Knowledge-Based Systems 143 (2018) 19–29

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Hybrid EGU-based group event participation prediction in event-based social networks

Shuo Zhang*, Qin Lv

University of Colorado Boulder, Boulder, CO, 80309, USA

ARTICLE INFO

Article history:
Received 12 January 2017
Revised 29 November 2017
Accepted 1 December 2017
Available online 8 December 2017

Keywords:
Event-based social networks;
Event attendance prediction
Behavior analysis

ABSTRACT

The increased popularity of event-based social networks (EBSNs) connects online social communities with offline event activities, and makes it interesting and necessary to understand users' event participation behaviors on this new type of online social platform, especially when groups are explicitly defined as event organizers and have declared memberships from individual users. Accurate event participation prediction can help guide more effective event participation, event organization, and community development. In this work, using data collected from the popular Meetup.com EBSN spanning three major cities with diverse cultures, we first conduct detailed analysis on group- and user-specified event behaviors. Based on this analysis, we propose a group-based event participation prediction framework that uses personalized random walk with restart on a hybrid EGU (event-group category-user) network to capture intrinsic social relationships, and fuses that with newly designed content and contextual features to predict event participation by group members. Detailed evaluations using the Meetup datasets demonstrate that our prediction frameworks achieves high prediction performance with the proposed features.

1. Introduction

Event participation is collaborative work in nature, and understanding people's event participation behaviors can offer valuable new insights for effective event participation, event organization, and community development. The recent years have seen increased popularity of event-based social networks (EBSNs), such as Meetup [1], Facebook Events [2], and Douban Events [3]. These social websites allow individuals or groups to easily create, manage, or participate in diverse types of events. As such, a rich set of event-related information becomes available, connecting people's online and offline activities. For instance, Meetup has over 440,000 monthly meetups (events) and almost 3 million monthly RSVPs [1]. Fig. 1 illustrates the key elements in Meetup: users, meetup groups, and events (meetups), as well as the types of information associated with these elements. Users can become members of different groups and RSVP for events hosted by different groups. New event activities of a group will be pushed to the group members' main pages; and group members can invite any registered user to attend the event.[1] The burst of social event media enriches user experience and people's awareness when interacting with the event [4]. Effective event participation not only enriches users' experience, productivity, and quality of life, but also allows event organizers to attract more people and increase the impact of their events. Accurate prediction of user's event participation can be beneficial in terms of planning, advertising and crowd management, and for individual users in terms of recommendation and coordination among different groups of users and events. The challenge lies in the fact that a large number of diverse events occur at different locations and time, while users with diverse interests and location/time constraints or preferences cannot easily identify the right set of events to attend. To support effective event search, recommendation, and organization, it is fundamentally important to understand and predict users' event participation based on the most influential factors.

While there has been a lot of research on location based social networks (LBSNs) [5–7], only a few studies exist for EBSNs on event participation prediction and event recommendation [8–11]. Different methods have been proposed in terms of content, contextual, and social feature extraction and multi-feature fusion. Most of these works focus on the interactions between individual users and events. The notions of group and group membership are either non-existent or directly embedded in the user-event modeling process without an explicit group-based analysis or design.

In this paper, to study group- and user-specific event behaviors, we have collected 12-month Meetup data in three major cities

* Corresponding author.
E-mail addresses: jasonzhang@colorado.edu (S. Zhang), qin.lv@colorado.edu (Q. Lv).

[1] A user does not need to be a group member to RSVP for that group's events. But this is rare in our Meetup datasets.

https://doi.org/10.1016/j.knosys.2017.12.002
0950-7051/© 2017 Elsevier B.V. All rights reserved.

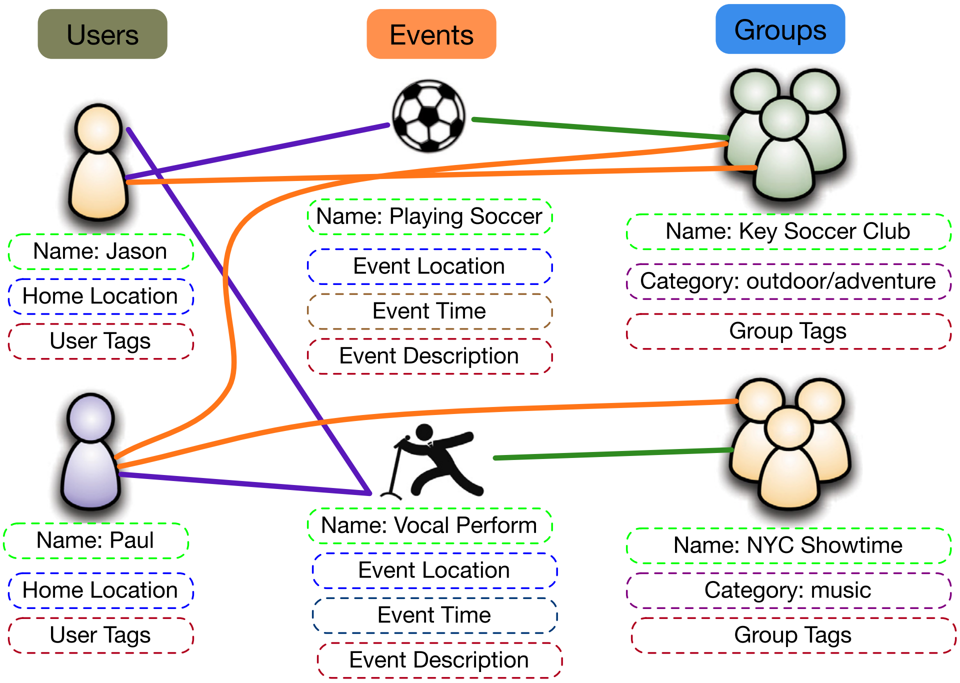

Fig. 1. An illustration of the key elements in Meetup.com.

across three continents: New York, London, and Sydney. Using these datasets, we analyze and validate that (a) groups have focused themes, (b) users have diverse interests, and (c) social factors play an important role in offline events attendance. Based on this analysis, we construct a hybrid network connecting users, group categories, and offline events, and utilize a personalized random walk with restart method to model group-based social relationship between offline events and users. This social feature is further integrated with our newly-designed content, spatial, and temporal features, resulting in a unified framework for group-based event participation prediction in EBSNs.

Our paper makes the following contributions:

(1) We have conducted detailed analysis using Meetup EBSN data covering three diverse cities to identify the key group- and user-specific features for event participation. Our detailed observations and analysis help us to discover patterns that play important roles in user's event attendance decision making and can facilitate further studies on event-based user and group behaviors, community development, etc.

(2) We have designed a hybrid EGU (event–group category–user) network and proposed the use of personalized random walk with restart [12], an extension of the popular link prediction algorithm RWR, to model the group-based social closeness between offline events and users.

(3) We have proposed a group-based event participation prediction framework. The framework effectively embeds and connects group context features and social related features that we discovered with historical event attendance logs.

