Knowl Inf Syst
https://doi.org/10.1007/s10115-018-1194-x

REGULAR PAPER

Who should I invite: predicting event participants for a host user

Jyun-Yu Jiang[1] · **Cheng-Te Li**[2,3]

Received: 2 October 2016 / Revised: 28 December 2017 / Accepted: 18 April 2018

Abstract While users can interact with others online, more and more social networking services can help people to organize various offline social events, such as dinner parties and study groups, on the Internet. The hosts can invite friends or strangers to participate in their events in either manual or collaborative manner. However, such invitation manners may cost substantial time. Besides, the invitees may be uninterested or even unexpectedly contain spammers. In this paper, we aim at developing a predictive model to accurate recommend event participants. Specifically, given the host who initializes a social event, along with its event contexts, including the underlying social network, categories, and geolocations, our model will recommend a ranked list of candidate participants with the highest participation potential. We propose a feature-based matrix factorization model that optimizes pairwise errors of user rankings for training events, using six categories of features that represent the tendency of a user to attend the event. Experiments conducted on two event-based social networks Meetup and Plancast and Twitter retweet data exhibit the promising performance of our approach, together with an extensive study to analyze the factors affecting users' event participation.

Keywords Event participants · Event-based social network · Participant prediction · Online social ties · Offline geographical activities

✉ Cheng-Te Li
reliefli@gmail.com

[1] Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA

[2] Department of Statistics, National Cheng Kung University, Tainan, Taiwan

[3] Institute of Data Science, National Cheng Kung University, Tainan, Taiwan

Published online: 26 April 2018

1 Introduction

Event-based social networking (EBSN) services, such as Meetup[1], Plancast[2], and Facebook Events, are getting popular recently. EBSNs provide a rich set of functions for users to manage social connections and communicate with one another in their online platforms, and enable users to keep track of their offline geographical activities via mobile devices in the real world. The most important feature of EBSNs is "event." Users can initiate diverse kinds of social events (or called activities interchangeably), such as gathering, study group, business meeting, hiking, and dining out, by specifying its objective, tags, and geolocation, and send RSVP (i.e., a request for a response from the invited person). The user who organizes the event is termed *host*. Then, they can invite their friends or anyone of interest to join their hosted events. By responding to the RSVP in EBSNs, those feel interested and available can accept the invitations. To attract more relevant users or make the event successful, the host user will spend much time to identify which users can significantly fit the event and have higher possibility to accept the invitation. However, manual invitations by the host may neglect some important users and need too much time to complete the process of event organization. Although some EBSN services allow the invited users to further invite other users, human subjectivity or personal bias may lead to include either irrelevant users. As mentioned above, we believe it is highly demanded and useful to have a smart participant recommender system so that the discovery of proper event participants can be more satisfying.

In this paper, to facilitate the formation of event participants in EBSNs, we aim at developing a predictive model to recommend a ranked list of candidate participants for the host/organizer of a social event. Specifically, given a host user who initiates an event, the event information describing the detailed settings, and the historical interactions among users, our goal is to recommend the event participants by predicting which users will participate the event. The event information specified by the host includes a set of tags describing the theme and the geographical location. The historical interactions among users include the online social connections among users and the offline geographical geographical footprints of each user. Many real-world scenarios need the recommendation of event participants. For example, someone intends to host a cocktail party with the old-school style at home and desires friends or friends of friends who are available, interested, or nearby can come to participate. An enthusiastic guy plans to organize a study group about social network analysis (SNA) at a certain coffee shop and wishes participants can not only have knowledge about SNA but also be willing to share the knowledge in person. A group leader is planning to start a campaign against nuclear power at a particular plaza and wants to attract those who feel interested to participate physically.

It is important to note that one may argue that such problem setting does not consider the semantics of events, because different events (e.g., bachelorette party, hiking activity, and study group for social computing) need to specify different topics or demographic information to attract diverse groups of people to attend. In fact, in our problem setting, the host is allowed to specify his/her tags. Therefore, the topics and demographic information about the event can be provided via the tags of the host. In other words, the proposed problem can be easily generalized to deal with different kinds of events with various topics.

The most relevant work is *Event-Centric Diffusion* (ECD) [11], in which the problem is: given a set of early participants of an event, predicting which users in the EBSN will accept to attend the event in the future. However, in the real-world cases, general events

[1] Meetup: http://www.meetup.com/.

[2] Plancast: http://plancast.com/.

