From Face-to-Face Gathering To Social Structure

Chunyan Wang
Dept. of Applied Physics
Stanford University
California, USA
chunyan@stanford.edu

Mao Ye
Social Computing Group
HP Labs
California, USA
mxy177@cse.psu.edu

Wang-chien Lee
Dept. of Computer Science
and Engineering
The Pennsylvania State
University
Pennsylvania, USA
wlee@cse.psu.edu

ABSTRACT

The rapid development of on-line social networking sites has dramatically changed the way people live and communicate. One particularly interesting phenomena came along with this development is the prominent role of various on-line networking portals played in scheduling and organizing off-line group events and activities. In this paper, we focus on studying the face-to-face(f2f) group formed through, or facilitated by, on-line portals. We first show the distinct characteristics of such f2f groups by analyzing datasets collected from Whrrl and Meetup. Next, we propose a dynamic model for group gathering based on the process of friend invitation to interpret how a f2f group is formed on-line. The results of our model are confirmed by empirical observations. Finally, we demonstrate that using such group information can effectively improve the accuracies of social tie inference and friend recommendation.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavior Sciences; G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms

Human Factors,Theory, Measurement

Keywords

group-gathering, social networks, social tie inference

1. INTRODUCTION

The process by which people come together, attract new members, and form groups has long been regarded as a crucial research issue due to its potential impacts on opinion formation, viral marketing and decision making [1, 2, 3, 4]. In the Web, characterization of various on-line groups has also attracted a lot of research interests recently. Existing works on on-line groups mostly focus on studying "cyber groups" (e.g., on-line communities [5, 6], social media websites [7, 8] and forums [9]), where group members interact with each other and thus do not need to present themselves at a "physical location". One special kind of groups largely unexplored thus far is the ones that are formed through on-line portals but taking place off-line at a real physical location. In fact, recent growth of social networking sites such as Meetup[1], Whrrl[2] and Facebook[3] have greatly facilitated people to schedule and organize such off-line group activities, from celebration parties, social networking banquet, to even larger events, such as demonstrations. Before, during and after the physical activities, users may share information such as location and group members through these social websites, which results in an unprecedented opportunities for researchers to better understand the behavior of group user at geographical locations.

Different from on-line "cyber groups" and individual activities in location based social networks (LBSN) [10, 11, 12, 13, 14, 15], the f2f group (group-gathering) in this study is constrained not only by spatial and temporal factors, but also social factors like influences from each other in the group. It would thus be interesting to better understand the user behaviors in these aspects and to answer the following research questions: (i) What are the differences between group activities compared with individual ones in the LBSN? (ii) What is the social structure within such a group? (iii) How is the group for face-to-face gathering formed through on-line portals? (iv) What kind of information can we extract from studying the characteristics of group-gatherings? In this paper, we focus on addressing these problems through investigating the "group check-in" dataset form Whrrl and the "group invitation" data form Meetup. The contributions of our study can be summarized in three complementary dimensions:

- **Characterization of the f2f groups.** We demonstrate the basic characteristics of group-gathering discovered from analysis of real datasets. Interestingly, we find that group-gathering activities exhibit striking differences compared with individual activities, both in spatial and temporal aspects. By further examining the social structures within groups, we find that many people joining the same group do not actually know

[1]http://meetup.com
[2]http://whrrl.com
[3]http://facebook.com

each other beforehand. The probability of friendship is inversely related with the size of the group.

- **Modeling of the f2f group formation process.** We propose a model to capture the dynamic process of f2f group formation based on an on-line friend invitation process. Based on our model, the size distribution of groups is expected to follow a power law distribution, which is verified by empirical observations. The model also reproduces the exponential duration of invitation process, and the inverse relationship between group size and the probability of friends.

- **Applications of the f2f group formation model.** We demonstrate that information exhibited in group-gathering activities has an implication regarding both the social closeness and geographical proximity features. We apply group information into the problem of social tie inference and friend recommendation. Through extensive experiments on real datasets, we show that such group information can significantly improve existing methods for social tie inference and friend recommendation.

The rest of the paper is organized in the following way. Section 2 introduces the observed characteristics of group-gathering. Section 3 introduces our model for group formation based on on-line friend invitation. Section 4 applies the group-gathering information to the problem of social tie inference and friend recommendation. Section 5 reviews the related work. Section 6 concludes the paper with discussions and future work.

2. CHARACTERISTICS OF GROUP

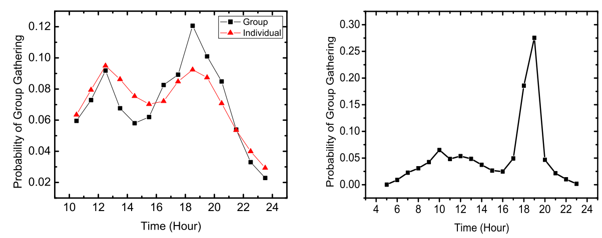
In this section, we focus on exploring interesting characteristics of group-gathering, by utilizing the check-in records in an LBSN site, Whrrl, as well as an on-line social networking portal for scheduling off-line events, Meetup. From the Whrrl dataset, we are able to compare differences between individual check-ins and group ones from the temporal, spatial and social aspects. The Meetup dataset also helps us to observe and confirm group-gathering patterns. We also present an analysis of social structures within the group at the end of this section.

2.1 Data

Whrrl is a popular LBSN that allows users to explore, rate and share points-of-interest. It also allows users to declare friendship among socially connected users and visits through check-ins at a physical place. Users could check in by using a mobile application on a GPS-equipped smart phone. Types of places include restaurant, hotel, bookstore and so on. A distinctive feature of this dataset is that a user could check-in by herself or with a group of other people, which is referred as *group check-in*. Users of the site are identified by unique user-ids. In our study, we crawl a friendship network consisting of 24,002 users and 145,228 social ties and collect the check-in records of these users' activities from January 2009 to January 2011. The undirected graph has an average degree of 12.101 and an average shortest-path length 4.718, which is a typical small-world social network. We crawled all the 357,393 check-in records over 120,726 different places associated with these users in the observation period. For each check-in record, we have information such as the exact location (i.e., longitude and latitude), time of check-in and the users involved (i.e., there may be more than one user-id for group check-ins).