(4) We have implemented state-of-the-art context-aware recommendation systems to evaluate the proposed prediction framework using real-world Meetup data spanning three major cities. The results show our hybrid prediction framework provides high-quality predictions for all three cities, and for users/groups with different active levels.

The rest of this paper is organized as follows. First we discuss related work. Then we present the problem formulation, the datasets we collected, and detailed data analysis to capture latent factors in people's event participation behaviors. Next we present the design of the EGU-based social graph to capture latent social features. Detailed evaluation results demonstrate high prediction performance of our method and contributions of the features we propose. Finally, we conclude this paper with discussion of some future work.

2. Related work

Location-based social networks (LBSNs) have been a topic of popular research in recent years, focusing mainly on activity recognition and prediction, as well as point-of-interest detection and recommendation [5–7,13,14]. Please see [15] for a survey on recommendations in LBSNs. Although related to EBSNs, LBSNs usually focus on static venues that do not change, while the cold-start problem is much more severe for EBSNs due to the highly dynamic nature of events. The work by Georgiev et al. aimed to recommend large events based on Foursquare check-ins and Twitter following/follower graph [13]. Recent work by Guha et al. concentrated on location surveillance concerns with check-in type location sharing features [5]. This represents some early work in LBSNs to study events, but the settings generally differ from that of EBSNs due to the lack of explicit notions of groups, events, and user RSVPs.

Compared with the extensive research on LBSNs, there is limited research on EBSNs. The works by Du et al. and Yu et al. proposed solutions for predicting event participation and inviting more influential followers [8,11]. Their works were based on the Douban Events network, which offers good content, contextual, and host-participant social features, but there is no explicit notion of group or group membership. Macedo et al. proposed a context-aware solution for event recommendation in EBSNs [10]. They used the Meetup network and modeled the social factor by the frequency of users attended events held by a specific group and user group preferences. Compared with previous works on EBSNs, we analyze group-specific event behaviors, propose the construction of a hybrid EGU network and use personalized RWR to effectively capture the social relation, and fuse the social feature with newly-designed content and contextual features for accurate event participation prediction.

Table 1
Notations for the Meetup data in a given city.

Symbol	Meaning
U	the set of users
G	the set of Meetup groups
E	the set of offline events
C	the set of group categories
$C(g)$	the category of group $g \in G$
$E(u)$	the set of events attended by user u
$U_{G(e)}$	the set of users who are members of the group hosting event e
N_u^c	number of events attended by user u that are hosted by groups in category $c \in C$

Our work is also generally related to group behavior analysis. For example, Massimi et al. studied online health and support groups to understand the basis implications when designing technologies to support people who need physical or mental help [16,17]. Hsieh et al. explored a series of predictors of volunteer socializers in Reddit, an online social news-sharing community [18]. Zou et al. analyzed the difference of community voting behaviors in Doodle polls [19]. Zhang et al. studied group event scheduling via a newly designed OutWithFriendz mobile application [20]. This line of research is orthogonal to our work of event participation predication, and insights gained from event-based group behaviors could be leveraged by our framework for further improvement.

Context information such as social trust, location, and time has been studied frequently in recent years for improving the performance of recommender systems. For example, SoRec [21] and SocialMF [22] built a social relation graph to model social trust, which has been shown to further improve the performance of traditional collaborative filtering methods. Liu et al. used random decision trees to subgroup user-item with all kinds of context features [23]. To evaluate the performance of our latent factors, we implemented several of those methods, which will be discussed in the Evaluations section.

3. Meetup EBSN data analysis

In this section, we first present the problem formulation and introduce the datasets we have collected from Meetup. We then present a detailed analysis of group- and user-specific behaviors in event participation. The insights we gain in this analysis are used to guide the design of our event participation prediction framework.

3.1. Problem formulation

As illustrated in Fig. 1, we consider three types of first-class elements in the Meetup EBSN: users, online groups, and offline events. Each group is associated with a single group category and multiple tags. Each event is hosted by a specific group and described by its time, location, and text description. Each user has a home location and a set of tags. Each user can be members of multiple groups and RSVP for different events. Table 1 summarizes the key notations we use to represent the Meetup data for each given city. Given an event $e \in E$, for each user $u \in U_{G(e)}$ (i.e., u is a member of the group hosting e), our goal is to predict if user u would attend event e or not. Please note that we focus on making predictions for group members instead of randomly-selected users since events are rarely attended by non-group members.

3.2. Data collection

Established in 2002, Meetup.com has grown to be a very popular EBSN with active users and meetup events across many countries [1]. It aims to revitalize local communities and help people around the world to self-organize via Meetup groups and events. Using the Meetup REST API,[2] we have collected year-long (from July 2014 to June 2015) datasets covering groups, users, and events in three major cities: New York, London, and Sydney. These cities are located in three different countries with three different continents with various geographical and cultural differences. To ensure sufficient data coverage, we only consider groups that have organized at least seven events and events that had at least three participants. Table 2 summarizes the data statistics of the three cities.

Table 2
Dataset statistics.

City	#groups	#users	#events	#rsvps
New York	3478	180,941	101,653	894,353
London	2109	126,476	81,640	638,317
Sydney	705	55,767	50,294	353,148

Table 3
Meetup group categories.

Alternative lifestyle	Health/Wellbeing	Parents/Family
Career/Business	Hobbies/Crafts	Pets/Animals
Cars/Motorcycles	Language/Ethnic identity	Photography
Environment	LGBT	Religion/Beliefs
Dancing	Literature/Writing	Sci-fi/Fantasy
Education/Learning	Movements/Politics	Singles
Fashion/Beauty	Movies/Film	Socializing
Fine arts/Culture	Music	Sports/Recreation
Fitness	New age/Spirituality	Support
Food/Drink	Outdoors/Adventure	Tech
Games	Paranormal	Women

3.3. Group behavior patterns

To discover the underlying forces that attract users to specific groups and the events they organize, we first focus our analysis on group behavior patterns and make three key observations with the Meetup datasets we have collected.

Observation 1. Most groups focus on a single, specific theme.

Each Meetup group belongs to a specific group category (see list of group categories in Table 3. The group categories are predefined by Meetup and they are exhaustive (33 total). Each group can be associated with multiple descriptive tags. To understand the focuses of different groups, we construct a tag-to-tag graph where each node is a tag and there is a link between two tags if they are shared by the same group. We then apply the Newman–Girvan algorithm [24] on the graph to cluster the group tags. The Newman–Girvan algorithm is a global measure for community detection. Given a random network, some nodes can form a cluster if the number of edges between them is larger than expected. The

[2] The Meetup API provides both RESTful HTTP and streaming interfaces which returns nearly all the data in JSON format from Meetup.com after necessary privacy processing.