(e.g., gatherings, study groups, and outdoor activities) are usually initiated by few or only a single user. We think it can benefit the organization of social activities if we can accurately recommend participants for the single event host. Therefore, we concentrate on tackling the cold-start participant prediction in which *only the host* of an event is given. Another research line is *Event Recommendation*, which aims at recommending events to each users in EBSNs [5,18,23,28] considering "events" as so-called items in traditional recommendation settings. While event recommendation is to recommend *multiple general events* to users, our work is as predicting whether users will accept a *single-host-based event*, where general events refer to events without specified hosts.

The event participant prediction task is challenging. The reason is twofold. The first is *data sparsity*. Some users may participate in few events (e.g., < 30) in the history, and each event has only a small number of participants (e.g., < 30). Therefore, different from conventional recommenders where users often have a long consumption history, it is difficult to learn the correlation between the preferences or willingness of users and the events. The second is *cold-start*. Participants are recommended to *only one* host user who may either a newcomer (i.e., have rare or no experience on organizing events) or have ever failed to host events (e.g., attracting few or no participants). That says, we are facing a limited event organization history for host users. Note that although it is true that many events are organized by a group of persons, we think it is also very common for an ordinary user to initiate an event and thus target the problem at recommending for a host.

In order to predict participants for the host of a given event, we develop a novel *Participant-Ranking Matrix Factorization* (PRMF) model. PRMF is a kind of feature-based matrix factorization. The general idea of PRMF is to rank the participants at higher positions in the recommended list by optimizing pairwise errors of user rankings for training events. To model the preferences and willingness of users to attend events, we propose six categories of features using the semantic information (e.g., tags), geographical information (e.g., locations), and online social network among users. Each feature category is validated to be useful in an extensive analysis. We use Meetup, Plancast, and Twitter datasets to examine the performance of PRMF, comparing with a series of competitors, including the state-of-the-art method Event-Centric Diffusion (ECD) [11]. In addition, we further analyze the effectiveness of each feature set and discuss which factors about events and hosts have higher impact on effective event participation prediction.

Here, we highlight the contributions of this work as below.

- We propose the *event participant prediction* problem, predicting which users will participate in the event organized by a single host, along with the locations and tags to depict the event and the social connections to characterize the event host. Compared with existing recommender systems, recommending participants for single host with such spatial, semantic, and social information is novel and challenging since solving the problem needs to face the issues of data sparsity and cold-start.
- We develop the *Participant-Ranking Matrix Factorization* (PRMF) model, which can effectively make use of historical event participation to generate a list of recommended users. Six categories of features using geographical, semantic, and social information are proposed and extracted, along with the extensive data and feature analysis.
- Experiments conducted on Meetup, Plancast, and Twitter datasets show that PRMF can clearly outperform the state-of-the-art competing method. A series of parameter analysis and feature study are also performed, and the results exhibit the practical effectiveness under a variety of real event settings, including training size, number of participants, active/inactive host, and tag popularity.

This paper is organized as below. We first provide the literature review in Sect. 2, followed by a preliminary data analysis to uncover why users participate in events in Sect. 3. Then, in Sect. 4, we define the event participant prediction problem and present the technical details of our Participant-Ranking Matrix Factorization model, along with six categories of features. We present the experimental studies in Sect. 5 and conclude this work in Sect. 6.

2 Related work

The relevant studies about event participant prediction can be divided into two parts: *Event-Centric Diffusion Prediction* and *Event Group Recommendation*. Liu et al. [11] is the first attempt to analyze event-based social networks and first propose to predict event participation, termed *Event-Centric Diffusion* (ECD). Their general setting is that assume an event is created at time t_c and ends at time t_s, which is the time that the event takes place in the real world. Given the event e at time t, where $t_c < t < t_s$, the task is to predict who will accept the RSVP to event e between t and t_s. All the users who accept to participate event e between t_c and t can be regarded as positive training instances to build the predictive model. It is apparent that our setting is more challenging and difficult than ECD [11]: We aim to predict participants for the host single user who creates the event at time t_c, rather than for the early participants. We think that in real-world applications, it would be more efficient and effective for the host to organize the event if some mechanism can instantly and accurately recommend the most potential participants so that the invitations can be sent right after the time that the event was created. Nevertheless, we consider the solutions to ECD (i.e., *online weights* and *offline weights*) as a compared method in the experiments. Note that although ECD [11] also examines the cold-start case (i.e., only the host is used), the performance is not satisfying. Although a recent study [26] also aims to find the potential participants of for the host of an event, they formulate the problem as a preference-based influence maximization problem to select the participants that can lead to higher influence spread. We argue that although finding seed users by maximizing the influence spread can bring more audience to the event, such seed users and audience may not truly attend the event. In this paper, therefore, we formulate it as an prediction problem that aims to find those who will participate the event in the future. Also due to such essential difference, it is unfair to make the comparison with their work [26]. While some work [21] formulates and tackles a simpler problem similar to ours, recommending individuals to a given venue so that they will truly visit there. However, their proposed method cannot handle our event participant prediction problem since neither users' preferences nor social connections are considered.