Another dataset included in our study is from Meetup, an on-line social networking portal for scheduling off-line meetings. We collect over 11,423 scheduled group events taking place in Boston, Cupertino, Las Angeles, New York, Orlando and San Francisco. For each of these group activities, we keep records of the scheduled time, location, the number of people responding to attend, and also the time of their responses to the invitation. Different from Whrrl, the friendship structure on Meetup is not explicitly available, which is typical for many on-line and off-line communities.

2.2 Temporal and Spatial Characteristics



(a) Whrrl (b) Meetup

Figure 1: Density plot of group-gathering and individual activities at different hours of the day in (a) Whrrl and (b)Meetup.

We start by investigating the temporal and spatial characteristics of group-gathering. As mentioned before, one interesting feature of the group in our study is that it requires its members to appear at a specific time and a real location. Natural questions to ask regarding such groups are their temporal and spatial patterns. For example, what is the peak hour for group activities? How far group members would travel from their center of movement to join the group? Are they willingly to travel more or less for group activities compared with an individual one? To answer these questions, we divide our check-in dataset from Whrrl into two categories: individual check-in records and group check-in records. To study the temporal pattern of group-gathering, we make a density plot of an individual check-in and a group check-in at any given hour in the day, as shown in Figure 1. In this figure, we use black squares to denote group activities, and red triangles to denote individual ones. It could be observed from Figure 1(a) that there exists two peaks in the noon (around 13:00pm) and evening (around 19:00pm) for activities. More interestingly, the evening peak for group activities is higher than that of individual check-ins. This bias in evening hours for group activities is further confirmed by the time pattern of group activities in Meetup dataset as shown in Figure 1(b). One possible explanation is that the scheduled time for group events is constrained by all the members in the group. While it is more likely for all the members to be available in the evening than in the noon, a constraint is weaker for individual activities.

Another interesting observation about group-gathering is its spatial patterns. Particularly, we are interested in look-

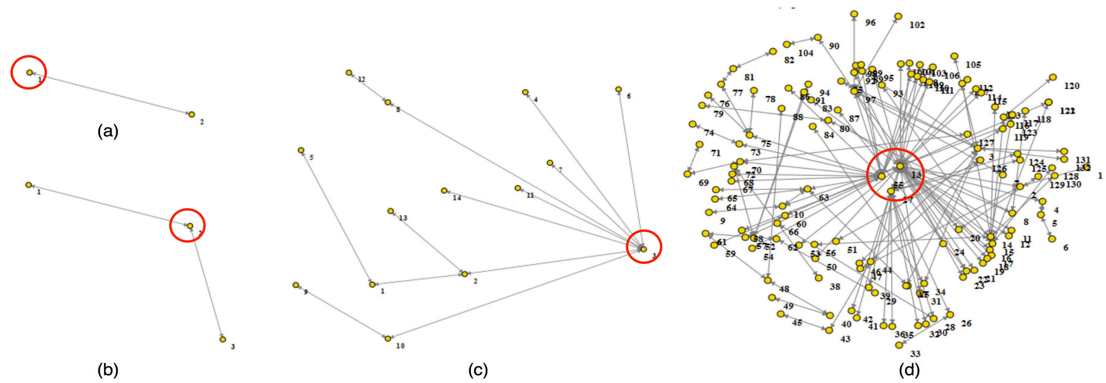
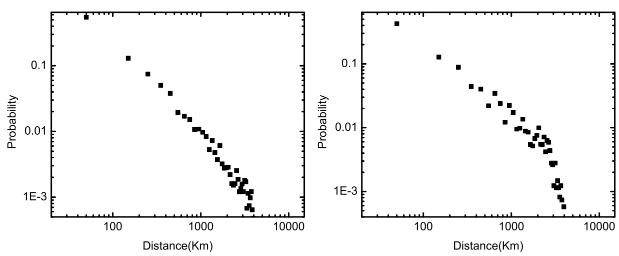


Figure 3: Four examples of social network structures within group-gathering events with different group sizes with (a) 2, (b) 3, (c) 14, (d) 135 from Whrrl. Red circles in the plot symbolifies the hubs in the social graph.



(a) Individual Activity (b) group-gathering

Figure 2: Density plot of the distance travelled from center of movement for (a) individuals and (b) groups.

ing at the geographical distances from the user's home location to places of check-ins. Here, center of all check-ins belonging to a certain user is used as an approximation of the user's home location [14]. For each check-in, we calculate its geographical distance from the center of movement. In Figure 2, the probability of distance traveled from home for individual activities (Figure 2(a)) and group events (Figure 2(b)) is plotted in a log-log scale. The linear relationship suggests a power-law distribution of the mobility pattern for individual movements. Using the maximum likelihood estimation (MLE), the exponent of power-law distribution has a value of -1.593. This result is consistent with prior works related to the levy flight of human mobility [16, 17, 10, 14, 18]. More importantly, for group-gathering events, a sharp drop in probability is noticeable when the distance from home reaches the spatial bound. The distribution of distance from home distinctly deviates from that of individuals with a shorter tail. This observation suggests a spatial bound from home location for people to join group events. One possible explanation for this observation is that people with a high degree of spatial proximity are more likely to form a group-gathering event. Due to the spatial bound coming from other group members, people are less freedom than they were alone. We will further discuss the interplay between group-gathering and spatial proximity in Section 4. For now, we focus on describing characteristics of group-gathering itself.

