Predicting Missing Contacts in Mobile Social Networks

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

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

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

***Abstract*—Experimentally measured contact traces, such as those obtained in a conference setting by using short range wireless sensors, are usually limited with respect to the practical number of sensors that can be deployed as well as available human volunteers. Moreover, most previous experiments in this field are partial since not everyone participating in the experiment is expected to carry a sensor device. Previously collected contact traces have significantly contributed to development of more realistic human mobility models. This in turn has influenced proposed routing algorithms for Delay Tolerant Networks where human contacts play a vital role in message delivery. By exploiting time-spatial properties of contact graphs as well as popularity and social information of mobile nodes, we propose a novel method to reconstruct the missing parts of contact graphs where only a subset of nodes are able to sense human contacts.**

***Keywords*-Mobile Social Networks; Contact Graph Reconstruction; Geographical Proximity; Social Profiles; Popularity;**

I. Introduction

The appearance of new wireless technologies have revolutionized the way people communicate and share their contents such as videos, photos, and messages. Delay Tolerant Networks (DTNs) in which nodes can exchange information only when they are in close proximity of each other have opened a new and exciting avenue for communication in the emerging social networks. In DTNs, the network is sparse and disconnected most of the time. Thus, most of known protocols for MANETs fail to operate in DTNs where successful delivery of a message strongly relies on human contact patterns.

The availability of contact traces such as [1], [2], [3] has allowed researchers to identify the fundamental properties of human mobility and to propose realistic mobility models [4], [5]. By using these mobility models, researchers have proposed efficient routing protocols for DTNs. In particular, SimBet [6], Bubble Rap [7], and Social-Greedy [8] routing algorithms are a few examples in which nodes exploit the underlying properties of contact traces for optimal routing. Therefore, the size and the reliability of contact traces are at the core of the ongoing research in DTNs.

Previously, researchers have distributed a limited number of short range wireless sensors among a set of people to record when they are in close proximity of each other. More specifically, whenever a person u who carries a sensor device comes into the close proximity of another person v who carries a wireless sensor or a Bluetooth enabled device, person u's sensor records a *contact event* with person v. In this paper, we only focus on those experiments in which wireless sensors are carried by a set of people to collect their contact events. We can represent the set of events by a directed graph called *contact graph*, where the nodes are people and the edges are contact events. We call the nodes which carry a sensor device *internal nodes* and those which carry a Bluetooth enabled device such as cellphones or PDAs *external nodes*. Justification for having two different types of nodes will be shown in section V. Specifically, real data which were collected in various settings show that people with Bluetooth enabled devices (e.g. external nodes) by far outnumbered those with wireless sensor ones.

In experimental datasets that we have analyzed (see Table I,) we have found that internal nodes have recorded a quite large number of contacts with external nodes. It is clear that these contacts belong to those people who carry their own cellphones or PDAs. While internal nodes can record the presence of all other nodes including internal and external ones, the external nodes are not able to detect any contact event. As a result, a large portion of the sampled contact graphs, specifically the contacts among all external nodes, is missing. In this paper, we are interested in reconstructing those partial contact graphs that are collected in a real experiment. We formulate the problem of inferring the missing part of a contact graph as a contact prediction problem, and we propose several methods for predicting the missing contacts. Our proposed methods predict the missing contacts among external nodes by exploiting the underlying properties of the contact graphs.

In this work, we study a variety of contact traces collected from different social settings such as [1], [2], [3], [9] for our analysis. Based on the observed time-spatial properties of contact graphs we propose three different methods that make their contact predictions by computing similarities between neighbor sets of external nodes. We also investigate the effectiveness of using nodes' contact rates for predicting missing contacts. Furthermore, it has been shown that the

978-1-4577-0351-5/11/$26.00

contact probability between mobile wireless devices is influenced by their owners' social characteristics [1], [10]. We call the set of social characteristics for each user her *social profile*. In this paper, we also present two socially-based methods and study their performance for predicting missing contacts by using a contact trace collected from a conference setting [2].

Our results show that we can reliably reconstruct the missing parts of contact graphs by using the proposed methods which in turn enables researchers to expand the existing collected contact traces in order to include the contacts among external nodes as well. Our solution to the contact prediction problem is very valuable because it also sheds light on the way in which people move. While the problem of link prediction is not new in the context of social networks [11], [12], to the best of our knowledge our work is the first one that tries to address this problem in the context of mobile social networks, a network consisting of contacts among a set of mobile users. The main contributions of this paper can be summarized as follows:

1) We present the problem of contact prediction in the context of mobile social networks and show how we can study this problem by using real data from different social settings.
2) We propose several methods each of which makes use of one of time-spatial, popularity, or social information to reconstruct the missing part of a contact graph.
3) Finally, we integrate social information with time-spatial information to propose a more effective method for contact prediction.

The remainder of the paper is organized as follow: Section II reviews the recent work in the field. Section III defines the problem to be tackled. Section IV describes the important properties of contact graphs that can help us predict contacts as well as proposed methods for contact prediction. The performance results of our methods for different datasets are presented in Section V. Finally, Section VI concludes the paper.