Event recommendation in event-based social networks (EBSN) aims at recommending either online social groups [27] or offline geographical events [5,18,23,28] to users and predict whether each user will be willing to join the groups and/or participate the events (held by other members in the group). Note the host's information is not considered in EBSN event recommendation. User spatiotemporal mobility patterns, place semantics, and social factors are jointly analyzed and used for event recommendation [5]. A heterogeneous graph-based recommender [17] is developed to jointly recommend groups and events to users. Some studies [3,14] further combine both content (event description) and context information (social, spatial, and temporal) to recommend events for a target user.

In addition, link prediction on social network can be taken into account to recommend friendships. Some studies [2] expand the friend list of each user for recommending potential friends. Meta information like tags [6] can be also applied to link users with similar properties.

Some work [8] model user rating behavior to infer social relations. However, all of them consider only friendship relations in online social network but not event participation. Some factors such as geographical and offline network information are not applied to infer user participation.

For single-host events, Jiang and Li [9] also show that these factors are important. A dynamic social influence approach [25], which models how user mutually influence their willingness on event participation, is devised to recommend events to users. Another recent study [7] quantifies degree of user engagement on events in Twitter for event recommendation for users. While studies in this task focus on treating "events" as "items" in the context of traditional recommendation, our study alternatively deals with another practical problem: recommending a set of potential users who will participate in the event hosted by a certain organizer.

3 Data analysis

In this section, we first conduct several data analysis to learn what factors may affect event participant prediction.

3.1 Data settings

For the data analysis, we adopt the real-world dataset of Meetup.com, which is an event-based social service. Users can express their interests to offline events by sending RSVPs. In the dataset, each RSVP represents that a user participated in a certain event. Each user has a registered location with its latitude and longitude. Events are also associated with the geographical information of event locations. To represent the personal interests, users can apply some tags such as "travel-photography" to describe themselves. Moreover, users can join online social groups and share comments and photographs with other members in the same groups. We provide the detailed description and statistics in Sect. 5.1.

We show the distribution of number of events for different event sizes in Fig. 1a. It can be observed that most events have very limited number of participants (e.g., < 50), and only few events are large scale (i.e., the participant number is higher than 400). In addition, the distribution of number of users for different numbers of event participations is shown in

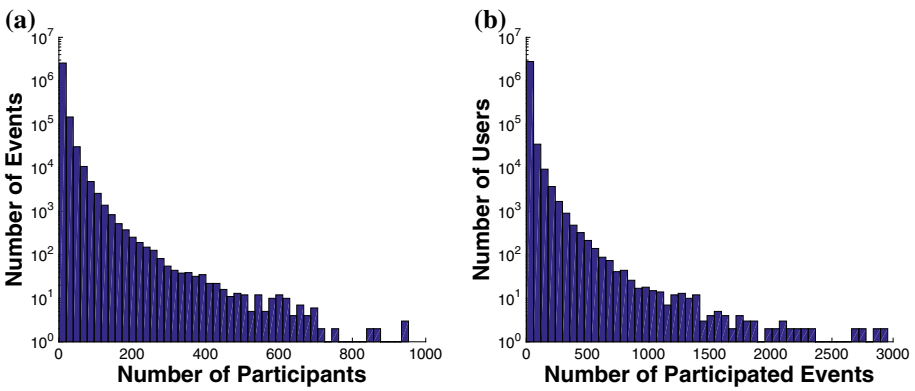


Fig. 1 The distributions of events and users. **a** Event distribution and **b** user distribution

Fig. 1b. Similarly, most users participate in fewer events (e.g., ≤ 50) while rare users attend events very frequently (e.g., ≥ 1000). These two distributions exhibit the severe data sparsity problem, along with our setting that the prediction is based on *only one* host user, which leads to the cold-start problem, so it is a very challenging task.