2.3 Social Characteristics

Thus far, we have examined the temporal and spatial characteristics of group-gathering. We found that group-gathering has special characteristics constrained both temporally and spatially, which characteristics distinct such f2f group from other on-line groups. Given a group or community, another aspect of interests is the social structure within the formed group. Next, we turn our attention to the social aspects of group-gathering both in microscopic and macroscopic scale. Figure 3 shows the detailed structure of friendship network graphs of four sampled group-gathering activities of different group sizes, ranging from (a) 2, (b) 3, (c) 14 to (d) 135 in Whrrl. What is different from our expectations is that not all the people attending the same activity know each other (declare friendship on-line) before they join the group. In fact, for large groups, as in Figure 3(d), most of the individuals joining the group do not have links with each other. One interesting phenomenon for group-gathering is the existence of hubs (marked by red circle) observed from the social graph. We define such hubs as the user in the group who has a social tie out-degree significantly larger than the average of other nodes in the group. In Figure 3(a), either of the two users can be the hub. The existence of such hubs itself could be related to the inequality in social status of groups. Here, we explain this observation by the following formation process of group-gathering: when one or a few persons want to initiate a group activity, the process starts by them asking their friends to join the event. As the group grows, the initiators of the event would thus know more people than others. In this way, the social graph would most likely have initiators well connected as hubs surrounded by individuals later invited.

From the perspective of macroscopic properties, we find that the probability of friendship scales inversely with the number of people joining the group. The red line in Figure 4 plots a sharp drop in probability of friendship within the group as the group size increasing. This drop suggests that the connectivity within a group becomes weaker for larger group activities compared with smaller ones. Nevertheless, even for very large groups, the probability of friendship re-

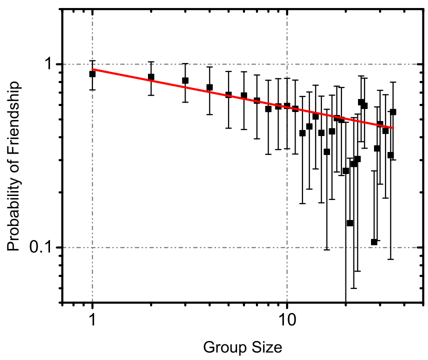


Figure 4: Probability of friendship within the aggregated group is inversely related to the group size.

mains over 0.20, which is much higher than 2.42×10^{-5}, the measured probability for two random nodes of being friends in the network. This fact suggests that even though people attending a group-gathering do not necessarily know each other, joining a group together itself is a strong indicator of friendship. We further study the strength of a social tie between two persons as a function of number of times the two persons joined group together in Figure 5. We use the number of common friends divided by the total number of friends of a pair of users as a measurement of tie strength. As can be seen from Figure 5, without much surprise, the social tie strength increases as the number of times people joined the same group and saturates around 4. On one hand, the observation suggests the important role of group activity in the process of friendship formation. On the other hand, the fact that a very small number of co-group can lead to orders-of-magnitude greater probabilities of friendship suggests the need for a deeper understanding of the underlying group formation mechanism and also further explorations of its potential application, as will be described in the following sections.

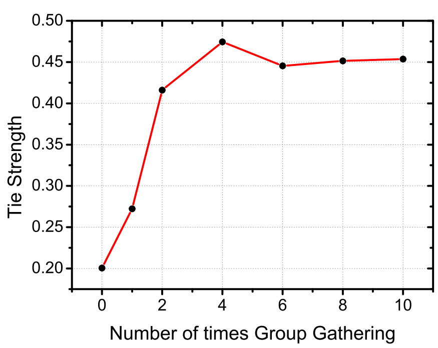


Figure 5: Social tie strength as a function of number of times group-gathering.

3. MODEL FOR F2F GROUP

In the last section, we describe our observations about the distinct characteristics of group-gathering from different aspects. From the observations, especially the social structure aspect, we find interesting properties such as the existence of hubs, the inverse scaling relationship between probability of friendship and group size. One would naturally ask how this kind of off-line groups is formed through on-line portals. A simple dynamic model that could capture the key structures of the group and also reproduce the observed properties in this case can help us to better understand the formation of group-gathering. Thus, we aim to capture the group formation process by proposing a model based on online invitation process. In the following, we demonstrate that the properties of group-gathering we described are robust, in that they can arise even on very simple assumptions of group formation. Our model capture these basic effects, and also yields a power-law scaling of group size distribution, which is generally recognized as one of the basic properties for group behavior [19, 20].

We begin by assuming that for each group event, there is one or a few initiators of the group, e.g., the first person who starts an event by asking his friends to join at time 0. At any point of time t, users in the existing group would attempt to invite one more friend to join the group. For simplicity, we also assume that the person receiving invitation has a fixed probability α to join. We use N to denote the total number of users in the network, $k(t)$ to denote the number of people in the group at time t, K to denote the final size of the group. We simplify the problem by considering the case that each user has the same number of friends and the same probability to accept invitation. If there are $k(t)$ individuals already in the group, the increment of group size at the time t should be $\alpha k(t)$. We have

$$\frac{\partial k(t)}{\partial t} = \alpha k(t). \tag{1}$$

Given that when $t = 0$, $k(0)$ is equal to 1, we have $k(t) = e^{\alpha t}$ and $K = e^{\alpha T}$, in which T is the duration of the invitation process. Thus we have

$$\Pr(K < l) = \Pr(T < \frac{\ln l}{\alpha}). \tag{2}$$

If there is a probability λ for the group to cease the inviting new users at each time step, it naturally follows that the distribution of T should be a geometric distribution with

$$\Pr(T = m) = (1 - \lambda)^{m-1}\lambda. \tag{3}$$

When m is large and λ is small, the distribution could be approximated by its continuous counter part, the exponential distribution. So in our continuous model, the distribution of T has a cumulative form of

$$\Pr(T < m) = 1 - e^{-\lambda m}. \tag{4}$$

Next, by combining Eq. (4) and Eq. (2), we have

$$\Pr(K < l) = 1 - e^{-\frac{\lambda}{\alpha}\ln l} = 1 - l^{-\frac{\lambda}{\alpha}}. \tag{5}$$

Since $\Pr(K = l) = \frac{\partial \Pr(K<l)}{\partial l}$, by taking the differentiation we obtain that

$$\Pr(K = l) \sim l^{-\frac{\lambda}{\alpha}-1}, \tag{6}$$

which is a power law distribution for group size. We will later show that this property is consistent with empirical observations.