II. Related Work

Researchers have proposed several synthetic mobility models based on underlying properties of contact graphs. Musolesi et al. have proposed a community-based mobility model (CMM) in which nodes tend to contact other nodes from their own community with higher probability than nodes from different communities [4]. They have used real contact traces to validate the CMM. By analyzing human contact and Wireless LAN traces, Hsu et al. have also introduced the Time-variant Community Model for human mobility [5].

The underlying properties of contact graphs also play a vital role in the performance of routing algorithms for DTNs. By using complex network analysis, researchers have found patterns in contact graphs that are similar to those in social graphs. Specifically, they have discovered communities and heterogeneous centralities in contact graphs obtained from contact traces. SimBet [6] and Bubble Rap [7] routing algorithms have been proposed based on these observations.

In [8] we proposed a routing algorithm called Social-Greedy that exploits the offline social profiles of people for routing messages in DTNs. In Social-Greedy each mobile node carrying a message forwards its message to those encountered nodes that are socially closer to the messages' destinations than itself. We have studied the performance of Social-Greedy by using real data collected from a conference. However, in this paper we are proposing a more effective measure to compute the social similarity among nodes. We then show the effectiveness of our new social measure for predicting missing contacts in mobile social networks.

Nowell and Kleinberg have addressed the link prediction problem in a citation network to predict future collaborations among scientists [11]. They have proposed several predictors based on properties of social networks. Authors in [12] have assessed the confidence of experimentally collected interactions among proteins by using small-world properties of protein networks. However, in this paper our goal is to study the underlying properties of evolving contact graphs to see if we can predict the contacts among external nodes along with the time intervals when those contacts actually happen.

III. Problem Definition

A contact event between two users *u* and *v* can be shown by a quadruple (u, v, t_s, t_e) implying that user *u*'s device has detected user *v*'s device in its close proximity in the $[t_s, t_e]$ time interval. We assume that every human contact between *u* and *v* is recorded as a contact event by one of the sensors carried by *u* or *v*. It is important to note that not every observed contact between two devices necessarily means a social interaction between people who carry the devices. For the rest of paper, we only focus on analyzing those contacts that are collected by wireless sensors in an experiment.

We can show all contacts recorded by internal nodes during an experiment using a directed contact graph $G = (V, E_{known})$. Here, we denote the set of all people participating in the experiment with $V = V_{int} \cup V_{ext}$ where V_{int} and V_{ext} are the sets of internal and external nodes, respectively. We assume that $|V| = N$ and $|V_{int}| = N'$ where $N' < N$. We also denote the set of known edges by E_{known} where we translate every observed contact such as (u, v, t_s, t_e) by the internal node u to a directed edge that connects node u to node v in G. From the data provided by experiments such as those listed in Table I, we can only construct a partial contact graph. In other words, while the set of edges in $E_{known} \subset V_{int} \times (V_{int} \cup V_{ext})$ is known, all edges between external nodes are missing. Our problem is to

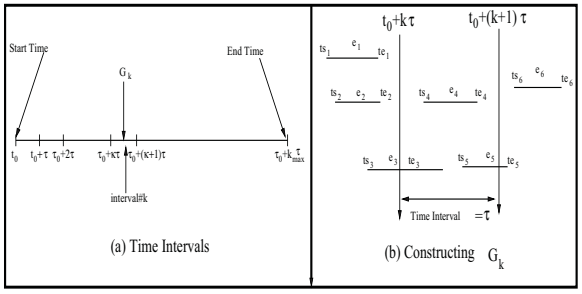

Figure 1. Constructing the partial contact graph G_k

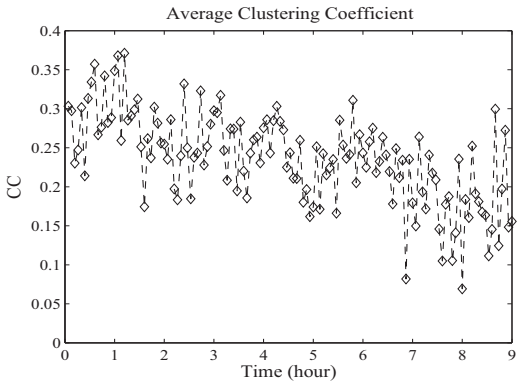

Figure 2. The average clustering coefficient (Infocom06: 9:00 AM to 6:00 PM)

infer the missing edges among external nodes by using the available information from known part of the graph (e.g. E_{known}). More specifically, our problem is to predict the missing edges $E_{unknown} \subset V_{ext} \times V_{ext}$ that are exactly the unobserved edges among external nodes.

IV. Reconstructing the Contact Graph

In this section we describe the essential properties of contact graphs that are useful for contact prediction. We also propose three different sets of prediction methods based on the underlying properties of contact graphs. Finally, we explain our main algorithm that makes use of our proposed methods to infer the missing contacts among external nodes.