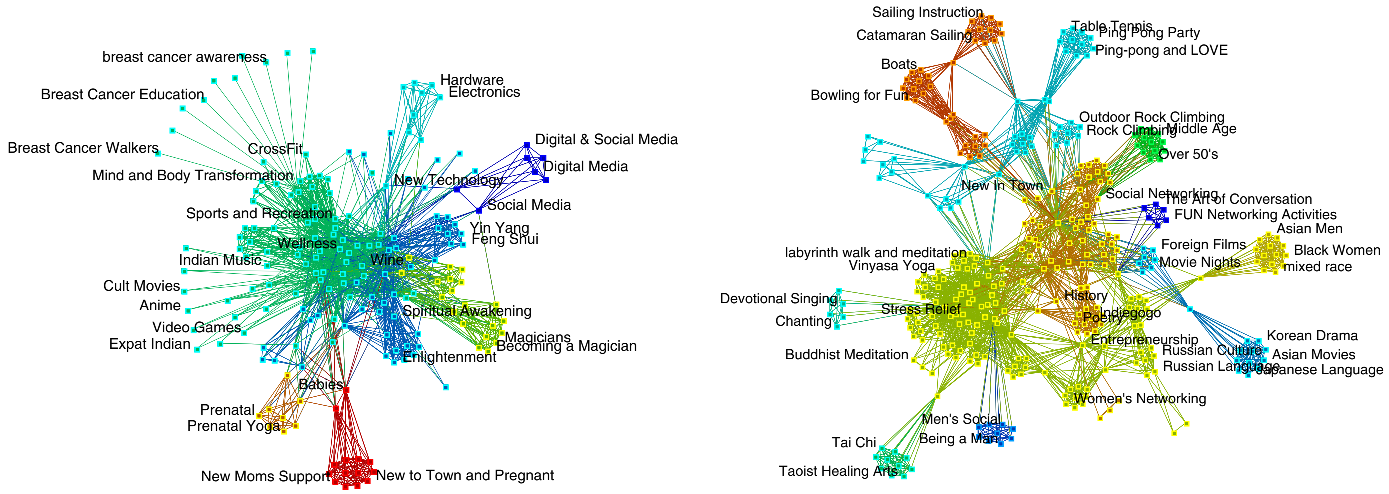

Fig. 2. Tag-to-tag network for New York (Left) and London (Right).

Table 4
Tag-to-Tag network modularity.

City	New York	London	Sydney
Modularity	0.63	0.71	0.55

Table 5
Group co-membership network modularity.

City	New York	London	Sydney
Modularity	0.13	0.13	0.08

algorithm works iteratively by computing the betweenness scores for the remaining network and removing the edge with the highest betweenness score. Based on the communities detected, Newman and Girvan also proposed an effective quality function for computing network modularity [25], which can be used to measure the strength of network division. The modularity can be computed as follows:

$$Q(C) = \sum_{i,j \in N} [P_{i,j} - \frac{M_{i,j}}{2m}]\delta(c_i, c_j) \tag{1}$$

where N is the set of nodes, m is the sum of the weight matrix M, $P_{i,j} = k_i^{out} k_j^{in} / m^2$ denotes the expected number of connections between nodes i and j, and k_i^{out} and k_j^{in} are the out degree of node i and in degree of node j, respectively. Function $\delta(c_i, c_j) = 1$ if node i and node j are in the same community, and 0 otherwise.

Fig. 2 shows the communities detected by the Newman–Girvan algorithm, where tags with the same color belong to the same community. Due to space limit, the figure only shows the results for New York and London, while the Sydney dataset shows similar characteristics. As can be seen in the figure, the group tags form several large connected components (clusters) that are distinct from each other. This is also evident given the high modularity values shown in Table 4. For instance, for the New York network shown in Fig. 2, in the upper area, the “Breast Cancer” cluster is only connected to more general concepts such as “Wellness”. This may be explained by the fact that groups which care about breast cancer would also be interested in healthy lifestyle. Similarly, in the bottom area, the cluster “Magicians” only connects to more general concepts like “Spirituality”. Based on these results, we can conclude that most groups in Meetup focus on a single, specific theme.

Observation 2. Users’ interests are motivated by a complex set of factors.

Each Meetup user can join multiple groups. We can then construct the group co-membership network, where nodes represent groups and two groups are connected if they have at least 15 members in common. The reason to pick 15 as a threshold is that we experimented with different threshold values: 10, 15, 20, 25. Using 10 and 15 results in similar group networks while increasing the threshold from 15 to 20 had a significant reduction of group-to-group edges (from 159,911 to 88,802) and thus limited group network information. Therefore we choose 15 as the threshold. Similarly to the tag-to-tag network, we apply the Newman–Girvan algorithm to color the group nodes with different clusters and compute the corresponding modularity. The results are shown in Fig. 3 and Table 5. Compared with the tag-to-tag networks in Fig. 2, although we still see large connected components representing different clusters, the clusters are much more spread out and entangled with each other. And the modularity values low for the group co-membership networks, indicating that people’s interests cannot be captured by a single topic. Instead, users’ personal interests are motivated by a much more complex set of factors. For instance, for a business woman, her career related to business does not define her. She may also be interested in breast cancer prevention and detection, and parenting.

Observation 3. Social factors play an important role in users’ participation of offline events.

Given the specific focus of each group (Observation 1) and diverse interest of each user (Observation 2), we consider representing users’ interests by the combined group categories of each user’s groups. Furthermore, groups are online, but the events they organize are offline. By participating in those offline events, users can have many face-to-face interactions, either seeing friends they already know in the group, or meeting new friends with similar interests. According to a previous study [26], users at Meetup tend to focus on three main topics: making new friends, finding people with similar hobbies, and networking. Therefore, even though no explicit relationships are declared among Meetup users, the co-existence of group memberships and co-attendance of group events are good indicators of the social ties among users. We employ the χ^2 test [27], which is a statistical test widely used to evaluate the likelihood that any observed difference between sets is happend by chance, to measure the significance of social relation in event attendance.

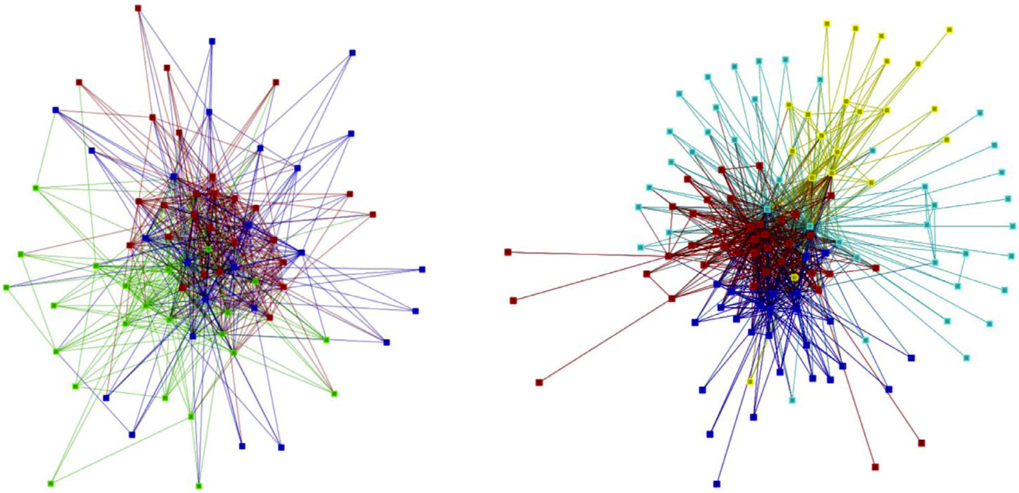

Fig. 3. Group co-membership network for New York (Left) and London (Right).

