Inferring Geographic Coincidence in Ephemeral Social Networks *

Honglei Zhuang[†], Alvin Chin[‡], Sen Wu[†], Wei Wang[‡]
Xia Wang[‡], and Jie Tang[†]

[†] Department of Computer Science and Technology, Tsinghua University
[‡] Nokia Research Center Beijing
zhl09@mails.tsinghua.edu.cn, jietang@tsinghua.edu.cn, ronaldosen@gmail.com,
{alvin.chin,wei.41.wang,xia.s.wang}@nokia.com

Abstract. We study users' behavioral patterns in ephemeral social networks, which are temporarily built based on events such as conferences. From the data distribution and social theory perspectives, we found several interesting patterns. For example, the duration of two random persons staying at the same place and at the same time obeys a two-stage power-law distribution. We develop a framework to infer the likelihood of two users to meet together, and we apply the framework to two mobile social networks: UbiComp and Reality. The former is formed by researchers attending UbiComp 2011 and the latter is a network of students published by MIT. On both networks, we validate the proposed predictive framework, which significantly improve the accuracy for predicting geographic coincidence by comparing with two baseline methods.

1 Introduction

An ephemeral social network indicates a social network temporarily created because of an event such as conference, game, or banquet. Such social networks are usually formed quickly and dissolve in minutes as well. Ephemeral social networks exist in both online and offline domains. In fact, these networks play an important role to expand users' social circle and strengthen social ties [16]. Different persons have very different behaviors in the ephemeral social networks. It is interesting and also important to understand what are the driving force for persons to select targets to meet.

There has been a few related works. For example, Eagle et al. [9] studied how friend relationships are formed by tracing users' geographic information through Wi-Fi, GPS and Bluetooth. They found that friends demonstrate distinctive temporal and spatial patterns in their physical proximity and calling

* The work is supported by Nokia Research Center and is also in part supported by the Natural Science Foundation of China (No. 61073073 , No. 60973102 , No. 61170061), Chinese National Key Foundation Research (No. 60933013, No.61035004), a special fund for Fast Sharing of Science Paper in Net Era by CSTD, and National Basic Research Program of China (No. 2011CB302302).

Fig. 1. An ephemeral social network via the Find & Connect system on Ubicomp'11 [5]. The left figure shows the recommended users for "Chin"; the right figure shows the detailed information of a recommended user.

patterns. Crandall et al. [7] investigated how social ties between people can be inferred from co-occurrence in time and space. Tang et al. [25] developed a general learning framework for inferring the types of social ties in social networks; and [22] further extended the problem of inferring social ties across heterogeneous networks by incorporating social theories such as social balance theory and social statu theory. However, all the aforementioned work only consider the problem in normal social networks. The situation is very different in ephemeral social networks. In a normal social network, friends tend to meet together to share recent experiences. However, in an ephemeral network, people are often inclined to make new friends. For example, in an academic conference, people may want to build new research collaborations with people who they may do not know before. An interesting question is: how likely are two random persons in an ephemeral social network to gather together, and how does the likelihood depend on users' personal information and their onsite spatial information?

We use an example to clearly motivate this work. Figure 1 shows the interface of our developed Find & Connect system on mobile phone. The system is designed for facilitating social interactions in ephemeral social networks, and has been deployed in several real scenarios including Ubicomp'11, Nokia Beijing Research Center, and Tsinghua Centenary ceremony. Employing the conference Ubicomp'11 as the example, the system allows the user to locate friends, check attendees in surrounding areas. One important feature of the system is to recommend people to meet. The left of Figure 1 shows the recommendation results for user "Chin". The user can check each recommended user (as shown in the right figure). Obviously, an accurate recommendation algorithm should consider not only social networking information, but also the onsite location information.

We formalize the problem of inferring geographic coincidences in ephemeral social networks. The goal is to investigate the underlying patterns that drive people to meet together, and to predict how likely a geographic coincidence would happen in the near future. The problem presents a set of challenges.

- *Making new friends.* As stated before, an important objective of users joining an ephemeral network (event) is to build new connections. It is important to predict new friendships in social networks.
- *Combining normal networks.* An ephemeral social network is not standalone. For example, attendees of an academic conference can be connected to academic social networks such as ResearchGate or Arnetminer. However, it is unclear how to combine the various normal networks for better predict the geographic coincidences.
- *Partially observed.* The ephemeral social network is always partially observed. Even the best organized event, there might be a portion of missing data due to various reasons, e.g., device failure and privacy protection. How to build a predictive model by considering the unlabeled data is a challenge.

To address the above challenges, we first study the behavioral patterns on how users meet together. We have found several interesting phenomena from both data distribution and social theory aspects. The duration of two persons staying at the same place and at the same time obeys a two-stage power-law distribution. Ten-minutes seems to be a boundary for users to staying together. From another perspective, ephemeral social networks represent more elite-related activities: elite users tend to meet together and ordinary users are also inclined to meet elite users. We also study two important social theories, homophily and structural hole, in the ephemeral social network.