A. Constructing Partial Contact Graphs

Since the collected contact traces change over time (e.g. edges between nodes appear and disappear), 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 that is the time gap between two consecutive sensings. The coefficient c should be chosen carefully according to the dataset setting. Usually c is chosen to be either 1 or 2. Let $\Lambda_k = [t_0 + k\tau, t_0 + (k+1)\tau]$ denote the k^{th} time interval where $0 \leq k \leq k_{max}$ and t_0 is the starting time of the experiment (see part (a) of Figure 1). For each time interval Λ_k, we construct a contact graph G_k by collecting all contacts that have been observed by internal nodes in Λ_k. We show the k^{th} contact graph with $G_k = (V, E_k^{known})$ where $V = V_{int} \cup V_{ext}$ is the set of all nodes and E_k^{known} is the set of all known edges of G_k (e.g. observed contacts by the internal nodes). In Figure 1, G_k contains all three contacts e_3, e_4, and e_5 that are observed by internal nodes in Λ_k. We construct all contact graphs G_k's for k_{max} steps. Our goal is to predict missing edges $E_k^{unknown}$'s by exploiting the information about the known edges of G_k's where $E_k^{known} \subset V_{int} \times (V_{int} \cup V_{ext})$ and $E_k^{unknown} \subset V_{ext} \times V_{ext}$.

B. Contact Graph Properties

There are three elements that play essential roles in contact process: (1) time-spatial locality, (2) social similarity, and (3) popularity. Next, we discuss the importance of these elements in the structure of the contact graph.

Time-Spatial locality: A contact between two nodes *u* and *v* at time t means that *u* and *v* have been in close proximity of each other at time t. If node *u* records a contact with node *v* at time t and node *u* also records a contact with node *w* at time $t + \delta$, for a small δ we can say that nodes *v* and *w* are in a close distance of each other around time t. If two nodes are geographically close, we expect that they are more likely to meet each other in the near future.

Social similarity: The other element that plays an important role in the structure of a contact graph is the social dimension of nodes. Let $sim_{soc}(u, v)$ denote the social similarity between two nodes *u* and *v*. We assume that if node *u* is more socially similar to *v* than *w* (e.g. $sim_{soc}(u, w) < sim_{soc}(u, v)$), then *u* is more likely to contact *v* than *w*.

Popularity: Popularity in social networks are captured by nodes' degrees. Let us define a node's *contact rate* as the number of contacts a node has made in the last $Wlen$ seconds. A node's contact rate in mobile social networks is similar to a node's degree in a social network, in that it reflects the social role of the device's owner. For example, if node *u*'s owner is a conference organizer, it probably has a high contact rate. Thus, if node *u* has a higher contact rate than *v*, we can assume that node *u* has a higher probability to contact a given node *w* than *v* does.

C. Methods Based on Neighborhood Similarity

To predict contacts, one approach is to exploit the underlying properties of contact graphs. Here we want to show that contact graphs have a neighborhood-cohesiveness property in which neighbors of a given node have a high probability of being connected to each other. First, we construct the contact graph G_k's for all time interval Λ_k's as described earlier. The clustering coefficient of node *u* in G_k is the fraction of pairs of *u*'s neighbors that are connected to each other by edges [13]. We can compute the average clustering

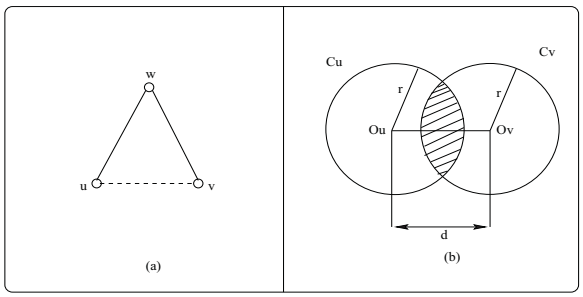

Figure 3. The effect of common neighbors on geographical proximity

coefficient of the contact graph G_k denoted with CC_k, by taking the average of clustering coefficients of all nodes in G_k. In Figure 2, we have shown the average clustering coefficients for the contact graphs during the second day of Infocom 2006 where $\tau = 2 \times T = 4$ minutes. Interestingly, the total average of all computed CC_k values is shown to be around 25%.

To justify the neighborhood-cohesiveness property of contact graphs, let us assume that we are running an experiment with N' sensors when all of them have similar sensing ranges of r. Moreover, suppose our nodes are uniformly distributed in the experimental region (e.g. a conference hotel). In this graph, there is an edge between two nodes *u* and *v* in Λ_k if their distance is less than r in that time interval (e.g. $d(u, v) < r$). Now, let us consider the simple scenario shown in part (a) of Figure 3 where two nodes u and v have detected a common neighbor w at the same time. The existence of links (u, w) and (v, w) implies a geographical proximity between u and v which in turn increases the chance that u and v also sense each other. It is not hard to see that number of common neighbors (*NCN*) between two given nodes u and v is proportional to the intersection area between their radial circles as shown in part (b) of Figure 3. A large *NCN* for two given nodes implies a large intersection area between their radial circles which in turn indicates a geographical closeness between them.

