Predicting Missing Contacts in Mobile Social Networks

Kazem Jahanbakhsh, Valerie King, and Gholamali C. Shoja

Computer Science Department, University of Victoria

June 2011

Outline I

Human Contact Prediction

- Modeling human mobility is a challenging but interesting problem.
- Knowing how people move can help us in:
  - Designing efficient routing algorithms for DTNs.
  - Proposing accurate human mobility models.
  - Designing mobile social applications.
  - Traffic planning in cities.
  - Modeling epidemic disease.

Human Contact Prediction Problem

- We say two people are *in contact* if they are in each other's proximity ($< 10m$).
- A contact can be detected by a wireless sensor (Bluetooth).
- A contact has time/spatial information.
- Predicting where and when people are going to contact each other is an interesting problem.
- For this we need to collect contact trace data.

Why predicting human contacts?

- Researchers have studied the properties of human mobility by using real data.
- There are several available contact traces which are collected by Bluetooth sensors.
- MIT Reality Mining, Infocom 05/06, Rollernet, and Cambridge datasets are few examples.
- All of these real datasets contain contacts among only a limited number of nodes.

Our Observations

- MIT dataset only includes contacts among 100 nodes.
- Issue: not everybody carries a wireless sensor (price/technical issues).
- However, most of people carry their cellphones.
- We have found that most of real datasets contain a large number of contacts from cellphones.
- Cellphones cannot record any contacts.
- Therefore, a large portion of contacts are missing.
- How can we infer the contacts among cellphones (i.e. external devices)?

Definitions and Assumptions

- A contact event between two nodes u and v is shown by a quadruple (u, v, t_s, t_e).
- *Internal nodes*: nodes which carry sensor devices (V_{int}).
- *External nodes*: Bluetooth enabled devices (cellphones and PDAs: V_{ext}).
- Contact events among people can be shown by a directed graph called *Contact Graph*.
- In contact graph $G = (V, E)$, V is the set of nodes and E is the set of contacts among people.
- We assume that $V = V_{int} \cup V_{ext}$.

Reconstructing the Contact Graph

- We assume only internal nodes can sample contacts.
- All edges in $E_{known} \subset V_{int} \times (V_{int} \cup V_{ext})$ are known.
- All edges in $E_{unknown} \subset V_{ext} \times V_{ext}$ are missing.
- Our problem is to infer the edges among external nodes (edges in $E_{unknown}$).

Related Work

- Several human mobility models have been proposed: community-based mobility model by Musolesi et al.
- Daly et al. and Hui et al. proposed routing algorithms which exploit contact graphs properties.
- Nowell and Kleinberg have studied the problem of link prediction in citation networks.
- Goldberg et al. have used cohesive neighborhoods between proteins for assessing the confidence of observed interactions among them.
- Our work is the first one that addresses contact prediction problem in the context of mobile social networks.

Contact Graph Properties

- We can compute the contact probability among people by exploiting contact graph properties:
- **Time-Spatial locality**: exploiting time-spatial properties of contact graphs.
- **Popularity**: exploiting the contact rates of mobile nodes.
- **Social similarity**: using offline social information about people who carry wireless devices.

Number of Common Neighbors and Geographical Proximity

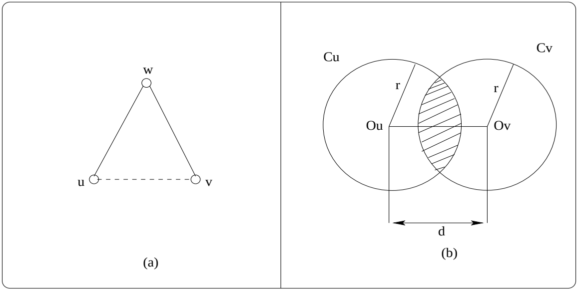

Figure 1: The effect of common neighbors on geographical proximity

Measuring Geographical Closeness

- We can compute the geographical closeness between two nodes by analyzing their neighbor sets over time.
- $$sim_{ncn}^{k}(u,v) = \left|N^k(u) \cap N^k(v)\right| \quad (1)$$
- $$sim_{jac}^{k}(u,v) = \frac{\left|N^k(u) \cap N^k(v)\right|}{\left|N^k(u) \cup N^k(v)\right|} \quad (2)$$
- $$sim_{min}^{k}(u,v) = \frac{\left|N^k(u) \cap N^k(v)\right|}{min(\left|N^k(u)\right|, \left|N^k(v)\right|)} \quad (3)$$

Popularity

- Preferential attachment: in social networks the probability of connection is proportional to nodes' degrees.
- Contact rates of mobile nodes play a similar role as node degrees in social networks.
- We assume that the contact probability between two nodes is proportional to the product of their contact rates:
- $$sim^k_{pop}(u, v) = \lambda_u.\lambda_v \tag{4}$$
- λ_u: the number of contacts of node u during time interval Λ_k.