Table 6
χ^2 test for social relation (Significance Value = 0.05).

City	New York	London	Sydney
Significance Rate	0.70	0.72	0.72

According to [28], users with more common relations have stronger homophily between them. We can then infer that two users are more likely to be friends or share similar interests if: (1) they joined more online Meetup groups in common and/or (2) they attended more offline events in common.

$$Score(u, v) = \alpha \frac{|E(u) \cap E(v)|}{|E(v)|} + (1 - \alpha) \frac{|G(u) \cap G(v)|}{|G(v)|} \tag{2}$$

Eq. (2) defines the closeness between two users u and v, where $E(u)$ and $E(v)$ denote the set of past events that users u and v have attended, $G(u)$ and $G(v)$ denote the set of groups that users u and v have joined, and α is the parameter to balance the weights between common events and common groups. If the normed closeness is high (larger than 0.1), which means that the two users co-attended many events and groups, they would have a higher probability of knowing each other. Then for each user and the events he/she attended in the test set,[3] we compare the attendance rate between his/her friends and non-friends using χ^2 test. The results are shown in Table 6. We can see that for all three cities, social relation is significant for more than 70% of the users. This motivates our design of the hybrid EGU network and the use of personalized random work with restart to model the social influence of event participation, which will be described in the graph design section.

3.4. Spatial factors

Many real-world networks are known to exhibit some heavy-tail properties. Previous studies have shown that users in LBSNs are more likely to have check-ins near their homes and the check-in distance probability follows the power-law distribution [13,29]. Using the Meetup datasets we have collected, we analyze the distance between event locations and attendees' home locations,[4] and discover that the probability of users' event participation given specific home-event distance also follows the power-law distribution. As shown in Fig. 4, for home-event distances that are within 30 miles, which cover more than 85% of the events in all three cities, there is a clear linear relationship in the log-log scale plot between home-event distance and probability of event participation. Based on this analysis, we propose the following model to capture the spatial factors for event participation:

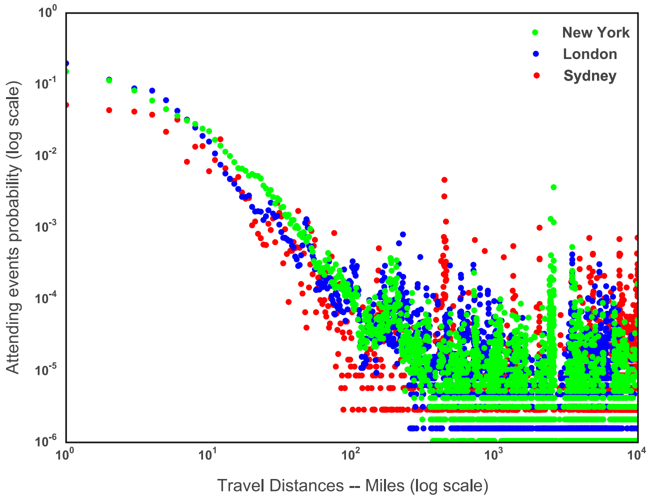

Fig. 4. Distribution of home-event distance vs. probability of event participation.

$$p = k \times d^b \tag{3}$$

where d is the distance, p is the probability of a user attending an event given the user-event distance, k and b are unknown parameters. When considered in the log scale, Eq. (3) can be transformed to Eq. (4), which follows the linear relationship:

$$\log p = w_0 + w_1 \log d \tag{4}$$

where $w_0 = \log k$ and $w_1 = b$. Let $p' = \log p$ and $d' = \log d$, then

$$p' = w_0 + w_1 \times d' \tag{5}$$

The coefficients of the model can be learned via linear regression. Then for a user $u \in U$ and an event $e \in E$, we can define the spatial score as follows:

$$S_l(u, e) = k \times d(u, e)^b \tag{6}$$

where $d(u, e)$ is the distance between u's home and e's location.

[3] Details of splitting the training and test sets will be explained in the Evaluations section.

[4] The user home location obtained through the Meetup API is the (estimated) latitude and longitude when a user registers at Meetup. As such, the home location may not be the true home location and the estimation can be coarse. More precise home location information can be helpful but is beyond the scope of this paper.

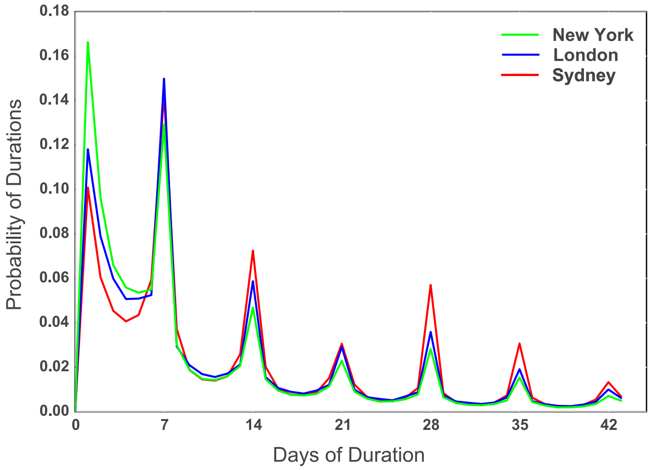

Fig. 5. Distribution of time duration between two consecutive events participated by the same user.

3.5. Temporal factors

Event time also plays an important role in users' decisions to participate in an event or not. Earlier work has discovered that on weekdays, most Meetup events are held in the evening, around 8 pm, and on weekends, the events are distributed more evenly during the day [9]. Besides confirming this in our Meetup datasets, we further analyze the time duration between each user's consecutive event participation. The results are shown in Fig. 5. We can see that there is a peak every 7 days, indicating that users tend to participate in events with a regular weekly pattern such as every week, every other week, every month, etc.

Based on the analysis above, we represent each event time as a 24×7 dimensional vector $\vec{e_t}$. For instance, if an event starts at 19:00 on Monday, then in its event time vector, the $24 + 19 = 43$ th element would be 1 (we assume Sunday is the first day of the week), and all other elements would be 0.

Then the temporal preference of each user u can be represented by the average of his/her past events attendance with the time decay:

$$\vec{u_t} = \sum_{e \in E_u} \frac{1}{(1+\beta)^{\theta(e)}} \vec{e_t} \tag{7}$$

where β is the time decay factor and $\theta(e)$ denotes the number of days past. Integrating time decay here is meaningful because users' temporal preferences may change over time and more recent information can better reflect users' preferences. Then the temporal score for user u and event e can be computed by cosine similarity:

$$S_t(u, e) = Cosine(\vec{u_t}, \vec{e_t}) \tag{8}$$

The main idea is based on: If event e starting time matches user u's temporal preferences, the corresponding $S_t(u, e)$ would be higher and vice versa. This is similar to the method used in [10].