For now, noting that to arrive at the power law scaling of group size distribution, one fundamental assumption is the exponential distribution of invitation duration, which point needs to be verified empirically. To this end, we measure the distribution of invitation duration. For an approximation of the duration, in our dataset, we record the time when the event was first set up by the initiator and the time when the last person responses to the event. We use the time interval in between as the invitation duration. We find that the distribution could be well approximated by an exponential distribution as shown in the semi-log plot in Figure 6. The exponential fit of red line in the figure yields $\lambda = 731.73$. The exponential distribution is repeated for similar but different definitions of invitation duration.

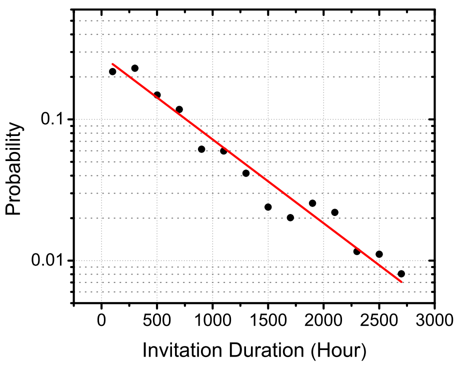


Figure 6: Distribution of invitation duration in semi-log scale. The red straight line in the figure suggests an exponential distribution of the duration.

As an extension of our model, we derive the inverse relationship between group size and the probability of friendship as shown in Figure 4. To arrive at the probability of friendship, we need to know the expected number of total existing links in the K size group, which equals to $\frac{K\langle m\rangle}{2}$, where $\langle m\rangle$ is the expected number of links from a user. According to our model, the number of users in the group that are invited by a specific user s is proportional to the duration that s stays in the propagation process. For simplicity, here we consider the case of a sparse graph, in which the probability of s has a link with users other than the ones invited by him or invite him goes to zero. Thus if we use d_s to denote the duration between the time point that person s joins the group and the time point T that the process ends, then $m_s \sim d_s$. And we have $d_s = T - t_s$, in which t_s is the time point that user s joins the group. For a random selected user in the group, based on Eq. (1), we have

$$\Pr(t_s < l) = \frac{e^{\alpha l}}{e^{\alpha T}}. \tag{7}$$

By taking the derivative on the cumulative distribution,

$$\Pr(t_s = l) = \frac{\alpha e^{\alpha l}}{e^{\alpha T}}. \tag{8}$$

This yields the distribution of arrival time of different users. Since we are interested in the expected number of links from a random selected user in the group, we take the average of d over s

$$\langle d\rangle_s = \int_0^T (T-l)\frac{\alpha e^{\alpha l}}{e^{\alpha T}}dl = \frac{1}{\alpha} - \frac{1}{\alpha}e^{-\alpha T} - Te^{-\alpha T}. \tag{9}$$

Taking the asymptotic form when T is large,

$$\langle d\rangle_s = \int_0^T (T-l)\frac{\alpha e^{\alpha l}}{e^{\alpha T}}dl \sim \frac{1}{\alpha}. \tag{10}$$

Thus the total number of edges $\frac{\langle m\rangle K}{2}$ would be proportional to $\langle d\rangle K \sim \frac{K}{\alpha}$, with a constant multiplier. And since there are $\frac{1}{2}K(K-1)$ potential edges in the group, the probability of two persons being friends in a K size group should scale as

$$\Pr(F) \sim (K-1)^{-1}. \tag{11}$$

The probability of friendship within a formed group is thus inversely related to the size of the group. Despite of the discrepancies in fitted exponents and predictions, the model qualitatively explains the inverse scaling between friendship and group size described in Figure 4.

(a) Whrrl.com

(b) Meetup.com

Figure 7: People attend group-gathering activities for different purposes, e.g., to hang out for recreation, to attend a meeting, or to march for a protest. Word clouds for the purposes of group-gathering activities.

From the above derivations, we see that the model captures key characteristics of the on-line invitation process and also the inverse scaling relationship described in Section 2. One direct implication of the model is the power law distribution of group size, which is suppose to be irrelevant with the purposes of the group, due to the fact that groups with different purposes share the same underlying formation mechanisms. Here, for a verification of our model, we analyze the purposes of group-gatherings and their impact on group size. The word cloud in Figure 7 shows the most frequent types of group-gathering purposes in Whrrl and Meetup datasets. It could be seen that, the most frequently visited places of the two sites are not identically the same.

We measure size distributions of groups in Whrrl and Meetup. Figure 8 (a)-(d) show the group size distribution from Whrrl dataset. Figure 8(a) is the size distribution of all groups in the dataset. A linear fit of the data in log-log scale has R-square value of 0.946. The power law exponents of group size are estimated to be 3.13 from MLE. For the size distribution of groups belonging to different categories, they share similar exponent values. Here, we use tags associated with the group check-in record to identify the categories of