In order to make the prediction task less difficult, we do not consider those users who participate in < 10 events and those events with < 10 participants. Since the data do not provide the time that a user participates in an event, in the experiments, we randomly select a user as the host for each event. Besides, to eliminate the location bias, we separate users and events into the subsets of eight cities, including New York City (NYC), Log Angeles (LA), Chicago (CHI), San Diego (SD), San Jose (SJ), Phoenix (PHX), London (LDN) and Paris (PA), by the user locations. More statistics about the used Meetup data are shown in Table 4.

3.2 Semantic information

The tags can be treated as the semantic information for determining users' interests. In other words, these information is able to be a hint to find the relationships between users and the host. Users may be more likely to attend the events of the hosts with more common interests. For example, people who like outdoor activities may join the events created by the hosts who also enjoy going outside. In contrast, if the host has no any common interest, it will be so absurd for users to participate in the event created by that host.

For users with tags in their profiles, Fig. 2 shows the average number of common tags to the host in each dataset. It is so obvious that participants generally have more common tags to the host than non-participants in every dataset. In the NYC dataset, the participants averagely have more than five times as the number of common tags to the event host than the non-participants. The results also demonstrate that the tag similarity between users and the host may be very important to finding potential participants. If the tag information of users can be well incorporated, the system will be able to understand more semantic knowledge.

3.3 Geographical information

The geographical limitation may be an important decision whether a user participates in the event. If the event location is far from users' homes, they may be unwilling to pay the enormous effort of transportation for attending the event. On the contrary, people may be

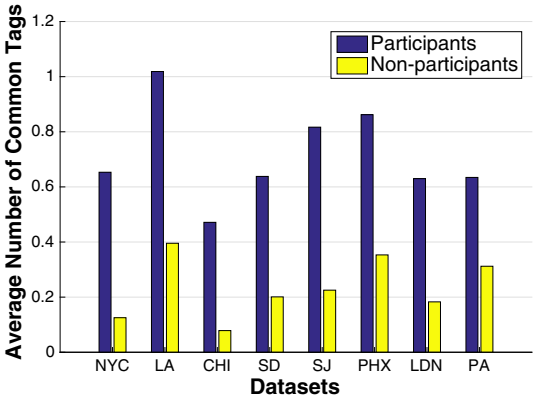


Fig. 2 The average numbers of common tags to the host in eight datasets for all users whose profiles include tags

Table 1 The median distances (miles) to the event locations for participants and non-participants in eight datasets

Dataset	NYC	LA	CHI	SD
Participants	**10.8973**	**14.9825**	**6.2850**	**19.7326**
Non-participants	11.0904	19.8551	7.6397	20.5993
Dataset	SJ	PHX	LDN	PA
Participants	**29.9376**	**20.7983**	40.2062	36.5179
Non-participants	30.4025	22.8540	**40.2024**	36.5179

Bold value indicates smaller distances

more willing to participate in the events hosted in closer places so that they can save the transportation costs in both of time and money.

Table 1 shows the median distance from users' homes to event locations in each dataset. We adopt the great-circle distance [15] to calculate the distance on the geographical coordinate system. To avoid the bias from extremely long or short distance, we compute the median distance instead of the average distance. Except the datasets of two European cities, participants are closer to the event locations than non-participants in almost of all datasets. The two European datasets contradict the others because users in the two datasets registered their locations so close to each other. The results show that people actually tend to join events held in closer places. Hence, the users who live in closer locations should have higher priority to be considered as participants. We think a method can more precisely learn users' preference if considering their location information.

Note that the results in Table 1 exhibit geodistance is correlated with event participation, although they are not so significant. Nevertheless, we think we cannot say geodistance does not matter because it is only one of the factors that can influence event participation. In other words, the key of event participation involves multiple factors. So it is unconvincing to separately discuss the performance of geodistance in recommending participants. Geodistance needs to be considered together with other factors such as semantic information in Sect. 3.2 and social information in Sect. 3.4. We will discuss how different combination of factors affect the performance of predicting event participants in Sect. 5.6 via the leave-one-out evaluation.

3.4 Social network information

The online social groups in Meetup.com can well form an online social network. The interaction among users in that social network may affect users' participation. Users may be more likely to participate in the events hosted by users in the same online social groups. On the contrary, people may not like to join an offline event hosted by an unknown person.