Now, we discuss three methods that measure the geographical closeness between two nodes based on the intersection between their neighbor sets. In all the following equations, $sim^k(u, v)$ denotes the similarity between two nodes *u* and *v* in Λ_k, and $N^k(u)$ represents the neighbour set of node *u* which is the set of contacted nodes by *u* in time interval Λ_k.

$$sim^k_{ncn}(u, v) = \left|N^k(u) \cap N^k(v)\right| \tag{1}$$

$$sim^k_{jac}(u, v) = \frac{\left|N^k(u) \cap N^k(v)\right|}{|N^k(u) \cup N^k(v)|} \tag{2}$$

$$sim^k_{min}(u, v) = \frac{\left|N^k(u) \cap N^k(v)\right|}{min(|N^k(u)|, |N^k(v)|)} \tag{3}$$

While Equation 1 simply calculates the *NCN* between a pair of nodes to estimate their closeness, the Equations 2 and 3 are the normalized versions of the NCN method. Equation 2 (*Jacard* method) makes use of the Jacard index to measure the similarity between the neighbor sets of the given nodes [14]. To clarify the main difference between Equations 2 and 3 (*Min* method), let us consider two scenarios. In the first scenario, let us assume that two nodes *u* and *v* contact two similar nodes in time interval Λ_k. Also suppose node *u* has only contacted these two nodes while node *v* has contacted ten nodes including these two common nodes in Λ_k. We might desire a higher significance for the described scenario than if each of *u* and *v* had seen six nodes in Λ_k where two of these six nodes are in common. Although the Jacard index gives $\frac{1}{5}$ for both scenarios, the Min method assigns a higher score to the first scenario [12].

D. Methods Based on Social Similarity

Considering the importance of homophily principle in link formation process in social networks [15], we want to test the power of social similarity for predicting the contacts among nodes in a mobile social network. In our previous work, we have studied the influence of the similarity of people's social profiles on their contact probabilities in a conference environment [8]. Eagle et al. [1] and Mitbaa [10] have also found a close relation between human mobility and their friendship networks. In this work, we have access to brief social profiles of people who attended Infocom 2006 conference. In these social profiles, people have reported their social characteristics. We can present each social characteristic i (*social dimension*) for node *u* with a set Γ^i_u called the social characteristic set of node *u*. For example, node *u*'s research interests can be shown with a set of topics as $\Gamma^{interests}_u = \{1, 2, 3\}$ where 1, 2, 3 can represent DTN, MANET, and Social Networking areas.

1) Social Similarity Based on Jacard Index: By employing the Jacard index, we can compute the similarity between two nodes with respect to each social dimension as follow [8]:

$$\sigma^i_{jacard}(u, v) = \frac{\left|\Gamma^i_u \bigcap \Gamma^i_v\right|}{\left|\Gamma^i_u \bigcup \Gamma^i_v\right|}, \tag{4}$$

where Γ^i_u is the social characteristic set of node u for social dimension i, and $\left|\Gamma^i_u\right|$ is its cardinality. Assuming that we have d different social dimensions for each node, we can compute the total social similarity between two nodes by calculating the total average over all d dimensions. Therefore, we can obtain the total social similarity between two nodes u and v as below:

$$sim^{soc}_{jac}(u, v) = \sum_{i=1}^{d} \frac{\sigma^i_{jacard}(u, v)}{d}, \tag{5}$$

where d is the number of social dimensions. Note that for Infocom 2006 data, the number of available dimensions are 6 as it will be shown later ($d = 6$).

2) Social Similarity Based on Foci Distance: Suppose in a conference there are two people *u* and *v* who are interested in the *Routing* research area. Moreover, assume that these two people do not have any other similar social characteristic (or *social focus*). More specifically, they have different affiliations, were born in different countries, speak in different languages and so on. Roughly speaking, there is a high probability for these two people to meet each other in the conference because both of them may attend the same sessions. We also can see that their contact probability has an inverse relationship with the number of people who are interested in the *Routing* area. Thus, we can define the *Foci distance* between two given nodes as the cardinality of the smallest social focus that both of them share with each other. In our previous example, the social distance between *u* and *v* with respect to research interest is equal to the number of conference participants who are also interested in *Routing*. We can write the Foci distance between two given nodes *u* and *v* as below [16]:

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

where F is the common social focus of *u* and *v*. Note that there is a supergroup that contains all nodes of the network. Considering Equation 6, we can define the *Foci similarity* between two nodes as follow:

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

E. Method Based on Popularity

Motivated by the preferential attachment model for social networks [17] in which a node *u* connects to another node *v* with a probability that is proportional to *v*'s degree, we can assume that the contact probability between two nodes in a mobile social network depends on their individual contact rates. Let λ_u denote the contact rate of node *u* that is, the number of contacts in an interval. We assume that the combined contact rate between two nodes *u* and *v* is proportional to the product of their individual contact rates. Thus, we define the popularity measure between two nodes *u* and *v* in time interval Λ_k as below:

$$sim^{k}_{pop}(u, v) = \lambda_u.\lambda_v \tag{8}$$

To compute node u's contact rate we count the number of u's contacts in the last $Wlen$ seconds. In Section V, we employ all six Equations 1, 2, 3, 5, 7, and 8 as our proposed prediction methods to reconstruct the missing parts of contact graphs.