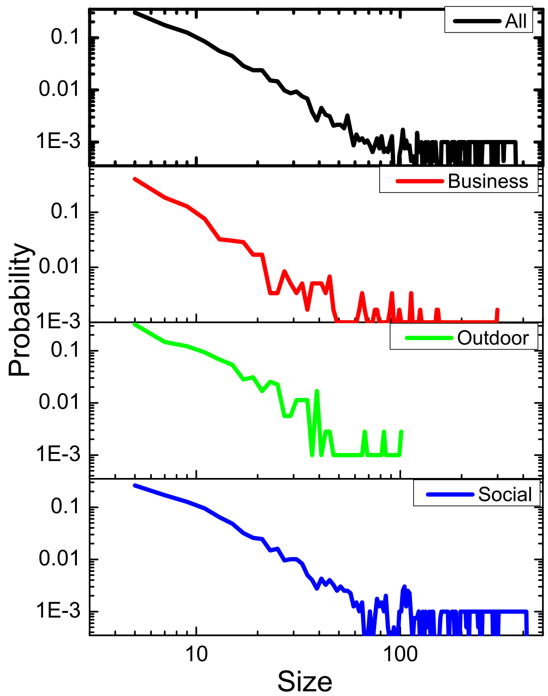


Figure 8: Distribution of group size for different categories of activity types. (a) all groups on Whrrl and (b)-(d) size distribution of certain types of group. The distribution of group size in different types are all governed by the same type of distribution.

the group [21, 22]. As can be observed from Figure 8(b)-(d) the size distribution is universal for all types of activities, even though the estimated exponent value may vary. Similarly, for an analysis on the Meetup dataset, we also obtain the exact same scaling relationship, which is consistent with our model. Due to the limitation of space, we omit figures for the group size distribution of Meetup.

4. APPLICATIONS

Our investigations of f2f group so far have been focused on studying and explaining the phenomena. From these investigations, we demonstrated that group gathering contains rich information that could be applied to solve key problems in social network analysis. In this section, we demonstrate two applications that leverage out the knowledge discovered from group-gathering. The first one is the *social tie inference* problem, where friendship structure is not explicitly declared such as in criminal networks [23] and people need to infer the social structure of the network [24]. The second application is for *friend recommendation*, where people use existing social or non-social information to recommend a list of users for given user to make friends with.

4.1 Group-Gathering Information

Inspired by prior works [24, 23, 18], we first look at the interplay of various similarity features, including group-gathering, spatial and social features on the network scale. From now on, we use GG to denote the number of times that two persons join group-gathering event together. We explore the connections of GG to a series of similarity features including social proximity and geographic proximity measures, by calculating the Pearson correlation coefficients. For network proximity in the social graph, we select three representative quantities, which have been proven to be reasonably good indicators of social closeness. In the following, we let V denotes the set of users and E denotes the set of edges.

- **Common friends (C).** The number of friends that nodes u and v have in common.

$$C(u, v) = |\Gamma(u) \cap \Gamma(v)|, \tag{12}$$

  where $\Gamma(x) = y|y \in V(x, y) \in E$ is the set of friends of x.

- **Weighted common friends (aa) [25].** A weighted common friends based on their degrees

$$aa(u, v) = \sum\nolimits_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(w)|}. \tag{13}$$

- **Jaccard's coefficient (jc).** Defined as the number of the inter-section of the friends of two nodes divided by the size of union

$$jc(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}. \tag{14}$$

To measure the geographic or spatial closeness [23, 18] of two users, we use the following indicators.

- **Distance.** We use the great circle distance of any two users' home coordinates as a measure of geographic distance between them.
- **Number of times of Co-location.** In our observation period, the number of times that two users u and v visit the same geographic grid ($0.1°$) at the same time frame (one day). Note that two users co-located may to attend their separate activities. For instance, two users may both work at two companies close to one another, i.e., they are co-located, but they may never meet each other or join group activities together.
- **Spatial cosine similarity (Scos).** The cosine similarity of two users' trajectories, capturing how similar their visiting frequencies to places are, i.e., the angle between the two vectors of number of visits at each location for user u and v.

The two categories of closeness indicators listed above together with GG are all measurements of similarity. The correlation of each pair of the indicators is as shown in Table 1. In the table, social closeness features have a strong correlation with each other, and mobility closeness features are also well correlated with each other. More importantly, the GG indicator has a strong correlation with both of the two types of indicators. GG has a stronger correlation with social closeness indicators than geographic similarity measurements do, and vice versa. This stronger correlation suggests that GG indicator is a combination of information from both geographic, movement, and social closeness. It would thus be of great interests to apply GG into similarity based predictions and recommendations.

	GG	Co-location	Distance	Scos	Common friend	aa	jc
GG	1	0.7275	-0.08072	0.2411	0.7118	0.7403	0.6904
Co-location	0.7275	1	-0.09209	0.1912	0.6399	0.6318	0.5966
Distance	-0.08072	-0.09209	1	-0.09765	-0.1302	-0.1176	-0.138
Scos	0.2411	0.1912	-0.09765	1	0.2767	0.2625	0.3117
Common friend	0.7118	0.6399	-0.1302	0.2767	1	0.9862	0.9326
aa	0.7403	0.6318	-0.1176	0.2625	0.9862	1	0.9164
jc	0.6904	0.5966	-0.138	0.3117	0.9326	0.9164	1

Table 1: Pearson Correlation of different indicators of mobile closeness(Co-location and distance) and social closeness(Scos, common friend, aa and jc). group-gathering(GG) has a strong correlation with both of the two types of indicators.

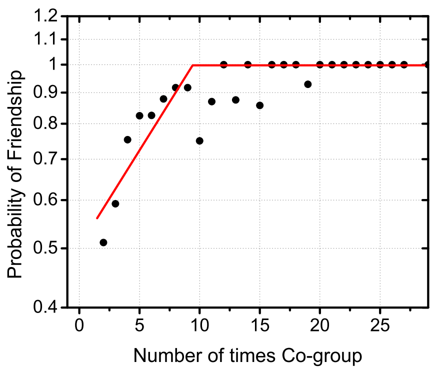


Figure 9: Probability of friendship as a function of the number of times group-gathering together, with probability in log scale.

4.2 Social Tie Inference

In this section, we employ the group-gathering information to infer social ties. Given that two persons have appeared on the same group-gathering event, how likely are they to be friends of each other? Social link declared explicitly on Whrrl is used as proximity of the ground truth of friendship. Figure 9 demonstrates empirical observations from the Whrrl dataset. For each pair of two users, we count the number of times they attended group-gathering events together in a period of one year. From Figure 9, we find that the probability of friendship grows exponentially fast with the number of times that the two users join group-gathering.