Based on the discovered behavioral patterns, we present a semi-supervised predictive framework, which incorporates the various patterns in unified model. An efficient algorithm is developed to learn the framework. Our experiments on two different networks validate the effectiveness of the proposed methodologies. Comparing with several baseline methods using SVM and CRF, the proposed model can improve the prediction performance by 8-19% (in terms of F1-score).

2 Preliminaries

In this section, we first define the ephemeral social network and present our problem formulation. Then we describe the data sets used in our empirical study.

2.1 Problem Formulation

An ephemeral social network is a temporary and dynamic network. Generally, we can consider users from (different) normal social networks form the temporary structure and behaviors in the ephemeral network. For example, in a game, users may form different groups based on their relationships and intimacy, while in a conference people gather in a technical session according to their interest.

Let $G = (V, E, \mathbf{W})$ represent a normal social network, where V is a set of users, $E \subset V \times V$ is a set of relationships between users, and $\mathbf{W}$ is an attribute matrix associated with users V. An ephemeral social network can be defined as $G'(t) = (U^t, \mathbf{X}^t, Y^t)$, where $U \subset V$ is a subset of V indicating users forming the ephemeral social network come from a normal social network, $\mathbf{X}^t$ denotes an ephemeral attribute matrix for users in U^t, and Y^t denotes a set of user behaviors we want to predict, e.g., whether a user will join a seminar.

Without loss of generality, we employ the ephemeral network built in the UbiComp 2011 conference as the example to define our problem. Users of the ephemeral network are researchers from universities and companies. Their corresponding normal network can be defined as the coauthor network. The ephemeral attributes include where the user is, when the user will give a talk, what the user is doing, etc.

A usual predictive task in an ephemeral network is to predict users' future behavior by leveraging the normal social network and users' ephemeral attributes. In this work, we consider the problem of geographic coincidence prediction. The objective is to predict whether two users will meet together in the near future. Formally, the problem can be defined as:

Problem 1. **Geographic coincidence prediction.** Given a normal network $G = (V, E, \mathbf{W})$ and an ephemeral network $G'(t) = (U^t, \mathbf{X}^t, Y^t)$, the goal is to learn a predictive function:

$$f : \{G'(t), G\} \rightarrow Y^{(t+1)}$$

where $y_{ij}^{(t+1)} \in \{0, 1\}$ indicates whether user u_i and u_j will meet at time $(t+1)$.

Roughly speaking, we try to infer whether two users will gather at approximately the same place and at approximately the same time. More accurately, we say that two users u_i and u_j have a geographic coincidence (i.e., $y_{ij} = 1$) if their distance is shorter than a constant (D meters) for more than M minutes. The definition of geographic coincidence might be different in some other scenario. For example, in the MIT's Reality data set, users' coincidence are measured by Bluetooth devices.

2.2 Data Sets

We study the problem of geographic coincidence prediction on two different types of social networks: UbiComp and Reality.

UbiComp UbiComp data set is collected by Find & Connect[1], a social network platform built for participants of conferences or meetings for finding conference resources and people and connecting with them. With a positioning system based on RFID or Wi-Fi, Find & Connect records the indoor location data for each user and provides indoor location-based services such as finding where the paper or session is being held, who are the people attending the sessions, where people

[1] An ephemeral social networking system developed in Nokia Research Center.

are in the conference and when, and where was the last time that two users have met. Thereby we are able to acquire logs of physical proximity, which implies a probable encounter and interaction between users, as well as social networking connections. The system has been deployed at the UIC 2010 conference [28], Nokia GCJK internal marketing event and UbiComp 2011[5].

We use the UbiComp data set, which consists of 234 users and 69,844 location logs during the 3-day conference. The data set is divided into time intervals by day. The proximity encounters are recorded from mining locations of users equipped with RFID tags using RFID readers and a modified version of the LANDMARC algorithm [19]. Given this, we say that two users u_i and u_j have a geographic coincidence if their distance becomes shorter than D meters at a specific time, and remains within the range of $[0, D)$ for more than M minutes.[2]

Since most attendees of UbiComp are academic researchers, we can acquire their publication lists and coauthor relationships by their names in ArnetMiner[3] [24], which consists of 1,756,147 authors and 1,813,514 publications as well as the coauthor relationships between users. Finally, out of 234 UbiComp users, 206 of them are found in ArnetMiner. We thereby obtain their research profiles including their publications, co-authorship and attended conferences.

Reality Reality data set is collected from 106 users from September 2004 to June 2005 in MIT. A pre-installed software on each users' mobile phone will record their communication logs as well as physical proximity logs. The communication logs include voice calls and short messages. The physical proximity logs are recorded by Bluetooth sensor, which scans for other contacts on average every 5 minutes. If the Bluetooth sensor of a user detects another sensor at a certain time, a physical proximity event between these two users will be recorded. Reality data set contains 162,700 communication logs and more than 4 millions physical proximity logs in total. Similarly, Reality data set is divided into time intervals by day.