Table I
REAL DATA DESCRIPTION

Dataset	Inf 05	Inf 06	MIT	Camb	Roller
Mobile nodes	41	79	97	36	62
Length	3 days	4 days	246 days	11 days	3 hours
Scanning period	120 sec	120 sec	300 sec	600 sec	15 sec
External no	206	4321	20698	11367	1050
Total contacts	227657	28216	285512	41587	132511
Ext. contacts	57056	5757	183135	30714	72365
Ext. contacts %	25%	20%	64%	74%	55%

F. Reconstruction Algorithm

The following steps describe our algorithm for reconstructing the missing parts of G_k's by selecting one of the previously presented prediction methods:

1) First we generate partial contact graph G_k's for all k values as we have described in Section IV.
2) Next, we compute the similarity scores between all pairs of external nodes by using one of our prediction methods. These similarity scores basically estimate the contact probabilities among external nodes. Therefore, for each time interval Λ_k we obtain quadruples such as $(u, v, k, sim(u, v))$ where *u* and *v* are external nodes, $sim(u, v)$ is the computed similarity score for nodes *u* and *v*, and k is the time interval number. We store all quadruples whose similarity scores are greater than zero in a similarity list denoted by L_{sim} for the post-processing step. We repeat the same process for all intervals ($0 \leq k \leq k_{max}$) and store all the quadruples in the same list.
3) When we finish with all intervals, we sort the L_{sim} list in a descending order based on computed similarity scores. The sorted version of the similarity list is our predictor results.
4) To infer the missing contacts, we select the first $Rank$ number of predictions from the sorted list of L_{sim}.

V. PERFORMANCE EVALUATION

In this section we evaluate the performance of all proposed methods in the previous section by using the available datasets. Our ultimate goal is to see which method is more effective for predicting the missing contacts.

A. Real Data Description

Here, we are planning to use contact traces collected from four different social settings. Table I describes these datasets. Info 05 and Info 06 datasets were collected from Infocom conference in 2005 and 2006 respectively. Participants in Info 05's experiment belong to different social communities; however, in the Infocom 2006 participants were especially selected such that 34 people out of 79 were from four research groups [18]. In Info 06 dataset, participants also reported a brief version of their social profiles which consisted of (1) *nationality*, (2) *graduate school*, (3) *languages*,

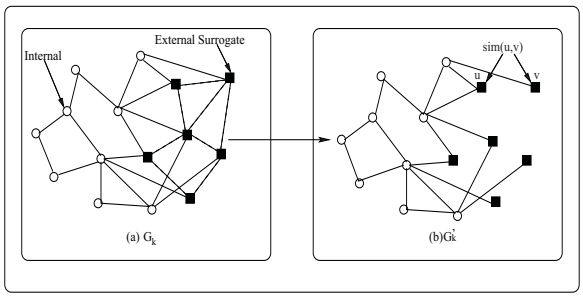

Figure 4. Simulating a partial contact graph

(4) *affiliations*, (5) *city & country of residence*, and (6) *research interests* [2]. In Cambridge dataset (Camb), the wireless sensors were distributed mainly between two groups of undergraduate students from University of Cambridge, and some graduate students from a research lab [9]. Note that both Infocom 06 and Cambridge datasets also include data for stationary wireless sensors; however, in this paper we only use the sampled contacts by mobile sensors. While Rollernet dataset (Roller) contains the contacts from a set of people who participated in a rollerblading tour in Paris [3], Reality dataset (MIT) includes contact data of students and staff at MIT for a period of 9 months [1].

As we can see from Table I, the number of external nodes in all datasets are much larger than the number of internal nodes. Moreover, the ratios of sampled contacts with external nodes to total contacts are also significant. This observation is our main motivation to propose a prediction algorithm for reconstructing the missing parts of partially sampled contact traces.

B. Testing Reconstruction Algorithm Using Real Data

Since the external nodes do not carry sensors, there is no way to validate the inferred edges between them. To get around this issue, we choose a random subset of the internal nodes and label them as the *surrogates* of external nodes. These surrogates are going to act as external nodes. We remove all the recorded contacts between surrogates from our G_k's. This process is shown in Figure 4 where the surrogates are shown as squares in the original graph G_k before edge removal. To generate the partial graph G'_k, we remove all contacts that have been recorded by surrogates, but we still keep the contacts recorded by the remainder of internal nodes (i.e. circle nodes in Figure 4). These partial G'_k's are the inputs for our reconstruction algorithm.

To validate an inferred contact, we examine the original contact trace that includes all contacts among all internal nodes to see if we can find a match. For example, for an inferred contact such as $(u, v, k, sim(u, v))$, we search through our complete dataset to see if there have been any contacts between *u* and *v* in the $[t_0 + k\tau, t_0 + (k+1)\tau]$ time interval. We are able to do this because the surrogates are actually internal nodes and we have all their contact data.