To construct the online social network, we first treat each user as a node in a social network. For each pair of users, an edge between two users will be created if they are the members of the same online social group. A direct way to measure the relationship between two nodes in a social network is the distance on the graph. If a user has closer relationship to the host, the network distance between them may be also shorter. Figure 3a shows the average network distance between users and the event hosts in each dataset. Note that we define the network distance between two nodes as the number of nodes in the shortest path in the network. The results show that the network distances of participants are significantly shorter than the distances of non-participants. We also measure the relationship between users by the number of co-friends on the social network. Here, we define the friends are the users who are the

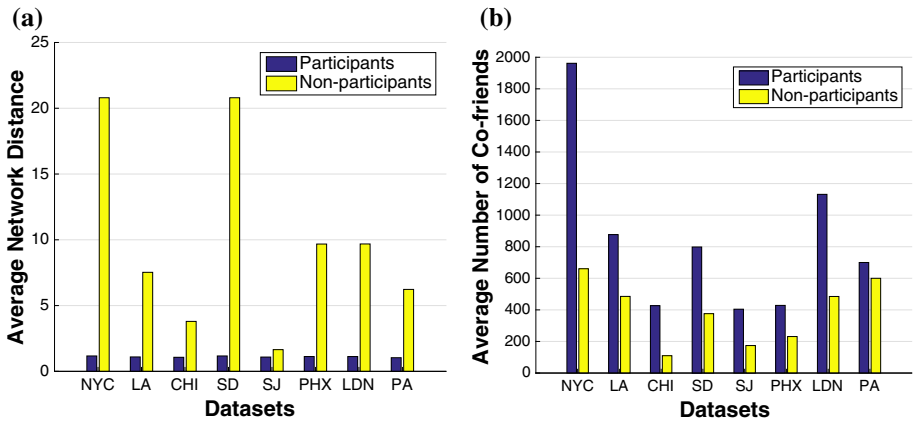


Fig. 3 The average network distances and numbers of co-friends between the event host and users in the online social networks (groups) of eight datasets. **a** Network distance and **b** number of co-friends

members in the same online social group. Figure 3b represents the average number of co-friends between users and the event hosts in each dataset. The participants generally have more co-friends on the online social networks of all datasets. In sum, both of the measures show that the online social network may be helpful to discover the users who will join the offline events.

In addition to online social networks, the event participation history of users may also form an offline social network. It is intuitive that users would like to participate in the events held by people who had ever attended the same events. To conduct the data analysis, we use 50% events in each dataset to construct the offline social network and present some statistics with the remaining events. For each pair of users, an edge is created if they had ever participated in the same offline events. Figure 4a, b represents the average network distance and the number of co-friends between the event host and users in the offline social network. The results are consistent with the analysis of online social networks. Participants of events have also shorter distances and more co-friends to the hosts than non-participants. Moreover, the difference on the offline social networks between participants and non-participants is much more significant than the difference on the online social networks. Hence, offline social networks may be more effective than online social networks for determining the event participants.

To summarize the results of data analysis, all three kinds of information may be so helpful to find event participants. If we can incorporate these useful information into our approach, the performance of predicting event participants may be much boosted.

4 Event participant prediction

Problem formulation We first formally define the problem of event participant prediction for a host. Let U be the set of users and E be the set of training events. Table 2 indicates the clear definitions of notations used in this work. For each user $u \in U$, $L(u) = (\text{lo}(u), \text{la}(u))$ represents the longitude and the latitude of user's registered location. $T(u)$ and $G(u)$ represent the set of semantic tags and participated online groups of the user u. The set $F(u) \subseteq U \setminus u$ states u's friends on the Internet. For each event e, $H(e) \in U$ denotes the host of e, and $R(e)$ is the set of its participants. Like users with registered locations, each event also has