3.6. Content factors

As we have mentioned earlier, each Meetup group and each user are associated with a set of tags, and each event has a text description. Instead of measuring the content similarity between two events (e.g., a new event and a previous event attended by a user), we propose to measure the content similarity between users and events, which better captures the latent semantic relationship between events and user interests. Specifically, each event document consists of the event description, and each user document consists of the user tags and the user's groups' tags. Using Latent Dirichlet Allocation (LDA) [30], each event and each user are then represented as a topic distribution vector that represents the corresponding semantic information. The content score between a candidate event $e \in E$ and target user $u \in U$ can be measured by the cosine similarity of their vectors:

$$S_c(u, e) = Cosine(\vec{u}, \vec{e}) \tag{9}$$

Table 7
User–Category Score: an example.

Meetup user: Marc	
Food/drink	0.357
Tech	0.283
Literature/writing	0.236
Career/business	0.123

4. Design of hybrid EGU graph and social score

In this section, we present the detailed design of our hybrid EGU social graph for event participation prediction which effectively fuses together group-based social factors.

Our analysis has demonstrated that group categories and group/event-enabled user friendships play major roles in offline event attendance. Inspired by this analysis, we aim to quantify the level of attractiveness of events for users by measuring the closeness between them. This is accomplished by constructing a hybrid EGU network consisting of users, group categories, and events, and running personalized random walk with restart in the network to compute the event-specific attraction scores for different users.

Note that we also considered using groups or group tags instead of group categories in the network. The issue is that there exist thousands of Meetup groups and over 18,000 tags, which can be very specific. Using groups or group tags directly would result in much larger graphs and may generate over-fitted prediction models. Furthermore, such highly-specific and highly-sparse graphs add little cross referencing information beyond the standard user-event participation information. In contrast, group categories (33 total) are globally defined at Meetup.com and when used in our social graph model strikes a good balance between social specificity (e.g., groups as communities) and interest commonality (e.g., shared interest within or across group categories).

Next, we explain in detail the proposed scoring functions for four different types of relationships in the hybrid EGU network, the network structure, and the actual random work process.

4.1. User–Category Score

This score captures a user's preference for different group categories, and higher values indicate that the user is more interested in a group category and this user interest is significant for that group category. In other words, a user's interest in a not-so-popular group category carries more weight than his/her interest in a highly popular group category. This notion is similar to TF-IDF. If we treat users as documents and group categories as words, then the importance of a category c for user u can be defined as:

$$Score_u^c = \frac{N_u^c}{\max\{N_u^{c'} : c' \in C\}} \times \ln \frac{|U|}{|\{v \in U : N_v^c > 0\}|} \tag{10}$$

where N_u^c denotes the number of past events attended by user u that belong to group category c. A user's interests can then be modeled by his/her personalized scores for different group categories. One example is shown in Table 7. The Meetup user Marc marked himself with the following tags: "Craft Beer", "JavaScript",

Table 8
Event-category score: an example.

Meetup event: Tibetan singing bowl meditation	
Health/wellbeing	0.424
Community/environment	0.215
Outdoors/adventure	0.183
New age/spiturality	0.082
Food/drink	0.033
Singles	0.022
Education/learning	0.021
Career/business	0.018

"Data Visualization", "Classic Poems", and "Technology Startups", which are adequately captured by his user-category scores for "food/drink", "tech", "literature/writing", and "career/business".

4.2. Event–Category Score

This score captures the importance of a group category for a given event. For a given event e, let $U_{G(e)}$ be the set of users who are members of the group that hosts event e (i.e., event e's group members). If category c is considered more important by event e's group members than other users, then category c is more important for event e. Specifically,

$$Score_e^c = \frac{|\{u \in U_{G(e)} : N_u^c > 0\}|}{|u \in U_{G(e)}|} \times \frac{\sum_{v \in U_{G(e)}} N_v^c}{\sum_{v \in U} N_v^c} \tag{11}$$

One example is shown in Table 8. The Meetup offline event: "Tibetan Singing Bowl Meditation" has different event–category scores for eight group categories. Among them, the group category "health/wellbeing" has the highest score. This is reasonable since this event is organized by the Meetup group "Sacred Karma Yoga", which focuses on physical health and practices. Besides that, its group members also show high interests in topics like "community/environment" and "outdoors/adventure".

4.3. Category–User Score

This score captures the importance of a user for a given group category. The calculation is similar to that of the user–category score, with a switch of order. Now group categories are regarded as documents and users are regarded as words:

$$Score_c^u = \frac{N_u^c}{\max\{N_v^c : v \in U\}} \times \ln \frac{|C|}{|c' \in C : N_u^{c'} > 0|} \tag{12}$$

Intuitively, for each group category c, high weights are given to the users who attended more events hosted by groups in category c compared with other users for the same category and other categories for the same user.

4.4. User–User Score

In Meetup, users do not have following or follower relationships as Twitter does. Specifically, we define the closeness between two users u and v as Eq. (2). Please note that in our equation, u's closeness to v is different from v's closeness to u. This is important since u and v may be interested in different numbers of groups and participate in different numbers of events. Therefore, having one event in common may be significant for one user (who only participated in a few events), and not so important for another user (who participated in many events).

4.5. Structure of the hybrid EGU network

To model the four kinds of social relationships discussed above, we design a hybrid network which connects users, group categories and offline events, shown in Fig. 6. Nodes represents offline events (left), group categories (middle) and users (right). Especially, User–User links, User–Category links, and Category–User links are bidirectional, and can have different weights for each direction. To model the probability of user's attendance, we normalize the weights on each link by dividing the sum of weights on the out links. Table 9 summarizes the statistics of hybrid networks constructed for each of the three cities, where the number of nodes include offline events, group categories and users, and sparsity is defined as the percentage of possible missing links between nodes. The graphs are expected to have high sparsity because: (1) there are no links between events; (2) there are no links between group categories; (3) there are no users if they are not co-member of any group.

Table 9
Statistics of the constructed hybrid networks.

City	#Nodes	#Edges	#Sparsity
New York	282,430	1,340,945	99.94%
London	205,558	1,078,389	99.84%
Sydney	108,245	675,962	99.46%

4.6. EGU-based social score calculation

Random walk has been successfully applied in a wide range of areas such as recommender systems [31,32] and information retrieval [33,34]. It is an effective way to measure the relevance between nodes. The random walk with restart (RWR) approach assumes that a random walker would sometimes jump back and restart at some node. At each step, the walker would either restart at one node (with probability $1-\beta$) or move to a neighboring node (with probability β). One iteration involves taking an estimated PageRank vector $\vec{v}$ [35] and computing the next vector by:

$$v' = \beta M\vec{v} + (1-\beta)\vec{e} \tag{13}$$

where M denotes the transition matrix, $\vec{e}$ is the vector with jump back available nodes set to 1 and the others set to 0.