Table II
THE PERCENTAGE OF MISSING PART OF CONTACT TRACES

Dataset	Inf 05	Inf 06	MIT	Camb	Roller
Edge Loss %	52%	56%	61%	56%	55%

C. Contact Prediction Using Time-Spatial and Popularity Information

To evaluate the performance of our prediction methods, for all our datasets we randomly choose 75% of internal nodes and label them as surrogates. We then construct the partial contact graph G'_k's as described earlier. The resulting partial graphs include only a small subset of nodes. Table II shows the average percentage of contacts that we discard by labeling 75% of the internal nodes as surrogates. Our goal is to infer the missing contacts between the surrogates of the partial contact graph G'_k's. Note that for Infocom 2005 and 2006 datasets, we use only the collected contacts on the first day of the main conference; however, for the Cambridge and Rollernet datasets we use all contacts for our analysis. We also use 35 days of MIT data for our analysis. We repeatedly run our prediction algorithm with 20 different random sets of surrogates, and the presented results are the average values.

Let us first focus on the performance of methods that are based on neighborhood similarity and popularity. To infer the missing contacts, we pick the first $Rank$ number of predictions from our sorted L_{sim} list as the most confident predictions. We then compute the percentage of matches (i.e. *true positives*) between our prediction results and real data by inspecting our database. We increase the $Rank$ value to see how different methods operate as we increase the number of predictions. For comparison purposes, we use as our baseline a simple *random predictor* that randomly selects a pair of surrogates and a time slot as a prediction. Figures 5, 6, 7, and 8 show the percentages of true positives for Infocom 2006, Cambridge, Rollernet, and MIT datasets respectively. We have not shown our results for Infocom 2005 because of the space limitation. However, Infocom 2005's results are very similar to the presented results for 2006.

From the given figures, we make several interesting observations. First, NCN, Jacard, Min, and Popularity predictors significantly outperform the random one, proving that there is indeed useful information even in partial contact graphs which can be used for prediction purposes. Second, it is evident that in most of our evaluations, the methods based on neighborhood similarity perform better than the popularity method which again proves the importance of using time-spatial locality for predicting the missing parts of contact traces. The reader should note that the popularity method does not contain any location information, unlike the neighborhood similarity methods. Third, as we increase the $Rank$ value, the percentage of true positives drops, while the percentage of false positives increases. This is because

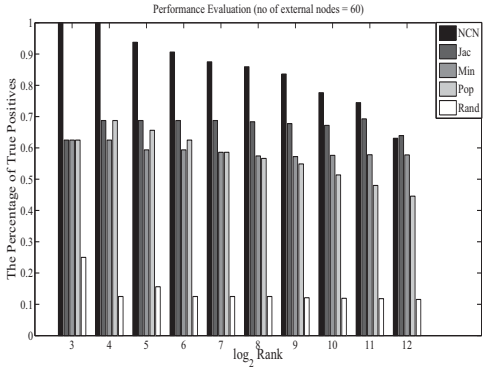

Figure 5. Percentage of true positives for contact predictions (Info 06)

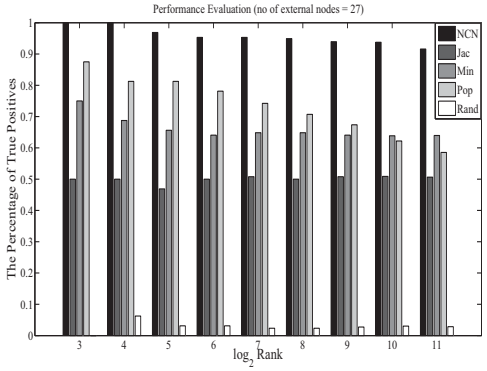

Figure 6. Percentage of true positives for contact predictions (Camb)

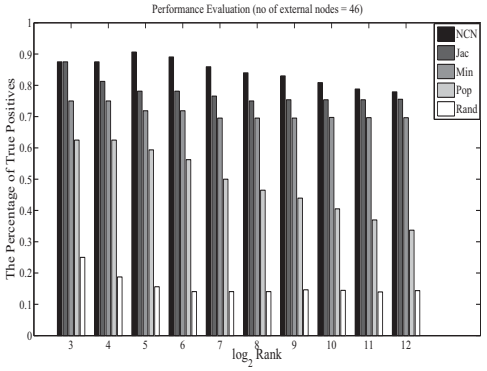

Figure 7. Percentage of true positives for contact predictions (Roller)

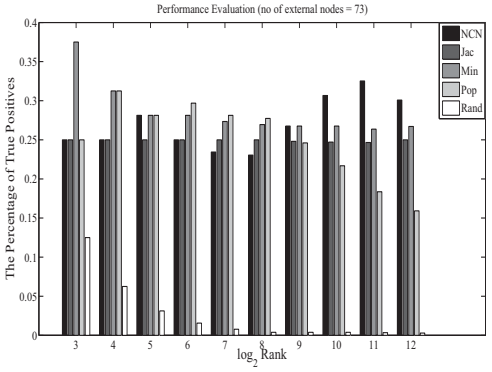

Figure 8. Percentage of true positives for contact predictions (MIT)

we sort our similarity list in a descending order.