Social Profiles (Infoccom 2006)

- The social profiles contain information about 6 different social dimensions.
- Each social dimension can be shown with a set of social features.
- Suppose node u speaks English and Spanish.
- Let us denote *English* and *Spanish* with 1 and 2 respectively.
- We can show the spoken languages of node u with a feature set $\Gamma u = \{1, 2\}$.

Jacard Social Similarity

- Social similarity between two nodes with respect to dimension i can be computed by using Jacard index:
- $$\sigma^i_{jacard}(u,v) = \frac{|\Gamma^i_u \bigcap \Gamma^i_v|}{|\Gamma^i_u \bigcup \Gamma^i_v|} \tag{5}$$
- Total similarity between two nodes is computed as the average over all dimensions:
- $$sim_{jac}(u,v) = \sum_{i=1}^{d} \frac{\sigma^i_{jacard}(u,v)}{d} \tag{6}$$

Foci Social Similarity

- The social distance between two nodes u and v can be defined as the size of the smallest social feature set that includes both of them:
- $$d_{foc}(u, v) = min \left| \{F | u, v \in F\} \right| \quad (7)$$
- Here, F is the focus set to which both u and v belong.
- Using the foci distance, the foci similarity between two nodes u and v is:
- $$sim_{foc}(u, v) = \frac{1}{d_{foc}(u, v)} \quad (8)$$

Reconstruction Algorithm

- First, we generate partial contact graph G_k's for all k's.
- Next, we compute the similarity scores between all pairs of external nodes by using one of our methods.
- For each time interval Λ_k, we obtain quadruples such as $(u, v, k, sim(u, v))$.
- We store all quadruples in a similarity list (L_{sim}).
- We sort L_{sim} list in a descending order based on computed similarity scores.
- To infer the missing contacts, we select the first *Rank* number of predictions from L_{sim}.

Real Data Descriptions

- The first two datasets are from Infocom 2005/2006 where 41 and 79 participants attended.
- The third dataset is collected at University of Cambridge (36 sensors).
- Rollernet dataset contains contacts from a set of people who participated in rollerblading (62 nodes).
- MIT dataset lasted for 9 months and includes contacts among 97 nodes.
- All of these datasets were sampled by Bluetooth sensors (e.g. < 10 meters).

How to Test Reconstruction Algorithm Using Real Data?

- External nodes donot carry any sensors.
- There is not any way to validate the predicted contacts among them.
- We can choose a subset of internal nodes and pretend that they are external nodes.
- We call these nodes as *surrogates* of external nodes.
- We remove all the contacts among surrogates.
- For our analysis we choose 75% of nodes in random as surrogates.
- We use our prediction methods to infer contacts among surrogates.

Simulating Partial Contact Graphs

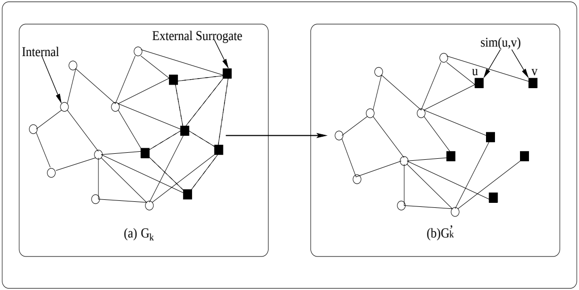

Figure 2: Simulating a partial contact graph

Percentage of True Positives for Infocom 2006

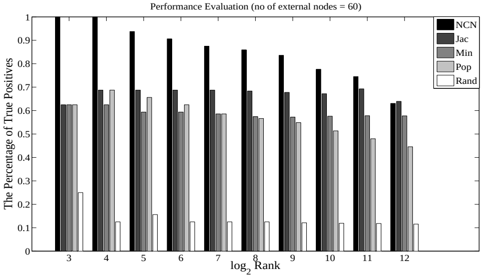

Figure 3: Percentage of true positives for contact predictions (Info 06)