In addition to the geographic coincidences, the Reality data also contains the friendships between two users collected by querying the users, which form a friendship network between all the users. In Reality data set, we directly regard each proximity log as a geographic coincidence since the detection range of a Bluetooth sensor is approximately 5-10 meters, which is close enough for a geographic coincidence.

3 Observations

In this section, we conduct following observations based on the UbiComp data in order to get a better understanding on the users' behavioral patterns and structural properties of ephemeral social network:

[2] We empirically set $D = 3$ and $M = 10$, which is based on the observation in [11] and the "ten-minutes" phenomenon we discovered in observations (Cf. §3).

[3] http://arnetminer.org

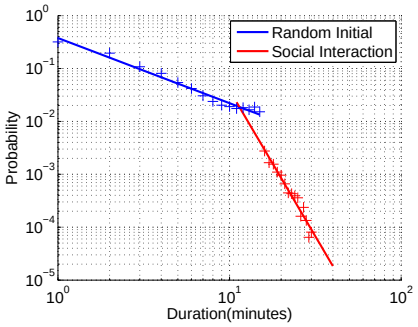


Fig. 2. Duration distribution

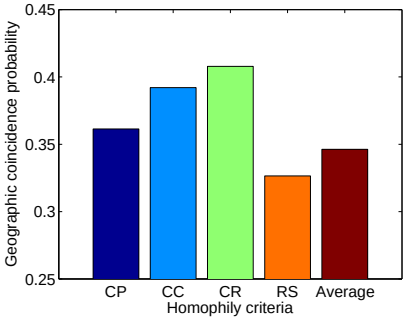


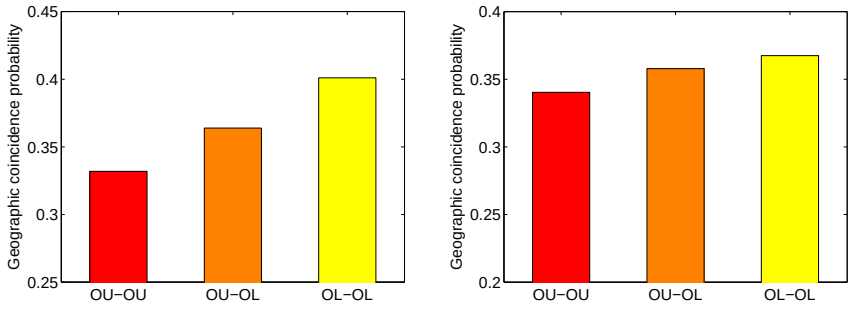
Fig. 3. Link Homophily

- *Two-stage power-law distribution.* We analyze the duration distribution of geographic coincidences and find that it satisfies a certain two-stage power-law distribution.
- *Link homophily.* How does the user similarity influence the geographic coincidences pattern?
- *Opinion leaders.* What is the role played by opinion leaders in ephemeral social network?
- *Structural hole.* Do users who span a structural hole have geographic coincidences with different people?

Two-stage power-law distribution. We first study the duration patterns of two users staying at approximately the same place and at approximately the same time. Figure 2 plots the distribution in a log-log space. It can be interestingly seen that the distribution can be described using a two-stage power-law and 10 minutes seems to be an inflexion point. When the duration time is less than 10 minutes, the exponent of the corresponding power-law is -1.2315, while, when the duration time increases to more than 10 minutes, the exponent becomes -5.5221. The phenomenon implies that a large portion of coincidences might be random based on users' location. For example acquaintances generally say hello when they meet and make small talk (less than 10 minutes). On the other hand, targeted meet may last a longer time.

Based on this observation, we set the duration threshold as $M = 10$ for the definition of geographic coincidence on UbiComp data set. That is to say, on Ubicomp data set we only consider geographic coincidences longer than 10 minutes since they are more likely to indicate actual social interactions.

Link homophily The principle of homophily [16] points out that users with higher similarity are more likely to establish relationships. In this work we mainly study the similarity in research area since most of the users are researchers and they attend the conference in order to get feedback or new collaboration in academia. The following criteria are employed to measure users' research similarity: (1) *Coauthored paper count (CP)*: It counts the coauthored publication number for each pair of users; (2) *Common coauthor count (CC)*: It counts the



(a) Publication count (b) H-index

Fig. 4. Observation of geographic coincidences between opinion leaders (OL) and ordinary users (OU).

number of common coauthors between two users; (3) *Common conference ratio (CR)*: We construct conference vectors for all users with their attendance times of different conferences. The common conference ratio is the cosine similarity of two users' conference vectors; (4) *Research similarity (RS)*: Jaccard similarity of the research interests of two users.

We rank all the user pairs by the above criteria and calculate the geographic coincidence probability of the top 600[4] pairs of users. The average geographic coincidence probability is also calculated for comparison, as shown in Fig. 3. We can observe that user pairs with highest CP, CC, or CR are more likely to have geographic coincidences than average. These results are expected. Users with more coauthored papers have direct connections between them and thus are more likely to meet each other; more common coauthors implies a strong effect of triadic closure [10], which influences the geographic coincidence probability and attending more common conferences increases their chance to know each other. However, a surprising observation is that geographic coincidence probability of user pairs with highest research similarity are approximately 2% lower than the average probability. This result indicates that attendees of an academic conference may tend to talk with people that have different research interests in order to get new ideas.