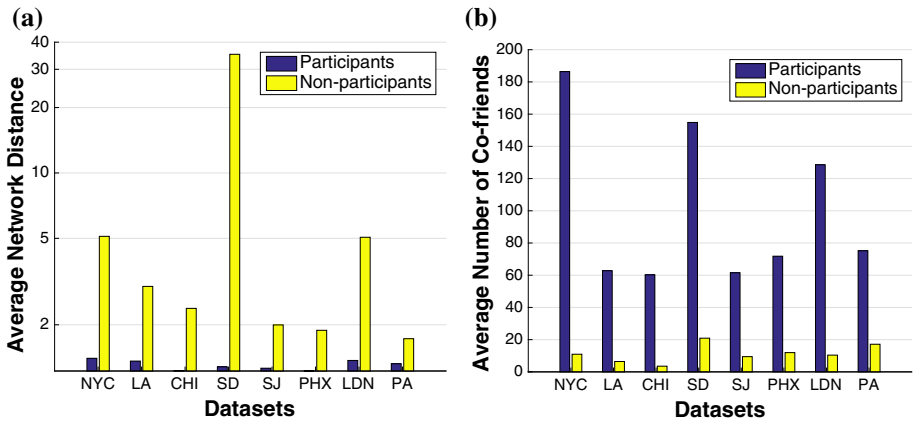


Fig. 4 The average network distances and numbers of co-friends between the event host and users in the offline social networks (events) of eight datasets. Note that here we build the offline social network with 50% events and do statistics with the remaining events. **a** Network distance and **b** number of co-friends

Table 2 The definitions of notations

Notation	Definition
u	A certain user
U	The set of users
e	A certain event
E	The set of events
$h = H(e)$	The host of the event e
$R(e)$	The actual participants of the event e
$L(u) = (\text{lo}(u), \text{la}(u))$	The geographical location of the user u
$T(u)$	The semantic tags of the user u
$G(u)$	The participated online groups of the user u
$F_{\text{online}}(u)$	The online friends of the user u
$F_{\text{offline}}(u)$	The offline friends of the user u

the location information $L(e)$, at which it holds. For the online and offline social networks, we denote $F_{\text{online}}(u)$ as the set of friends on the online network, and $F_{\text{offline}}(u)$ as the friends on the offline network. Given the host $h = H(e)$ of an event e and the training data E, our goal is to generate a ranking to all other users $u \in U \setminus h$ as candidate users so that ones with higher rankings are more likely to be in the actual participants $R(e)$.

Note that one may argue that such problem setting does not consider the semantics of events, because different events (e.g., bachelorette party, hiking activity, and study group for social computing) need to specify different topics or demographic information to attract diverse groups of people to attend. In fact, in the problem definition, the host is allowed to specify his/her tags. Therefore, the topics and demographic information about the event can be provided in the form of tags. In other words, the proposed problem can be generalized to deal with different kinds of events with various topics.

We propose a feature-based *Participant-Ranking Matrix Factorization* framework. The framework starts from the formation of the factorization machine [19] for incorporating user

collaboration as well as the knowledge in training events and social networks. By optimizing pairwise errors of user rankings for training events, our approach can discover appropriate parameters to rank participants in higher positions. To describe the characteristics of users and events, we also devise six sets of features.

4.1 Feature-based matrix factorization

To exploit the information of user collaboration, collaborative filtering such as matrix factorization [10] is one of the most applicable solutions. Although some existing methods [12,13] had incorporated social network structures with collaborative filtering for different kinds of recommendation tasks, the design of collaborative filtering-based approach for predicting event participants is still to be resolved. In this work, we adopt the concept of *Factorization Machine* (FM) [19] and extend FM to simultaneously consider user collaboration, user information, and the structure of event-based social network.

Although FMs can exploit additional information as several features in the model, there are still some drawbacks for our task. For an event, all users will be ranked for a single host, so the bias terms for hosts are meaningless. Hence, the model should be revised. We propose to modify the 2-way FM because it can capture all single and pairwise interactions between features and parameters. For the candidate user u and the host user h, the prediction can be formulated as follows:

$$\hat{r}_{hu} = b_u + \boldsymbol{w}^T \boldsymbol{x}^{\boldsymbol{hu}} + \frac{1}{2} \sum_i \left(2 \cdot x_i^{hu} \boldsymbol{v_i}^T \boldsymbol{z_u} + \sum_{i \neq j} x_i^{hu} x_j^{hu} \boldsymbol{v_i}^T \boldsymbol{v_j} \right),$$

where $\boldsymbol{x}^{\boldsymbol{hu}} \in \mathbb{R}^N$ is the vector of N extracted features for the host h and the user u; $b_u \in \mathbb{R}$ and $\boldsymbol{w} \in \mathbb{R}^N$ are the bias terms of the user u and each feature. For the user u and the i-th feature, $\boldsymbol{z_u}, \boldsymbol{v_i} \in \mathbb{R}^k$ are the latent parameters with k factors. All bias terms and latent parameters should be trained by the model and training data.