Personalized RWR is an extension of traditional RWR [12]. In the hybrid EGU network we constructed, the walker will start from an offline event node (left). In each step, the walker would jump to the adjacent nodes or jump back. Eq. (13) can be directly used in PRWR by setting other nodes in $\vec{e}$ to 0. We set parameter β to 0.85 by experience, which has been widely used. Our intuition is that if a user node can be easily reached via edges, the user has a higher probability to attend the offline event. Then the social score between user u and event e can be defined as:

$$S_s(u,e) = PageRank_e(u) \tag{14}$$

where $PageRank_e(u)$ is the Pagerank score of user u after running PRWR starting from event e.

5. Evaluations

In this section, we evaluate the event participation prediction performance of our proposed framework using the Meetup datasets that we have collected from New York, London, and Sydney.[5] We first discuss the evaluation methodology, before presenting the evaluation results in detail.

[5] We also experimented with Meetup data in Tokyo, Japan, which is much smaller and has only 1/10 of the users as New York. Still, our method achieved good performance with 0.83 accuracy.

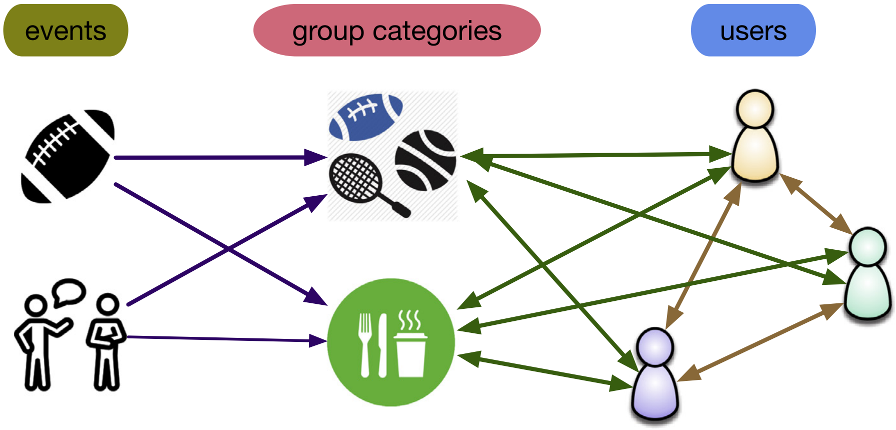

Fig. 6. An illustration of the hybrid EGU network.

5.1. Evaluation methodology

For the Meetup dataset collected for each of the three cities, we split the dataset into three subsets. For every group, we randomly select 80% offline events and attendees pairs that are used for training, 10% of the data are used for validation and parameter tuning, and the remaining 10% data are used for testing. Our social prediction model utilizes the knowledge that some users have already decided to attend an event (i.e., user RSVPs that can be seen by others), and random splitting make it possible to obtain partial user-event participation information for training. Meanwhile, the collected user-event pairs are regarded as positive samples. The number of positive user-event samples are 890,600 for New York, 631,212 for London, and 348,132 for Sydney. To train the classifier without bias, for each event, if the number of attendees is k, then we would randomly select another k users from that group who did not attend the event to form negative user-event pairs.[6] Using the testing set, for each event hosted by a group and each user who is a member of that group, a binary YES or NO decision is made in terms of whether the user would attend that event or not. We then compute four different metrics to evaluate the prediction performance: *accuracy, precision, recall, and F1 measure*. As stated before, we choose to evaluate the prediction performance for group members of an event instead of random users in the system, since users who are not members of a group rarely attend events hosted by that group.

Since our work focuses on the design of new features and their integration, to evaluate the end results of event participation prediction, we implemented several traditional or state-of-the-art recommender methods, each of which has been shown to be effective in its specific settings. We then incorporate our newly-designed features into these methods as appropriate.

- **PMF** [36] is the basic probabilistic Matrix Factorization method, which only considers user-event attendance results.
- **BiasedMF** [37] incorporates user and event biases into MF, which only considers user-event attendance results.
- **SoRec** [21] embeds social trust network into MF, which incorporates both user-event attendance results and simple social relationships.
- **SocialMF** [22] studies trust propagation with MF, which incorporates both user-event attendance results and simple social relationships. In SoRec and SocialMF, the social relation network is built by Eq. (2).
- **GLFM** [38] A latent factor model for group-aware event recommendation by using two kinds of latent factors to model the dual effect of groups: user- oriented perspective and event-oriented perspective.
- **MF-EUN** [39] A hybrid collaborative filtering model, namely, Matrix Factorization with Event-User Neighborhood by incorporating both event-based and user-based neighborhood methods into matrix factorization.
- **SoCo** [23] is the state-of-the-art context-aware recommender method. Random decision trees are applied to provide contextual based user-event subgrouping. All features discussed in the previous sections are incorporated.

5.2. Overall prediction performance

We first compare the overall prediction performance using different context-aware recommender methods. As discussed in previous sections, our proposed framework uses four types of features: content (C), location (L), time (T), and group-based social feature (S). Here we implemented four traditional methods as our baselines: PMF, BiasedMF, SoRec and SocialMF, which consider less contextual information as discussed. We also compare the performance with state-of-the-art context-aware recommender method: GLFM, MF-EUN. All these prediction/recommender algorithms have been widely used in recent years.

Fig. 7 shows the event participation prediction performance of different methods using four different metrics in three cities. As can be seen from the figure, SoCo, by integrating all the features we proposed, performs best across all four metrics in all three cities, achieving 0.91 accuracy in New York, 0.87 accuracy in London and Sydney. In comparison, methods with less context information do not provide good performance. For example, in New York, the accuracy of SoRec is 0.67 while SocialMF is 0.73, and the precision of SoRec is 0.68 while SocialMF is 0.73. We observe similar trends in all three cities.

The parameters used in our experiments are determined by a grid search using the 10% validation sets. The following equation parameters are set up for New York, London, and Sydney respectively: In Eq. (9), the number of topics T we choose for LDA was 80, 75, 60 for the three cities. The value of α is typically set to $50/T$ while β is set to 0.02. In Eq. (7), the time decay parameter α defines the decreasing speed of past event's importance according to the time span and is set to 0.03, 0.01, and 0.005 for the three cities.

[6] If less than k users exist in the group who did not attend the event, we randomly select the remaining users from other groups. This is very rare.