Opinion leader The two-step flow theory [2, 14, 16] suggests that ideas usually flow first to "opinion leaders" and then to more people from them. There are several algorithms to detect opinion leaders in social networks. In this work we use two different indicators to define opinion leaders: publication count and H-index. We rank all the users by their publication count or H-index and take the top 25% as opinion leaders. Fig. 4 presents the comparison of geographic coincidence probability between different types of user pairs. It is clearly shown that ordinary users (OU) and opinion leaders (OL) are more likely to have a geographic coincidence than two ordinary users, which implies that people tend to communicate with opinion leaders. We also find that two opinion leaders have

[4] Probability of top 200, 400 pairs of users yields similar results.

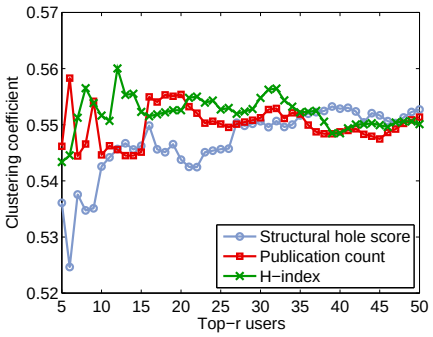


Fig. 5. Clustering coefficient

the highest probability of geographic coincidence. This is expected because in an academic conference, opinion leaders are more willing to exchange ideas and hence have more direct interactions.

Structural hole In a social network, a person is called to span a *structural hole* if she is connected to two people in different parts of networks that are otherwise not well connected to each other [3]. It is claimed that such nodes have an informational advantage with connection to people who are not linked to each other, and hence are exposed to a more diverse source of ideas. An interesting question is, whether a person who spans a structural hole in coauthor network will also present a higher diversity in its geographic coincidence pattern? In this paper, we simply define node A's "structural hole score" in a coauthor network by the number of author pairs (B, C) which satisfies that A is the only common coauthor. We rank all the users by their structural hole score, and calculate the average clustering coefficient of top-r users over ephemeral social networks of all the time intervals in UbiComp data set. We also rank the users by publication count and H-index to provide a comparison to opinion leaders. The result is presented in Fig. 5. It is shown that users with structural hole score ranking in the top 20 tend to have lower clustering coefficient (confirmed by paired t-test with 95% significance), but it turns out to be close to the average when taking the top 50 users into account. The clustering coefficient of opinion leaders, however, always remain consistent with average level. It indicates that users who have a higher structural hole score also tend to have geographic coincidences with a wide variety of people, but the difference is slight since in an ephemeral social network, a larger proportion of users seeks for new relationships and hence have geographic coincidences with various people.

4 Factor Graph Model

We employ a factor graph model to predict the geographic coincidences between users. The basic idea is to construct a graphical model by modeling each pair of users as a node. We then define different types of factor functions to incorporate different factors into the prediction task, and define an objective function based

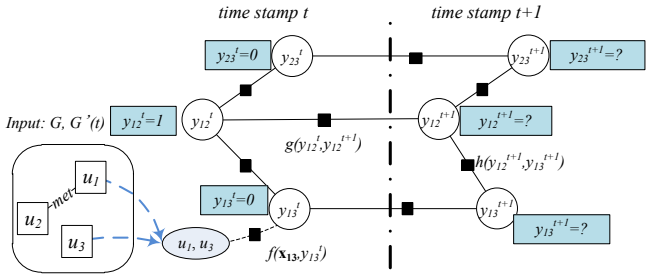


Fig. 6. A graphical representation of factor graph model

on the joint probability of the factor functions. The model can be trained by optimizing the objective function.

As Fig. 6 demonstrates, at time t, we map the event of geographic coincidence of every pair of users (u_i, u_j) as a node y_{ij}^t in our graphical model, corresponding to an event to predict in the ephemeral social network G^t. For the labeled set Y_L, we let $y_{ij}^t = 1$ if $y_{ij}^t \in Y^t$, otherwise $y_{ij}^t = 0$. For unlabeled edge set Y_U, we have $y_{ij}^t =?$ to predict. We use Y to represent the global set of all y_{ij}^t. The factor graph model was previously used for inferring social ties in social networks [25].

We define three different kinds of factor functions as follows:

- **Attribute factor function** $f(\mathbf{x}_{ij}, y_{ij}^t)$. It incorporates the attribute value $\mathbf{x}_{ij}$ of each pair of users corresponding to y_{ij}^t, where $\mathbf{x}_{ij} = [\mathbf{w}_{i\cdot}, \mathbf{w}_{j\cdot}]$ combines the attribute vector of both users.
- **Temporal correlation factor function** $g(y_{ij}^t, y_{ij}^{t+1})$. It represents the temporal dependencies between the geographic coincidences indicated by y_{ij}^t and y_{ij}^{t+1}.
- **Social correlation factor function** $h(Y_c^t)$. Y_c^t represents a clique which consists of a set of y_{ij}^t. It leverages the social correlation between user pairs.