Interestingly, in the Rollernet dataset we observe that the effect of popularity drops more significantly than other datasets as we increase our $Rank$. To justify this we should recall the social structure of people who participated in this experiment. One group consisting of 25 staff members were asked to stay at previously assigned positions in front and back of the tour. There was another group of friends with 11 nodes as well as a group of skilled skaters with 26 nodes. We believe that most of these people except skilled skaters were not very mobile and were located most of the time in the same relative position inside the tour. As a result, the performance of popularity method without location information drops faster than that of a conference setting in which people have a chance to meet each other at least every two hours during coffee breaks.

Furthermore, results from the MIT dataset show that for large geographical spaces (e.g. campus environments) the percentage of true positives for contact predictions is low because it is likely to have a subset of external nodes where there are not any internal nodes in their proximity. Therefore, the geographic based methods fail to predict the missing contacts among such external nodes and different prediction methods may have to be devised.

It is necessary to mention that most real contact traces are quite noisy, and in particular they can miss many contacts. Authors in [19] have listed several issues related to iMotes software which were used in the Infocom, Cambridge, and Rollernet experiments. The reset issue because of memory overflow, the synchronization issue because of random seeds, and the limitation for the number of responses returned in a Bluetooth scanning caused by the Bluetooth protocol stack are the most important ones. All of these issues can cause iMotes to miss some of real contacts. Thus, the number of computed false positives for our predictors may be overestimated because some of those false positives could have actually happened in the reality, but iMotes failed to capture them.

D. Contact Prediction Using Offline Social Information

As we have mentioned earlier, Infocom 2006 data also includes participants' social profiles. In this part of our analysis we assume that we do not know anything about the contact trace except the social profiles of participants. In other words, we assume that all internal nodes act as surrogates of the external nodes. For testing our social methods, we compute social similarities between all possible

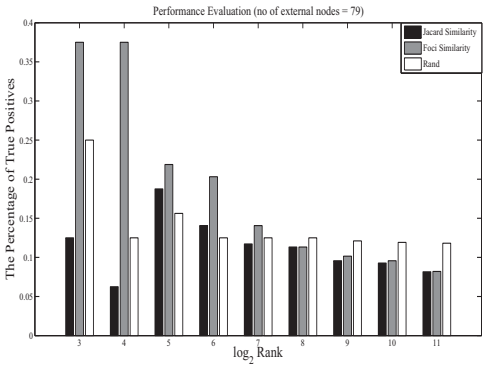

Figure 9. Percentage of true positives for contact predictions using social data (Info 06)

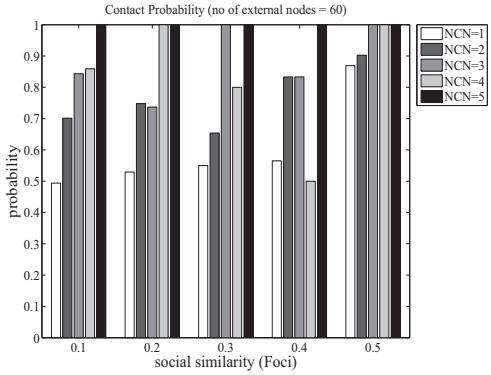

Figure 10. Contact probability as a function of social and proximity information (Info 06)

pairs of surrogates (e.g. $(u, v, sim(u, v))$ by using Equations 5 and 7 independently. We store the computed Jacard and Foci similarities in L_{sim}^{jac} and L_{sim}^{foc} lists, respectively. We sort both of these similarity lists in descending order. These two lists are our social predictor results for Jacard and Foci similarities. To validate a prediction based on social similarity such as $(u, v, sim(u, v))$, we randomly select a time interval Λ_k as the time step when a contact has happened between nodes u and v. For evaluation, we choose the first $Rank$ number of predictions from our sorted similarity lists and inspect them by using Infocom 2006's contact trace data.

Figure 9 shows the percentages of true positives regarding our prediction results when we only use social profiles. The figure shows that Foci similarity better reflects the similarity among people who attend a conference than Jacard does. From Figure 9, we can make the interesting observation that using social data without any time and proximity information can still be helpful for predicting missing contacts. The reader should note that the collected social profiles were only partial in that some people did not report their complete profiles, or any at all. By testing the distribution of social profiles, we have found that there are only around 100 pairs of nodes which are socially very similar (e.g. $sim_{foc}^{soc} >= 0.2$) while the majority of nodes are not. Therefore, we expect that for $Rank$ values greater than 7, social profiles lose their effect for prediction task. For the rest of our analysis we only use the Focus method to measure the social similarity between nodes.

We have already seen that NCN method outperforms others as it contains the proximity information. Now, the question is if we can propose a better predictor by using both social profiles and NCN information. One could make the case that once two users are in relative proximity (e.g. the same room), the probability of meeting each other is high if they are also friends. We need to characterize the effects of social focus and NCN on contact probability. We therefore select 75% of internal nodes as surrogates and repeat our evaluations as before by computing NCN scores between surrogates. We filter those quadruples that exactly have c number of nodes in common (e.g. $sim_{ncn}^{k}(.,.) = c$) and store all of them in L_{sim}^{c} list. Next, we use data binning to categorize all L_{sim}^{c}'s quadruples into five equally sized intervals based on Foci similarity. We choose our intervals as $thr_{soc} \leq sim_{foc}^{soc}(u, v) \leq thr_{soc} + 0.1$ where $thr_{soc} \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$. We then compute the percentage of surrogates that have actually met each other. This gives us the contact probabilities for different social similarity and NCN values. We need to repeat the same process for all possible NCN values (e.g. $1 \leq c \leq 5$).