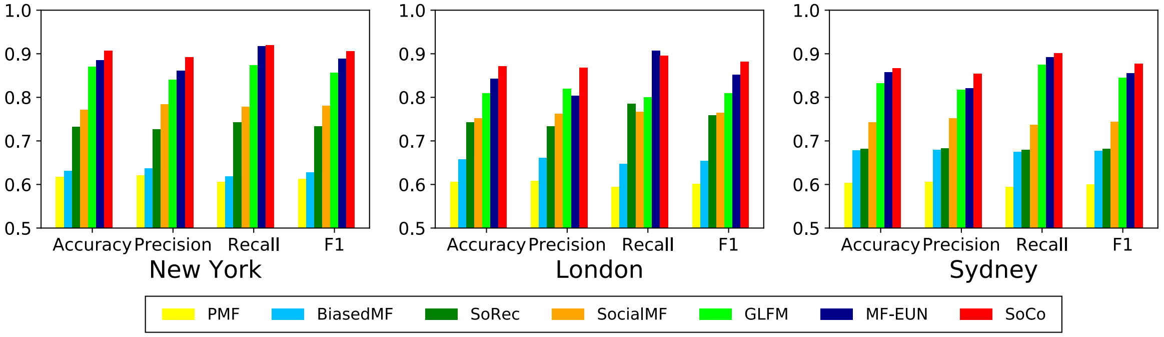

Fig. 7. Prediction performance comparison among different methods.

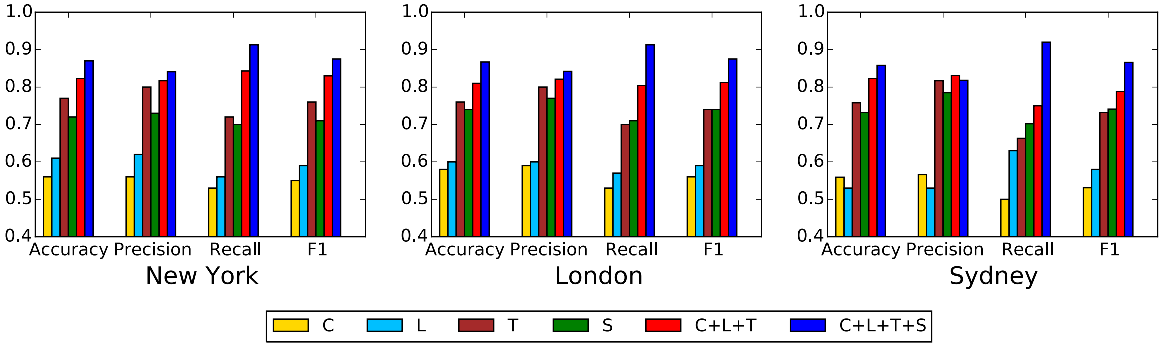

Fig. 8. Prediction performance comparison by SoCo among four individual features and different feature combinations: content (C), location (L), time (T), social (S), C+L+T fuses the C, L, and T features, and C+L+T+S fuses all four features (C, L, T, S).

5.3. Contributions of different features

Our proposed framework fuses together four different types of features: content (C), spatial (L), temporal (T) features, and the social feature (S) computed via hybrid EGU network. To understand the individual contributions of different features, we conduct experiments using individual C, L, T, S features, as well as the fused features of C+L+T and C+L+T+S (both using the well-performing SoCo algorithm for multi-feature fusion). We ran 10-fold cross validation under each experimental setup and report the average values for the four prediction performance metrics.

Fig. 8 shows the results for the three cities. We can see that all four individual features contribute to the prediction performance, while our proposed temporal feature with time decay and group-based social feature via RWR achieve better performance than that of content and spatial features. Using the social feature alone allows us to achieve >0.70 performance across all four metrics in all three cities. The temporal feature is the best-performing feature among all the four individual features, achieving >0.76 accuracy for all three cities. This result demonstrates the importance of the temporal feature for EBSNs: event time is an important factor for users to determine whether to attend an event or not; and the effect of time decay should be considered as recent patterns better predict users' preferences for event participation. Furthermore, C+L+T+S outperforms C+L+T significantly in all three cities for all four metrics, which demonstrates the effectiveness of the EGU-based social feature we have developed.

5.4. Impact of user and group activeness

In EBSNs, it is natural to have users who participate in more events than others, and groups that host more events than others. This is demonstrated by the large standard deviation values in Table 10 and Table 11, which show the average and standard deviation of the number of events attended per user and the number of events hosted per group for the three cities. The different activity levels of users and groups may lead to different performance when predicting different users' participation in events hosted by different groups. To understand how our proposed solution performs under these different scenarios, we evaluate the prediction performance for users and groups with different activities levels. Fig. 9 shows how the predication accuracy changes using SoCo for users who have attended different number of events and for groups who have hosted different number of events. We expect better performance for more active users and groups, since more training data are available for those users and groups. As shown in Fig. 9, the prediction accuracy increases from below 0.85 to above 0.90 as users' and groups' event activities increase. This is reasonable since more event activities by users and groups offer more insights into users' event participation behaviors, which can be captured by the proposed method. Besides that, it is good to note that even for users who attended at most 5 events and groups that hosted at most 25 events, our solution still achieves above 0.80 accuracy.

Table 10
#Events attended per user.

City	New York	London	Sydney
Average #Events / User	6.95	8.34	6.32
Standard Deviation	15.73	14.91	22.55

Table 11
#Events hosted per Group.

City	New York	London	Sydney
Average #Events / Group	29.82	38.75	39.16
Standard Deviation	54.15	61.73	65.54

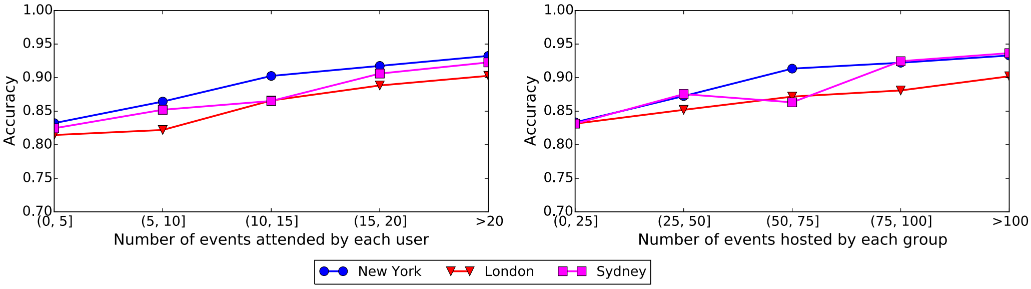

Fig. 9. Prediction performance comparison for (Left) users who have attended different number of events, and (Right) groups that have hosted different number of events.

6. Conclusions and future work

This work aims to address the problem of predicting users' event participation in event-based social networks. Using Meetup datasets collected from three cities with diverse cultures, we have conducted a detailed analysis of group-based event behaviors, as well as users' content, spatial, and temporal preferences for event participation. Based on this analysis, we propose a group-based event participation prediction framework that (1) constructs a hybrid EGU network and utilizes personalized random work with restart to infer the social closeness between users and events; and (2) fuses together four types of newly-designed features (content, location, time, social) to make a unified prediction regarding users' event participation. Evaluation results using the Meetup datasets demonstrate that (1) the proposed prediction framework achieves high prediction performance in all three cities; (2) the group-based social feature obtained via EGU is an important factor for predicting user event participation; and (3) the framework performs better for active users and active groups but still achieves good prediction performance for inactive users and inactive groups.