The three factor functions can be instantiated in different ways. In this work, we define them as exponential-linear functions. Formally, we define the attribute factor function as

$$f(\mathbf{x}_{ij}, y_{ij}^t) = \frac{1}{Z_1} \exp\{\alpha^T \mathbf{\Phi}(\mathbf{x}_{ij}, y_{ij}^t)\} \tag{1}$$

where α is the weighting vector; $\mathbf{\Phi}(\mathbf{x}_{ij}, y_{ij}^t)$ is the feature vector function.

The temporal correlation factor function can be defined as

$$g(y_{ij}^t, y_{ij}^{t+1}) = \frac{1}{Z_2} \exp\{\beta^T \mathbf{g}(y_{ij}^t, y_{ij}^{t+1})\} \tag{2}$$

where β is the weighting vector; $g(y_{ij}^t, y_{ij}^{t+1})$ is an indicator function.

We define the social correlation factor function in a similar way

$$h(Y_c^t) = \frac{1}{Z_3} \exp\{\lambda^T \mathbf{h}(Y_c^t)\} \tag{3}$$

where λ is the weighting vector; $\mathbf{h}(Y_c^t)$ is an indicator function, taking the geographic coincidences of a clique of user pairs as input. Z_1, Z_2 and Z_3 are normalizing factors. This definition of factor function was often used in a graphical models such as Markov Random Fields [12] or Conditional Random Fields [15].

The joint distribution over all the Y can be written as

$$P(Y|G,G') = \frac{1}{Z}\exp\{\sum_t\sum_{y_{ij}^t}\alpha^T\mathbf{\Phi}(\mathbf{x}_{ij},y_{ij}^t) + \sum_{i,j}\sum_t\beta^T\mathbf{g}(y_{ij}^t,y_{ij}^{t+1}) + \sum_t\sum_{Y_c^t}\lambda^T\mathbf{h}(Y_c^t)\} = \frac{1}{Z}\exp\{\theta^T\mathbf{S}\} \tag{4}$$

where $\theta = [\alpha^T,\beta^T,\lambda^T]^T$ is the parameter vector;
$\mathbf{S} = [\sum_t\sum_{y_{ij}^t}\mathbf{\Phi}(\mathbf{x}_{ij},y_{ij}^t), \sum_{i,j}\sum_t\mathbf{g}(y_{ij}^t,y_{ij}^{t+1}), \sum_t\sum_{Y_c^t}\mathbf{h}(Y_c^t)]^T$ denotes all the features and Z is the normalizing factor.

Model Learning We learn the FGM by estimating the parameter configuration θ to optimize the log-likelihood of observed data. The observed data could be incomplete and thus pose challenges to model learning. We regard the entire factor graph as a partially labeled graph. Let Y_L denote the set of known geographic coincidences, and Y_U as the set of unknown geographic coincidences. The learning task can be formally described as to find a parameter configuration θ^* such that $\theta^* = \text{argmax}_\theta P(Y_L|G,G')$.

We define the log-likelihood as the objective function

$$\mathcal{O}(\theta) = \log P(Y_L|G,G') = \log\sum_{Y|Y_L}\exp\theta^T\mathbf{S} - \log Z = \log\sum_{Y|Y_L}\exp\theta^T\mathbf{S} - \log\sum_Y\exp\theta^T\mathbf{S} \tag{5}$$

A gradient decent method (Newton-Raphson method) is used to optimize Eq. 5. The gradient for each parameter is

$$\frac{\partial\mathcal{O}(\theta)}{\partial\theta} = \frac{\sum_{Y|Y_L}\exp\theta^T\mathbf{S}\cdot\mathbf{S}}{\sum_{Y|Y_L}\exp\theta^T\mathbf{S}} - \frac{\sum_Y\exp\theta^T\mathbf{S}\cdot\mathbf{S}}{\sum_Y\exp\theta^T\mathbf{S}} = \mathbb{E}_{P(Y|Y_L,G,G')}\mathbf{S} - \mathbb{E}_{P(Y|G,G')}\mathbf{S} \tag{6}$$

We use Loopy Belief Propagation (LBP) to approximate the gradient and update θ iteratively.

Predicting geographic coincidences With the learned parameter configuration θ, the prediction task is to find a Y_U^* which optimizes the objective function, i.e., $Y_U^* = \text{argmax}_{Y_U}P(Y|G,G')$.

We employ similar methodology in this optimization task. Instead of calculating the joint probability, we calculate the marginal probability for each $y_{ij}^{(t+1)}$ and predict them as positive when the marginal probability is greater than 0.5, otherwise the event will be predicted as negative.