To be more precise, we aim to estimate representations for both individual users and context features in the identical hidden space. Users will have their own vectors $\boldsymbol{v_i}$ to describe preferences and characteristics while each feature will also have another vector representing information in the same space. With these representations, the interactions among users and features can be revealed, thereby estimating the potential for users to attend an event.

4.2 Optimization for participant ranking

With the representation of the prediction task, the next step is to optimize the parameters $\boldsymbol{w}$, $\boldsymbol{v}$ and $\boldsymbol{z}$ so that the predicted values of participants are greater than ones of absent users. As the approach in [20], we attempt to minimize the pairwise errors in rankings by enumerating participants and other users absent from events. The objective function of our approach can be described as follows:

$$\underset{\boldsymbol{w},\boldsymbol{v},\boldsymbol{z}}{\operatorname{argmin}} \sum_{e \in E} \sum_{u_i \in R(e)} \sum_{u_j \in U - R(e)} -\ln(\sigma(\hat{r}_{hu_i} - \hat{r}_{hu_j})) + \frac{\lambda}{2} \left(||\boldsymbol{w}||_F^2 + ||\boldsymbol{v}||_F^2 + ||\boldsymbol{z}||_F^2 \right),$$

where σ is the logistic sigmoid $\sigma(x) = \frac{1}{1+e^{-x}}$; $||\cdot||_F$ represents the Frobenius norm for regularization; and λ is the regularization parameter. The difference between predictions $\hat{r}_{hu_i} - \hat{r}_{hu_j}$ captures the relationship between the candidate users u_i and u_j for the host user h, so the objective function can well estimate the pairwise errors. With optimization methods

Table 3 Defined features and the corresponding formulas

Category	Feature class	# of features	Formulas
User latent features	User feature	1	u
	Host feature	1	h
User statistics	Participation frequency	2	$\lvert\{e : u \in R(e), \forall e \in E\}\rvert$, $\lvert\{e : h \in R(e), \forall e \in E\}\rvert$
Semantic features	Number of tags	2	$\lvert T(u)\rvert$, $\lvert T(h)\rvert$
	Common tags	3	$\lvert T(u) \cap T(h)\rvert$, $\frac{\lvert T(u)\cap T(h)\rvert}{\lvert T(u)\rvert}$, $\frac{\lvert T(u)\cap T(h)\rvert}{\lvert T(h)\rvert}$
	Tag similarity	1	$\mathrm{Sim}_{\cos}(T(u), T(h))$
Online network features	Co-participated groups	1	$\lvert G(u) \cap G(h)\rvert$
	Online weights	1	$w^{\mathrm{on}}_{u,h} = \sum_{g \in G(u)\cap G(h)} \frac{1}{\lvert\{u' : g \in G(u'), \forall u' \in U\}\rvert}$
	Number of co-friends	1	$\lvert\{u' : u' \in F_{\mathrm{online}}(u), u' \in F_{\mathrm{online}}(h), \forall u' \in U\}\rvert$
	Network distance	1	$\mathrm{distance}_{\mathrm{online}}(u, h)$
Offline network features	Co-participated events	1	$\lvert\{e : u \in R(e), h \in R(e), \forall e \in E\}\rvert$
	Offline weights	1	$w^{\mathrm{off}}_{u,h} = \sum_{e \in \{e : u \in R(e), h \in R(e), \forall e \in E\}} \frac{1}{\lvert\{u' : u' \in R(e), \forall u' \in U\}\rvert}$
	Number of co-friends	1	$\lvert\{u' : u' \in F_{\mathrm{offline}}(u), u' \in F_{\mathrm{offline}}(h), \forall u' \in U\}\rvert$
	Network distance	1	$\mathrm{distance}_{\mathrm{offline}}(u, h)$
Location features	Host distance	1	$\mathrm{distance}(L(u), L(h))$
	Event distance	1	$\mathrm{distance}(L(u), L(e))$
	Preferred location	2	$\mathrm{distance}(L_{\mathrm{pre}}(u), L(h))$, $\mathrm{distance}(L_{\mathrm{pre}}(u), L(e))$

such as Gradient Descent, the model can reach a local minimum and correctly optimize the errors.