Figure 10 shows the contact probability among surrogates as a function of NCN and Foci similarity. We can observe that for NCN values of 1, 2 and 3, as Foci similarity between two given nodes increases, the contact probability also becomes greater. Figure 10 also shows that we cannot expect any improvement by adding social information when NCN is large. This is because for these cases the NCN acts as a dominant factor in contact probability. Our results are very encouraging as they provide incentive to incorporate social information with the NCN method to achieve a better performance.

One possible way for combining social information with NCN would be to compute prediction quadruples by using the NCN method. We then sort all quadruples in a descending order based on their NCN scores. Next, we use the social Foci similarity as the second dimension to rank all quadruples with the same NCN scores in a descending order. This second ranking would give a higher weight to those pairs that are socially closer. This two-fold sorting would provide better performance than when we use only one of NCN or Foci similarity. Incorporating nodes' contact rates with NCN information with a similar approach would also enhance the performance of the NCN predictor. We have plan to pursue this direction as future work.

VI. Conclusions And Future Work

In this paper, we have studied the problem of contact prediction in the context of mobile social networks. We have described different methods for predicting missing contacts. We have examined our methods by using real contact traces collected from different social settings. Our results show that time-spatial based scores provide the most reliable results for predicting missing contacts among external nodes. We have also studied the power of social profiles to predict human mobility. We have shown that combining social information with time-spatial information can provide better performance results than using each of them independently. We believe that our contributions have significant practical values because they allow researchers to study properties of large scale contact graphs by sampling only a portion of the original graphs. Our results are also important for mobility modeling since they explain how people move in different social settings such as conference and campus environments. For our future work, we plan to propose more efficient methods for predicting missing contacts in large geographical spaces as in MIT dataset.

References

[1] N. Eagle, A. Pentland, and D. Lazer, "Inferring social network structure using mobile phone data," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 106(36), pp. 15 274–15 278, July 2009.

[2] A. Chaintreau, P. Hui, J. Scott, R. Gass, J. Crowcroft, and C. Diot, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 606–620, June 2007.

[3] 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*. IEEE, april 2009, pp. 1116–1124.

[4] M. Musolesi and C. Mascolo, "A community based mobility model for ad hoc network research," in *Proceedings of the 2nd international workshop on Multi-hop ad hoc networks: from theory to reality*, ser. REALMAN '06. New York, NY, USA: ACM, 2006, pp. 31–38.

[5] W.-J. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, "Modeling Spatial and Temporal Dependencies of User Mobility in Wireless Mobile Networks," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1564–1577, October 2009.

[6] E. M. Daly and M. Haahr, "Social network analysis for routing in disconnected delay-tolerant manets," in *MobiHoc '07: Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*. New York, NY, USA: ACM, 2007, pp. 32–40.

[7] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: social-based forwarding in delay tolerant networks," in *MobiHoc '08: Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*. New York, NY, USA: ACM, 2008, pp. 241–250.

[8] K. Jahanbakhsh, G. C. Shoja, and V. King, "Social-greedy: a socially-based greedy routing algorithm for delay tolerant networks," in *MobiOpp '10: Proceedings of the Second International Workshop on Mobile Opportunistic Networking*. New York, NY, USA: ACM, 2010, pp. 159–162.

[9] J. Leguay, A. Lindgren, J. Scott, T. Friedman, and J. Crowcroft, "Opportunistic content distribution in an urban setting," in *Proceedings of the 2006 SIGCOMM workshop on Challenged networks*, ser. CHANTS '06, 2006, pp. 205–212.

[10] A. Mtibaa, A. Chaintreau, J. LeBrun, E. Oliver, A.-K. Pietilainen, and C. Diot, "Are you moved by your social network application?" in *Proceedings of the first workshop on Online social networks*, ser. WOSP '08. New York, NY, USA: ACM, 2008, pp. 67–72.

[11] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*. New York, NY, USA: ACM, 2003, pp. 556–559.

[12] D. S. Goldberg and F. P. Roth, "Assessing experimentally derived interactions in a small world," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 100, no. 8, pp. 4372–4376, February 2003.

[13] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks." *Nature*, vol. 393, no. 6684, pp. 440–442, June 1998.

[14] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bulletin del la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[15] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[16] J. Kleinberg, "Small-world phenomena and the dynamics of information," in *In Advances in Neural Information Processing Systems (NIPS) 14*. MIT Press, 2001, p. 2001.

[17] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science (New York, N.Y.)*, vol. 286, no. 5439, pp. 509–512, October 1999.

[18] 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*, ser. PERCOMW '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 65–70.

[19] E. Nordström, C. Diot, R. Gass, and P. Gunningberg, "Experiences from measuring human mobility using bluetooth inquiring devices," in *MobiEval '07: Proceedings of the 1st international workshop on System evaluation for mobile platforms*. New York, NY, USA: ACM, 2007, pp. 15–20.