Table 1. Statistics of UbiComp and Reality data sets

Data set	Users	Labeled samples	Unlabeled samples
UbiComp	243	17,391	23,871
Reality	106	7,384	2,140

5 Experimental Results

5.1 Experimental setup

Data sets We validate the effect of our proposed model on two different data sets: *UbiComp* and *Reality*, and compare the result with two baseline methods. A brief statistics of the data sets is shown in Table 1.

UbiComp data set includes user location logs on September 19th and September 21st. In this work, we divide the data set into two time intervals, namely the two days of the conference. We regard all the geographic coincidences on September 19th as labeled and predict the geographic coincidences on September 21st.

For Reality data, We select 12 consecutive days, each with more than 100 communication logs for our experiments. Then we define the first 10 days as labeled. The task is to predict the geographic coincidences in the last 2 days.

Baseline methods We define two baseline methods for the geographic coincidences task.

- *SVM.* This method only uses the users' attribute to train SVM and to predict the geographic coincidences.
- *CRF.* We consider the time correlation and establish sequential conditional random fields for each user pair.

We evaluate the performance of geographic coincidence inference in terms of precision, recall and F1-score.

Factor definitions For both data sets, we define the temporal correlation factors between two consecutive time intervals for each user pair.

In UbiComp data set, we also define four different types of social correlation factors according to the principle of homophily (Cf. Section 3): if two users are similar in some aspects, they will be more likely to have geographic coincidence with the same person. To define the social correlation factors based on homophily of coauthored paper count (CP), we first rank all the user pairs by CP and select those within top 150, denoted by (u_i, u_j). Then for every other user u_n, we add social correlation factors *CPInf* between e^k_{in} and e^k_{jn}. The other three homophily-based social correlation factors *CCInf*, *CRInf* and *RSInf* can be defined similarly.

In Reality data set, we define social correlation factors based on the structural balance theory [10]. It suggests that people in a social network tend to form into a balanced network structure. To be specific, for a triad, the balance theory

Table 2. Prediction performance comparison(%)

Date set	Method	Precision	Recall	F1-score
UbiComp	SVM	**34.5**	20.4	25.6
	CRF	33.2	39.4	36.0
	FGM	34.0	**65.4**	**44.7**
Reality	SVM	84.1	64.4	72.9
	CRF	73.6	**85.8**	79.2
	FGM	**85.1**	81.0	**83.0**

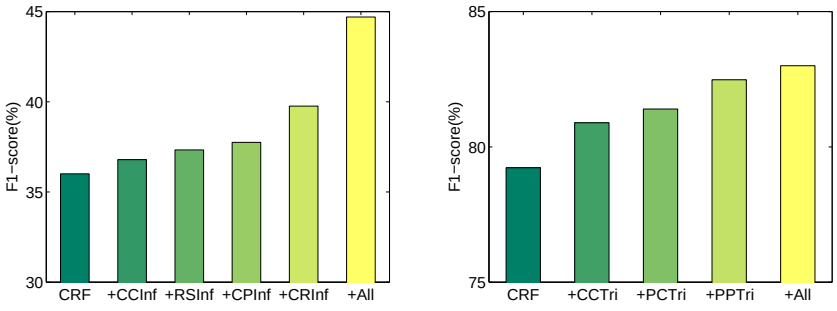
claims that all of the three users or only one pair of them should be friends. We employ two kinds of connection, physical proximity connection and calling connection (voice call or SMS), to identify triads. Then for each user pair (u_i, u_j) in a triad, we establish a social correlation factors involving all the three user pairs. Since there are two types of connections, we can define three different social correlation factors regarding the connection types of the other two user pairs: *CCTri* (both calls), *PCTri* (one call, one physical proximity) and *PPTri* (both physical proximity).

5.2 Results and discussion

Performance comparison We compare the prediction performance between our methods and the baselines, as shown in Table 2. It is shown that our model outperforms other methods in both two data sets. In UbiComp data set, FGM achieves an improvement of approximately 8-19% in terms of F1-score compared to the baselines, and also improves recall by approximately 26-45%. Although SVM shows a higher precision of 34.5%, the precision of the proposed model is very close to the baseline (34.0%). In Reality data set, FGM also gives a rise of 4-10% compared to the baselines in terms of F1-score. In addition, FGM achieves the highest precision among all the methods. We can also observe the effect of time correlation, employed by CRF. The time correlation factor improves the F1-score of CRF by about 10% in UbiComp data set and approximately 7% in Reality data set.

Contribution of social correlation factors To further investigate the contribution of different social correlation factors in the prediction task, we remove all the social correlation factors and evaluate the performance by adding each of them individually into the model. Thereby we can measure their contribution by the improvement they achieve to F1-score, as shown in Fig. 7.

It is shown that in both data sets, all social correlation factors improve the performance. In UbiComp data set, CRInf factor contributes the most to F1-score amongst the four social correlation factors by an average improvement of 3%. It implies that users who often attend common conferences may have a stronger implicit correlation since they probably have been in the same ephemeral social network before. Its effect is even stronger than those with ex-



(a) UbiComp (b) Reality

Fig. 7. Analysis of social correlation factors on F1-score

plicit coauthorship (CPInf). The effect of CCInf, RSInf and CPInf factors are also observable.