Percentage of True Positives for Cambridge

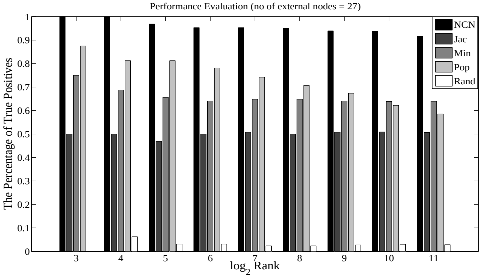

Figure 4: Percentage of true positives for contact predictions (Camb)

Percentage of True Positives for Rollernet

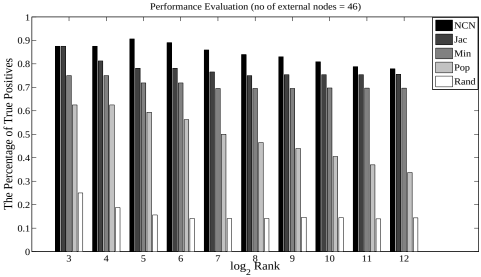

Figure 5: Percentage of true positives for contact predictions (Roller)

Percentage of True Positives for MIT

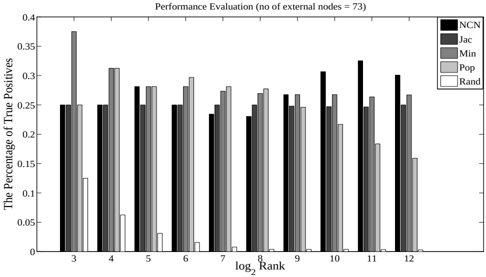

Figure 6: Percentage of true positives for contact predictions (MIT)

Using Offline Social Profiles

- On the next slides, we test the power of social profiles for contact prediction.
- For the first part of our analysis, we assume that we only have social profiles of nodes in V.
- We assume all edges of G are unknown.
- The problem is to infer edges in E by only using the social information of nodes.
- Then, we study the performance of combining social information with proximity data for contact prediction.

Percentage of True Positives Using Social Data

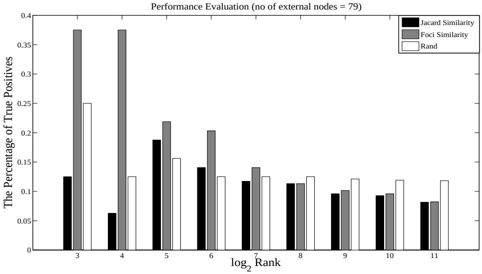

Figure 7: Percentage of true positives for contact predictions using social data (Info 06)

Percentage of True Positives Using Social/Proximity Data

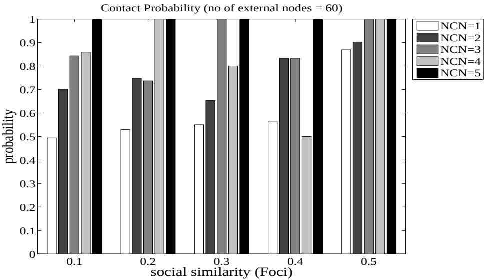

Figure 8: Contact probability as a function of social and proximity information (Info 06)

Discussion

- NCN, Jacard, Min, and Popularity outperform random predictor.
- Methods based on neighborhood similarity perform better than the popularity method.
- For large geographical spaces (MIT) the percentage of true positives is low.
- This is because it is likely to have a subset of external nodes where there are not any internal nodes in their proximity.
- Using social data without any time-proximity information is still helpful for contact prediction task.
- Foci similarity better reflects people mobility in a conference.

Conclusions and Future Work

- We have studied the problem of contact prediction in the context of mobile social networks
- Our results show that time-spatial based scores provide the most reliable results.
- We have shown that combining social information with time-spatial information provides better performance results.
- Our methods allow researchers to study properties of large scale contact graphs by sampling contacts among a subset of graph nodes.
- We plan to propose more efficient methods for predicting missing contacts in large geographical spaces (e.g. MIT).