 10-fold cross validation to see how accurately we predict if we choose external nodes differently. The performance results of two approaches have been shown in Table 6. We observe that choosing nodes with less centralities as sensors helps classifiers achieve better prediction results. These results are important because they illuminate how practitioners should choose the sensor nodes for sampling human contacts in order to achieve better predictions.

6. CONCLUSIONS

In this paper, we have employed the logistic regression and KNN classifiers from the machine learning area for predicting missing contacts. We have done this by extracting a set of different features. Interestingly, we have demonstrated that it is possible to use a contact dataset A as the training data in order to predict the missing contacts of a different dataset B. We have also shown that the number of common neighbors and the total overlap time play the most significant roles in predicting human contacts. Finally, we have shown that nodes with high centralities provide better prediction accuracy than nodes with low centralities. For our future work, we have plan to employ more sophisticated machine learning techniques such as Support Vector Machine in order to achieve better prediction results.

7. ACKNOWLEDGEMENTS

The authors would like to thank Yumi Moon for her editing assistance.

8. REFERENCES

[1] Weka 3 - data mining with open source machine learning software, 2012.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 106(36):15274–15278, July 2009.

[4] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006.

[5] P. Hui and J. Crowcroft. How small labels create big improvements. In *Proceedings of the Fifth IEEE International Conference on Pervasive Computing and Communications Workshops*, PERCOMW '07, pages 65–70, Washington, DC, USA, 2007. IEEE Computer Society.

[6] K. Jahanbakhsh, G. C. Shoja, and V. King. Human contact prediction using contact graph inference. In *International Symposium on Social Computing and Networking*, pages 813–818. IEEE, December 2010.

[7] K. Jahanbakhsh, G. C. Shoja, and V. King. Predicting missing contacts in mobile social networks. In *World of Wireless Mobile and Multimedia Networks*, pages 1–9. IEEE, June 2011.

[8] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 641–650, New York, NY, USA, 2010. ACM.

[9] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58:1019–1031, May 2007.

[10] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, Feb. 2010.

[11] P. U. Tournoux, J. Leguay, F. Benbadis, V. Conan, M. Dias de Amorim, and J. Whitbeck. The Accordion Phenomenon: Analysis, Characterization, and Impact on DTN Routing. In *IEEE INFOCOM 2009 - The 28th Conference on Computer Communications*, pages 1116–1124. IEEE, april 2009.

[12] L. Vu, Q. Do, and K. Nahrstedt. Jyotish: Constructive approach for context predictions of people movement from joint wifi/bluetooth trace. *Pervasive and Mobile Computing*, 7(6):690 – 704, 2011.

[13] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011. ACM.