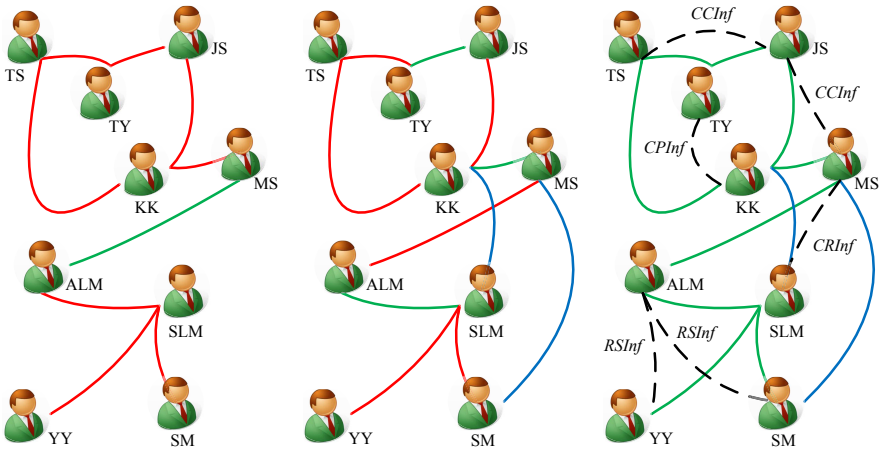
In Reality data set, it can be observed that PPTri achieves the highest improvement of approximately 3%. It implies that ephemeral social network probably obeys the structural balance theory and indeed helps improve the performance. PCTri and CCTri also contribute significantly to the improvement of performance. It indicates that joint with relationships in normal social network such as mobile social network, triadic social correlation factors based on structural balance theory still contributes to the prediction performance.

Case study We further conduct a case study to investigate why our proposed model outperforms other baseline methods. Fig. 8 presents the prediction result on a subset of UbiComp data generated by three different approaches: SVM, CRF and FGM. Green solid lines represent true positive samples; red solid lines for false negative samples and blue solid lines for false positive samples. In addition, we use black dash lines to point out the social correlations between users.

It can be observed that CRF tends to predict more geographic coincidences than SVM with the help of time correlation. It successfully detects more geographic coincidences (e.g. JS-TY and MS-KK), albeit few of them are incorrect (Cf. Fig. 8(b)). Our proposed approach further leverages the social correlation factors to improve the prediction result. For example, when MS and JS have a higher common coauthor count, geographic coincidences of MS and KK may increase the chance of a geographic coincidences between JS and KK. Our proposed model is able to capture such social correlations and infer the geographic coincidences between KK and JS from the prediction between MS and KK. The social correlation factors benefit the prediction result of FGM by significantly improving the recall, as shown in Fig. 8(c).

6 Related work

Dynamic behavior analysis There are several works on social dynamic behavior analysis. Zhang et al. [26] proposed a dynamic continuous factor graph



(a) SVM (b) CRF (c) FGM

Fig. 8. Case study

model to predict users' emotion states. Tan et al. [21] proposed a noise tolerant model for predicting user's actions in online social networks. Tang et al. [23] proposed a topical affinity propagation to quantify the social influence between users. However, these works did not leverage location information, while we focus on predicting geographic coincidences.

User mobility analysis Quite a few works on user mobility analysis have been conducted. Li et al. [17] designed a hierarchical-graph-based similarity measurement for estimating user similarity based on their location history. Liu et al. [18] proposed an approach to utilize information of mobile objects for the clustering task. Qian et al. [20] explore co-location mining pattern with dynamic neighborhood constraint. However, rather than analysis of user mobility, we focus on a prediction problem. Cho et al. [6] develop a Gaussian model by incorporating periodicity and influence of social network structure to predict human location tracks. Crandall et al. [7] studies geographic coincidences between users to infer social ties, while our work focus on prediction of geographic coincidences from social network. Zheng et al. [27] used a graph-based algorithm to infer user mobility based on GPS data. Tang et al. [25] developed a general learning framework for inferring the types of social ties in social networks; and [22] further extended the problem across heterogeneous networks. But none of these works provide an approach for prediction of interpersonal geographic coincidences.

Physical proximity analysis Physical proximity has been employed in many works to quantify users' behaviors. Eagle et al. [8] use GPS on mobile phones to analyze proximity of the users in order to present the properties of users' location tracks. However, different from tracking users' mobility, we aims to predict geographic coincidences between users in this work. There is also a host of conference proximity analysis in current literature. Isella et al. [13] use RFID badges to collect face-to-face proximity data of individuals at a scientific conference, and

analyze its static and dynamic properties. Atzmueller et al. [1] explore different roles of participants in a conference by examining their face-to-face interaction patterns. Similarly, Cattuto et al. [4] collect data from the office environment and academic congress.