To interpret the above observations and develop a framework for social tie inference from group-gathering, we provide a probabilistic model based on the frameworks of prior studies [26, 23, 18]. From the model derived in last section, each pair of users has a probability $\Pr(F) = \frac{M}{N-1}$ to be friends. Let us use C_l to denote the fact that two persons have a total number of l times attending the same group-gathering, i.e. group check-in together in this case, during a period of L days. We assume that each pair of friends has a probability of p_1 to attend a group event, and each pair of non-friends has a probability of p_2 to attend a group event on any given day. Based on our invitation model, one significant conclusion is that p_2 is non-zero, since there's always a chance that a pair of persons not knowing each other are both invited to the event by others. For simplification, we also assume that the network structure does not evolve within the L days of observation period. Whether a pair of user goes out together each day is independent event. So the number of times they co-group follows a binomial distribution.

$$\Pr(C_l|F) = \frac{L!}{l!(L-l)!} p_1^l (1-p_1)^{L-l}. \tag{15}$$

Similarly, for a pair of non-friends,

$$\Pr(C_l|\overline{F}) = \frac{L!}{l!(L-l)!} p_2^l (1-p_2)^{L-l}. \tag{16}$$

From Bayes' Law, we have

$$\Pr(F|C_l) = \frac{\Pr(C_l|F)\Pr(F)}{\Pr(C_l)}. \tag{17}$$

We could thus derive the probability of $\Pr(F|C_l)$ if we know $\Pr(C_l)$. Notice that

$$\Pr(C_l) = \Pr(C_l|F)\Pr(F) + \Pr(C_l|\overline{F})\Pr(\overline{F}), \tag{18}$$

take this back to $\Pr(F|C_l)$,

$$\begin{aligned}\Pr(F|C_l) = &\\ &\frac{\frac{M}{N-1}\frac{L!}{l!(L-l)!}p_1^l(1-p_1)^{L-l}}{\frac{M}{N-1}\frac{L!}{l!(L-l)!}p_1^l(1-p_1)^{L-l} + \frac{N-1-M}{N-1}\frac{L!}{l!(L-l)!}p_2^l(1-p_2)^{L-l}}\\ = &\frac{Mp_1^l(1-p_1)^{L-l}}{Mp_1^l(1-p_1)^{L-l} + (N-1-M)p_2^l(1-p_2)^{L-l}}.\end{aligned} \tag{19}$$

Given that $M \ll N$, the above equation could be approximated as

$$\begin{aligned}\Pr(F|C_l) &\approx \frac{M}{N}\left(\frac{1-p_1}{1-p_2}\right)^L\left[\frac{p_1(1-p_2)}{p_2(1-p_1)}\right]^l\\ &= \frac{M}{N}\left(\frac{1-p_1}{1-p_2}\right)^L e^{l\ln[\frac{p_1(1-p_2)}{p_2(1-p_1)}]}.\end{aligned} \tag{20}$$

This result explains the fact that the probability for two persons to be friends increases exponentially with the number of times joining group activities together observed in Figure 9, and provides us a framework to infer social ties based on group-gathering.

Next, we compare social tie inference from group-gathering with two baseline models from prior work that performs reasonably well. The first baseline model predicts friendship based on geographical distance between two persons [26]. The second one infers social ties based on geographic coincidences [23]. For each pair of users in the network, we use their center of movement to measure the distance between

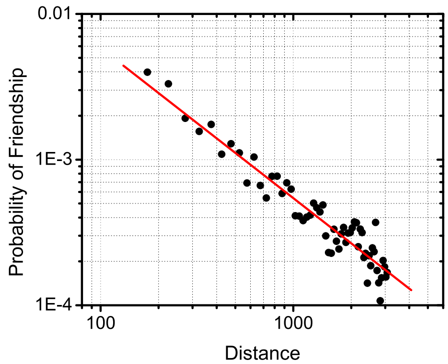


Figure 10: Probability of friendship as a function of the distance between two users, in log-log scale.

them and to obtain the probability of friendship by averaging all pairs of users within a given distance range. We use the number of times that two users both located within the same geographical grid (s) [26] on the same day in the one year observation period as the number of times of co-location. The probability of friendship as a function of the above two indicators is shown in Figure 10 and Figure 11. In Figure 10, the red line shows the distribution of power law with an exponent of -1.03. For the first method based on geographical distance, the best achievable probability is smaller than 0.01 for any distance over 100 km. In Figure 11, the number of times of co-location has to be greater than or equal to 100, in order to reach the saturation region of high probability in friendship. Compared to the baseline models, group-gathering method has advantages in its potential high accuracy prediction using very limited amount of information.

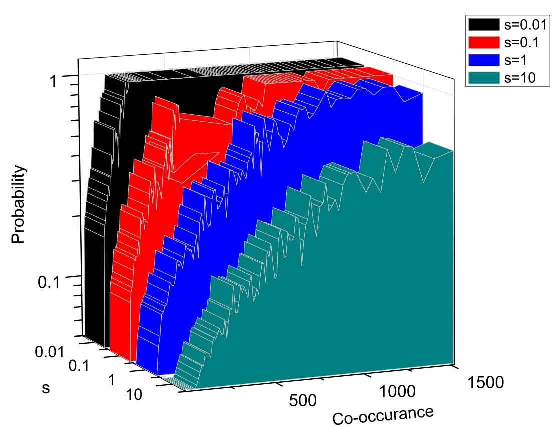
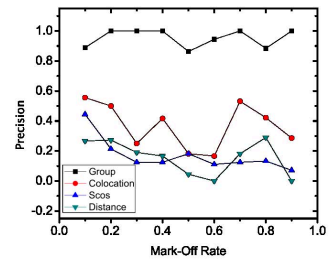
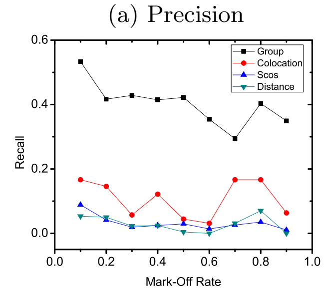


Figure 11: Probability of friendship as a function of co-location under different cell sizes, s, in a semi-log scale.