Our work offers valuable new insights into the increasingly more popular EBSNs and demonstrates the effectiveness of predicting user event participation. As our future work, we will continue studying group-based behavior patterns and exploit those patterns for further improvement of our prediction framework. We would like to investigate specific city, culture, group characteristics that impact event participation and integrate them for better prediction. Further analysis using more precise user home location information (e.g., via user mobility traces) would be helpful. Our current evaluation used random splitting of the user-event pairs, further investigation that strictly follows the temporal ordering would be interesting. Finally, we would like to investigate better fusion strategies and how the proposed prediction framework can be utilized for more effective event recommendation and organization.

Acknowledgment

This work was supported in part by the US National Science Foundation (NSF) through grant CNS 1528138.

References

[1] Meetup about, (http://www.meetup.com/about/).
[2] Facebook events, (http://events.fb.com/).
[3] Douban events, (http://www.douban.com/location/).
[4] Y. Hu, Event Analytics on Social Media: Challenges and Solutions, Ph.D. thesis, Arizona State University, 2014.
[5] S. Guha, S.B. Wicker, Do birds of a feather watch each other?: Homophily and social surveillance in location based social networks, in: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, 2015, pp. 1010–1020.
[6] Y. Huang, Y. Tang, Y. Wang, Emotion map: A location-based mobile social system for improving emotion awareness and regulation, in: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, 2015, pp. 130–142.
[7] R. Priedhorsky, A. Culotta, S.Y. Del Valle, Inferring the origin locations of tweets with quantitative confidence, in: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, 2014, pp. 1523–1536.
[8] R. Du, Z. Yu, T. Mei, Z. Wang, Z. Wang, B. Guo, Predicting activity attendance in event-based social networks: content, context and social influence, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2014, pp. 425–434.
[9] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, J. Han, Event-based social networks: linking the online and offline social worlds, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 1032–1040.
[10] A.Q. Macedo, L.B. Marinho, R.L. Santos, Context-aware event recommendation in event-based social networks, in: Proceedings of the 9th ACM Conference on Recommender Systems, ACM, 2015, pp. 123–130.
[11] Z. Yu, R. Du, B. Guo, H. Xu, T. Gu, Z. Wang, D. Zhang, Who should i invite for my party?: combining user preference and influence maximization for social events, in: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2015, pp. 879–883.
[12] H. Tong, C. Faloutsos, J.-Y. Pan, Fast random walk with restart and its applications, in: ICDM '06: Proceedings of the Sixth International Conference on Data Mining, IEEE, 2006, pp. 613–622.
[13] P. Georgiev, A. Noulas, C. Mascolo, The call of the crowd: Event participation in location-based social services, in: Proceedings of the 8th International Conference on Weblogs and Social Media, 2014, pp. 141–150.
[14] S. Zhang, Q. Lv, Event organization 101: Understanding latent factors of event popularity, in: International AAAI Conference on Web and Social Media, AAAI, 2017, pp. 716–719.
[15] J. Bao, Y. Zheng, D. Wilkie, M. Mokbel, Recommendations in location-based social networks: a survey, Geoinformatica 19 (3) (2015) 525–565.
[16] M. Massimi, Exploring remembrance and social support behavior in an online bereavement support group, in: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, ACM, 2013, pp. 1169–1180.
[17] M. Massimi, J.L. Bender, H.O. Witteman, O.H. Ahmed, Life transitions and online health communities: reflecting on adoption, use, and disengagement, in: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, 2014, pp. 1491–1501.
[18] G. Hsieh, Y. Hou, I. Chen, K.N. Truong, Welcome!: social and psychological predictors of volunteer socializers in online communities, in: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, ACM, 2013, pp. 827–838.
[19] J. Zou, R. Meir, D. Parkes, Strategic voting behavior in doodle polls, in: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, 2015, pp. 464–472.
[20] S. Zhang, K. Alanezi, M. Gartrell, R. Han, Q. Lv, S. Mishra, Understanding Group Event Scheduling via the OutWithFriendz Mobile Application, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1 (4) (2017) 175.
[21] H. Ma, H. Yang, M.R. Lyu, I. King, Sorec: social recommendation using probabilistic matrix factorization, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management, ACM, 2008, pp. 931–940.
[22] M. Jamali, M. Ester, A matrix factorization technique with trust propagation for recommendation in social networks, in: Proceedings of the Fourth ACM Conference on Recommender Systems, ACM, 2010, pp. 135–142.
[23] X. Liu, K. Aberer, SoCo: a social network aided context-aware recommender system, in: Proceedings of the 22nd International Conference on World Wide Web, ACM, 2013, pp. 781–802.
[24] M. Girvan, M.E. Newman, Community structure in social and biological networks, Proc. of the National. Academy of Sciences 99 (12) (2002) 7821–7826.
[25] M.E. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical Review E 69 (2) (2004) 026113.
[26] Meetup dataset, (http://kennyjoseph.github.io/2013/10/29/blog_post_pt1.html).
[27] F. Yates, Contingency tables involving small numbers and the χ 2 test, Supplement to the Journal of the Royal Statistical Society 1 (2) (1934) 217–235.
[28] C.S. Fischer, To Dwell Among Friends: Personal Networks in Town and City, University of Chicago Press, 1982.
[29] M. Ye, P. Yin, W.-C. Lee, D.-L. Lee, Exploiting geographical influence for collaborative point-of-interest recommendation, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011, pp. 325–334.

[30] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[31] R. Navigli, S. Faralli, A. Soroa, O. De Lacalle, E. Agirre, Two birds with one stone: learning semantic models for text categorization and word sense disambiguation, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 2317–2320.

[32] R.-H. Li, J.X. Yu, J. Liu, Link prediction: the power of maximal entropy random walk, in: Proceedings of the 20th ACM International conference on Information and Knowledge Management, 2011, pp. 1147–1156.

[33] S. Chang, V. Kumar, E. Gilbert, L.G. Terveen, Specialization, homophily, and gender in a social curation site: findings from Pinterest, in: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, 2014, pp. 674–686.

[34] M. Klein, T. Maillart, J. Chuang, The virtuous circle of Wikipedia: Recursive measures of collaboration structures, in: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 2015, pp. 1106–1115.

[35] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford InfoLab, 1999.

[36] A. Mnih, R.R. Salakhutdinov, Probabilistic matrix factorization, in: Advances in Neural Information Processing Systems, 2008, pp. 1257–1264.

[37] Y. Koren, Factorization meets the neighborhood: a multifaceted collaborative filtering model, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 426–434.

[38] Y. Jhamb, Y. Fang, A dual-perspective latent factor model for group-aware social event recommendation, Information Processing & Management 53 (3) (2017) 559–576.

[39] X. Li, X. Cheng, S. Su, S. Li, J. Yang, A hybrid collaborative filtering model for social influence prediction in event-based social networks, Neurocomputing 230 (2017) 197–209.