7 Conclusion

In this paper, we formally define the ephemeral social network and study to which extent we can predict geographic coincidences in an ephemeral social network. We conduct a series of observations on an ephemeral social network extracted from a data collected during an academic conference (UbiComp 2011). Based on link homophily, opinion leader and structural hole, we show the interplay between the normal social network (coauthor network) and users' behavioral pattern in the ephemeral social network. We then propose a Factor Graph Model (FGM) for the prediction task. Experimental results show that our model outperforms baseline on two data sets: UbiComp and Reality. Further analysis also suggests that social correlation factors help improve the performance.

A limitation of this work is that a geographic coincidence does not necessarily indicate an actual social interaction, e.g. conversation or discussion. We carefully select the parameters so that extracted geographic coincidence are very likely to be accompanied with actual social interaction, but the real situation is hard to detect without collecting additional context. Another flaw is the requirement of labeled data since we use supervised learning for our model. An unsupervised learning approach would further reduce the cost of geographic coincidences prediction.

References

1. M. Atzmueller, S. Doerfel, A. Hotho, F. Mitzlaff, and G. Stumme. Face-to-face contacts during a conference: Communities, roles, and key players. In *Workshop MUSE at ECML/PKDD*, 2011.
2. A. Booth. Personal influence networks and participation in professional association activities. *The Public Opinion Quarterly*, 33(4):pp. 611–614, 1969.
3. R. S. Burt. *Structural holes: The social structure of competition.* Harvard University Press, Cambridge, MA, 1992.
4. C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS ONE*, 5(7):e11596, 07 2010.
5. A. Chin, B. Xu, F. Yin, X. Wang, W. Wang, X. Fan, D. Hong, and Y. Wang. Using proximity and homophily to connect conference attendees in a mobile social network. In *Accepted to the 2nd International Workshop on Sensing, Networking and Computing with Smartphones*, pages 1–9. IEEE Computer Society, 2012.
6. E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD*, pages 1082–1090, 2011.
7. D. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 107(52):22436, 2010.

8. N. Eagle and A. Pentland. Social serendipity: Mobilizing social software. *IEEE Pervasive Computing*, 4(2):28–34, Apr. 2005.
9. N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *PNAS*, 106(36):15274, 2009.
10. D. A. Easley and J. M. Kleinberg. *Networks, Crowds, and Markets - Reasoning About a Highly Connected World*. Cambridge University Press, 2010.
11. E. T. Hall. A system for the notation of proxemic behaviour. *American Anthropologist*, 65:1003–1026, 1963.
12. J. M. Hammersley and P. Clifford. Markov field on finite graphs and lattices. *Unpublished manuscript*, 1971.
13. L. Isella, J. S. A. Barrat, C. Cattuto, J.-F. Pinton, and W. V. den Broeck. What's in a crowd? analysis of face-to-face behavioural networks. *Journal of Theoretical Biology*, pages 166–180, 2010.
14. E. Katz. The two-step flow of communication: An up-to-date report on an hypothesis. *The Public Opinion Quarterly*, 21(1):pp. 61–78, 1957.
15. J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.
16. P. F. Lazarsfeld, B. Berelson, and H. Gaudet. *The People's Choice. How the Voter Makes up his Mind in Presidential Campaign*. Columbia University Press, New York, 1944.
17. Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W. Ma. Mining user similarity based on location history. In *Workshop on Advances in Geographic Information Systems*, 2008.
18. S. Liu, Y. Liu, L. M. Ni, J. Fan, and M. Li. Towards mobility-based clustering. In *KDD*, pages 919–928, 2010.
19. L. M. Ni, Y. Liu, Y. C. Lau, and A. P. Patil. Landmarc: indoor location sensing using active rfid. *Wireless Networks*, 10:701–710, November 2004.
20. F. Qian, Q. He, and J. He. Mining spatial co-location patterns with dynamic neighborhood constraint. In *ECML/PKDD*, pages 238–253, 2009.
21. C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In *KDD*, pages 1049–1058, 2010.
22. J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogeneous networks. In *WSDM'12*, pages 743–752, 2012.
23. J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD*, pages 807–816, 2009.
24. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *KDD*, pages 990–998, 2008.
25. W. Tang, H. Zhuang, and J. Tang. Learning to infer social ties in large networks. In *ECML/PKDD'11*, pages 381–397, 2011.
26. Y. Zhang, J. Tang, J. Sun, Y. Chen, and J. Rao. Moodcast: Emotion prediction via dynamic continuous factor graph model. In *ICDM*, pages 1193–1198, 2010.
27. Y. Zheng, Q. Li, Y. Chen, X. Xie, and W. Ma. Understanding mobility based on GPS data. In *Ubiquitous Computing/Handheld and Ubiquitous Computing*, pages 312–321, 2008.
28. L. Zhu, A. Chin, K. Zhang, W. Xu, H. Wang, and L. Zhang. Managing workplace resources in office environments through ephemeral social networks. In *UIC*, pages 665–679, 2010.