To test the performance of our framework based on group-gathering for social tie inference, we use data from Whrrl to train an SVM for prediction of the social tie. We compare the prediction accuracy of group-gathering with three baseline features, including distance, number of times of co-location and spatial cosine similarity (Scos) [26, 23, 18]. In our experiments we select a strongly connected component of over $2,000$ users as a subset, hide different percentage (mark-off rate) of pairs for training, and use the rest for testing. Figure 12 shows the prediction accuracy as a function of mark-off rate. Interestingly, we can observe that the group-gathering feature outperforms other baseline features both in precision and recall. The experiments suggest that group-gathering information could improve the predictions accuracy in social tie inference problem.



(a) Precision

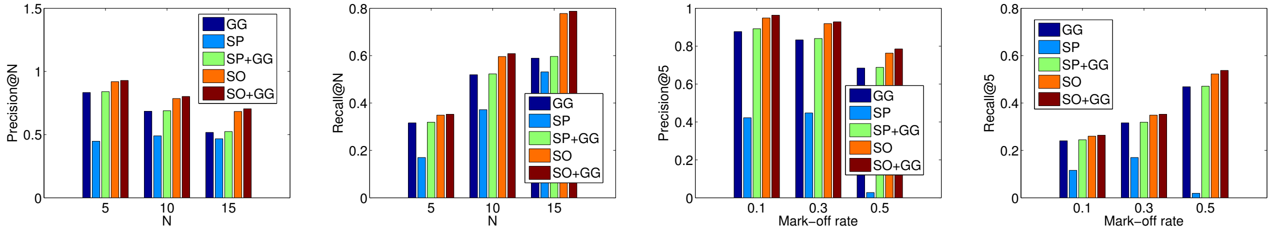


(b) Recall

Figure 12: Social tie inference accuracy as a function of mark-off rate. In the graph, the *GG* feature outperforms other base-line features in terms of both precision and recall.

4.3 Friend Recommendation

Next, we exploit the group-gathering information in the friend recommendation problem. Besides the simple strategy of suggesting friends-of-friends [27], several type of information, such as co-tagging and co-location records, that are related to pairs of users, have been explored to facilitate friend recommendation in [18, 27, 28]. To study the effectiveness of adding group-gathering information in friend recommendation, we perform a series of experiments in comparing the precision and recall of prediction results. Here, we use SP to denote a combination of spatial features including distance, co-location and $Scos$. We use SO to denote a combination of social features including common friend count, aa and jc in Section 3. In our experiments, we hide the same percentage of pairs for each person for testing and use the rest as training. By ordering the probability of these hidden pairs predicted by trained model from SVM, we conclude a recommended friendship if the pair's probability is among the N highest of all hidden pairs. We compare the precision and recall of the recommended friendship using different features. As shown in Figure 13, the GG feature alone outperforms the combination of SP features. More importantly,



(a) Precision@N (b) Recall@N (c) Precision@5 (d) Recall@5

Figure 13: (a) Precision as a function of N. (b) Recall as a function of N. The figures suggest a performance improvement by adding GG in friend recommendation. (c)-(d)Precision and recall as a function of mark-off rate. By adding GG in friend recommendation, the performance improves.

by combining GG with SO and SP, the prediction results are improved. This finding suggests a promising application of GG for friend recommendation.

5. RELATED WORK

In this section, we discuss some of the relevant studies categorized in three sub-areas.

5.1 On-line Group and LBSN

Research related to on-line groups and location based social networks are the two fields most related to our work. Both of the two attract great recent research interest for their novelty impacts and potential applications. As discussed in the introduction section however, most existing works on on-line groups focus on studying "cyber groups"[5, 6, 7, 8, 9]), where group members interact with each other in cyberspace and thus do not constrained geographically. This fact differentiates the special kind of group in our study from others. Here in our study, social networking websites are serving as subsidiaries in helping organize or form the group, rather than a venue or platform for communications. And also, from the perspective of research related to LBSN [10, 11, 12, 13, 14, 15], fruitful results have been obtained by applying individual check-in record in various prediction tasks such as place-of-interest recommendation and link prediction. Yet, few documented research has linked to the check-in record of a group. Thus exploring group behavior under the context of LBSN should be of interests for researchers from both of the two fields.

5.2 Social Tie Inference

Knowing patterns of relationship in a social network is very useful with various applications such as to investigate collaborations among criminals and to exploit social relationships for e-commerce promotion [29]. However, in various organizations, neighborhoods, on-line communities and social service websites, friendship structure is not explicitly declared. Existing works on social tie inference primarily use criteria such as self-report [30], communication [31, 24, 32], similarities of opinion [25] and spatial-temporal co-occurrence [29, 26, 23]. Particularly for the last criteria, recent works bring the interplay between human mobility and friendship into consideration [14, 18, 33].

5.3 Friend Recommendation

Recommending people in social networks is an important task for many recommender systems. Many prior work proposed to recommend potential friends to users or predict future links in a social network [18, 22, 27, 28, 34, 13]. Besides of the simple strategy of suggesting friends-of-friends [27], content-based approaches were proposed to match the content of user profiles and determine user similarities for recommendation in [27, 28, 34]. Furthermore, multiple types of information, such as co-tagging, co-commenting records and co-location records, that are related to pairs of users, have been explored to facilitate friend recommendation in [18, 27, 28]. In this paper, we aim to exploit group-gathering information to improve the friend recommendation performance. To our best knowledge, this type of information has not been explored before; while we consider this type information have strong indication about potential friendship, i.e., people join the same group-gathering activities usually have overlapped social circles and spatial closeness. It would thus be extremely interesting to study the effectiveness of group-gathering information applied into friend recommendation problem.

6. CONCLUSION AND FUTURE WORK

In this paper, we study the phenomena of (off-line) face-to-face group that are organized through on-line social portals, by using datasets collected from Whrrl and Meetup. First, based on the analytical study, we find that group-gathering activities exhibit striking differences compared with individual activities both in temporal and spatial aspects. We also analyze the social structure properties with-in such a group. Next, to understand the formation process of group-gathering, we propose a dynamic model based on the concept of on-line friend invitation process. The proposed model well reproduces the empirical observations. Finally, we discuss how to apply group-gathering information into problems of social tie inference and friend recommendation. Right now, our model incorporates only the social aspect of group gathering characteristics, potential refinements of the model may take into consideration the spatial and temporal constraints, and also the impact of microscopic differences between different users. In closing, we point out that although the focus in this paper has been on face-to-face groups that are formed through social networking websites, the framework of our model may be applicable to other types of groups and communities. The issue here raised in the paper - off-line user behavior facilitated or changed by on-line networking - is fundamentally important and therefore will provide ample opportunities for further future work.

7. ACKNOWLEDGMENTS

C. W. would like to thank HP Labs for financial support. C. W. and M. Y. would like to thank B. A. Huberman for fruitful discussions in this work.

8. REFERENCES

[1] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

[2] Roger Guimera, Brian Uzzi, Jarrett Spiro, and Luis A. Nunes Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005.

[3] G. Palla, A. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446:664–667, 2007.

[4] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD*, pages 44–54. ACM Press, 2006.

[5] N.F Johnson, Chen Xu, Zhenyuan Zhao, Nicolas Ducheneaut, Nicholas Yee, George Tita, and P.M. Hui. Human group formation in on-line guilds and off-line gangs driven by a common team dynamic. *Phys. Rev. E*, 79:066117, 2009.

[6] Lei Tang, Huan Liu, Jianping Zhang, and Zohreh Nazeri. Community evolution in dynamic multi-mode networks. In *KDD*, pages 677–685, 2008.

[7] Munmun De Choudhury. Modeling and predicting group activity over time in on-line social media. In *Hypertext*, pages 349–350, 2009.

[8] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. Influence and passivity in social media. In *WWW*, pages 113–114, 2011.

[9] Xiaolin Shi, Jun Zhu, Rui Cai, and Lei Zhang. User grouping behavior in on-line forums. In *KDD*, pages 777–786, 2009.

[10] Nan Li and Guanling Chen. Analysis of a location-based social network. In *CSE*, pages 263–270, 2009.

[11] Vincent Wenchen Zheng, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative location and activity recommendations with gps history data. In *WWW*, pages 1029–1038, 2010.

[12] Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. On the semantic annotation of places in location-based social networks. In *KDD*, pages 520–528, 2011.

[13] Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. Exploiting place features in link prediction on location-based social networks. In *KDD*, pages 1046–1054, 2011.

[14] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD*, pages 1082–1090, 2011.

[15] Huiji Gao, Jiliang Tang, and Huan Liu. Exploring social-historical ties on location-based social networks. In *ICWSM*, 2012.

[16] M.C. Gonzales, C.A. Hidalgo, and A.L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[17] Bin Jiang, Junjun Yin, and Sijian Zhao. Characterizing the human mobility pattern in a large street network. *Phys. Rev. E*, 80:021136, Aug 2009.

[18] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and A.L. Barabasi. Human mobility, social ties, and link prediction. In *KDD*, pages 1100–1108, 2011.

[19] E. Bonabeau, L. Dagorn, and P. Freon. Scaling in animal group-size distributions. *PNAS*, 96:4472–4477, 1999.

[20] D.J.T. Sumpter. Collective animal behavior. In *Princeton University Press*, 2010.

[21] Denis Helic, Markus Strohmaier, Christoph Trattner, Markus Muhr, and Kristina Lerman. Pragmatic evaluation of folksonomies. In *WWW*, pages 417–426, 2011.

[22] Rossano Schifanella, Alain Barrat, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Folks in folksonomies: social link prediction from shared metadata. In *WSDM*, pages 271–280, 2010.

[23] David J. Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 107(52):22436–22441, 2010.

[24] Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *PNAS*, 106(36):15274–15278, 2009.

[25] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2001.

[26] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, 2005.

[27] Jilin Chen, Werner Geyer, Casey Dugan, Michael J. Muller, and Ido Guy. Make new friends, but keep the old: recommending people on social networking sites. In *CHI*, pages 201–210, 2009.

[28] Ido Guy, Inbal Ronen, and Eric Wilcox. Do you know?: recommending people to invite into your social network. In *IUI*, pages 77–86, 2009.

[29] Hady Lauw, Ee-Peng Lim, HweeHwa Pang, and Teck-Tim Tan. Social network discovery by mining spatio-temporal events. *Computational and Mathematical Organization Theory*, 11:97–118, 2005.

[30] S. Wasserman and K. Faust. Social network analysis: Methods and applications. In *Cambridge University*, 1994.

[31] Munmun De Choudhury, Winter A. Mason, Jake M. Hofman, and Duncan J. Watts. Inferring relevant social networks from interpersonal communication. In *WWW*, pages 301–310, 2010.

[32] Honglei Zhuang Wenbin Tang and Jie Tang. Learning to infer social ties in large networks. In *ECML/PKDD*, pages 381–397, 2011.

[33] G. Mangesh and E. R. Tina. Measuring tie strength in implicit social networks. *Arxiv*, 2011.

[34] Daniele Quercia and Licia Capra. Friendsensing: recommending friends using mobile phones. In *RecSys*, pages 273–276, 2009.