In this paper, we apply *Stochastic Gradient Descent* (SGD) [1] to optimize the model. However, even though SGD is one of the fastest optimization methods, each iteration still costs $O(kN^2|E||U|^2)$. It will be so time-consuming when the users are numerous. Hence, in our implementation, the *bootstrap sampling* technique [4] is applied to accelerate the optimization process. In each iteration, for each participant of an event, we randomly sample C absent users for optimization. Note that the set of sampled users will be different in each iteration. Then, the time complexity of each iteration can be reduced to $O(kCN^2|E||U|)$. In fact, because the number of participants is usually limited, the optimization is generally faster than the expectation of its time complexity. We set the sample size of absent users $C = 1$ so that each training participant will be matched to a different absent user in each iteration for optimization.

4.3 Feature extraction

To model the preferences and willingness of users to participate events, we propose six categories of features. These features are supposed to recognize different situations for par-

ticipants and hosts and then adjust the model into a better status. Note that these features are denoted as $\boldsymbol{x}^{\boldsymbol{hu}}$ in Sect. 4.1. The completed list and the detailed formulas of each feature are listed in Table 3.

4.3.1 User latent features

In our approach, users have their bias terms and latent parameters, so the identity of users can be fairly treated as a kind of features in the model. More specifically, the candidate user u and the host user h in the model can be represented as two categorical features.

4.3.2 User statistics

The statistical information of user participation history may reveal their preference. By user preference, the system can further understand the behaviors of users. In this paper, we consider the participation frequency as the feature of user statistics.

- **Participation frequency** Different users may have different preferences for participating in offline events. Some active users may tend to participate in much more events. For users with similar interests, active users may be more likely to attend events. The activeness of the host may also affect the distribution of the participants. The participation frequencies of both the candidate user and the host will be calculated. Here, we use the number of participated events in the training data as a numerical feature.

4.3.3 Semantic features

Some EBSN services allow users to state some semantic information, such as interests and tags, in the profiles. The semantic information can not only understand users' characteristics but also discover candidate users whose interests are similar to the host user.

- **Number of tags** The size of the tag set $|T(u)|$ can reflect the diversity of user's interests. Users with more diverse interests may tend to attend events in different topics. The interest diversity of hosts may also affect user participation. Hence, we compute the numbers of tags for the user and the host as two features.
- **Common tags** The common tags of the user and the host $|T(u) \cap T(h)|$ can represent their shared interests. Users may be more likely to attend the event if the host has more interests. Besides, the influence of common tags may depend on the size of the tag set. For example, if a user with 10 tags has eight common tags to the host, these tags can cover 80% of interests; however, the same number of common tags can cover only 8% of interests for another user with 100 tags. Therefore, we compute not only the number of common tags, but also the ratio of common tags to the tag set for the user and the host.
- **Tag similarity** The set of tags can be treated as a bag of words, so we can apply the vector space model to compute the similarity between interests of the user and the host. Here, we calculate the cosine similarity of two tag sets $\mathrm{Sim}_{\cos}(T(u), T(h))$ to represent the tag similarity.

4.3.4 Online and offline social network features

The structures of social networks may be so helpful to discover event participants. User interactions on a EBSN can be used to construct social networks. We consider online and offline

social networks. For online networks, we focus on the online groups and social relationships of users on the Internet. The online social network is constructed based on online groups, in which users are nodes, and an edge is created for two users who have ever joined at least one group. For offline networks, the participated events are utilized to construct the real-world relationships between users. The offline social network is constructed based on offline events, in which users are nodes, and an edge is created for two users who have ever participated in at least one event. From either online or offline networks, we can have the following four features. Consequently, there are totally eight features in this section. Since both networks have the same four features, we describe them together in the following.

- **Co-participated groups and events** A direct solution to quantify the social relationship between the user and the host is to compute the numbers of co-participated groups and events. For online networks, users with more co-participated groups may have more common interests and more interactions with the host in the Internet. For offline networks, users with more co-participated events may be more likely to attend the offline events with the host. The formulas for computing features of co-participated groups and events are shown in Table 3, respectively.
- **Online and offline weights** To measure the social interactions between users, previous work also proposed several approaches to calculate the weights on social networks. For online networks, [16] uses the reciprocal of each co-participated group as the contribution from that group. For offline networks, [11] applies the same concept to treat each offline event as a group to compute the weights of offline social interactions. Here, we compute the above two weights as the features for representing the social interactions between the user and the host in both of online and offline social networks. The formulas for computing features of online and offline weights are shown in Table 3, respectively.
- **Number of co-friends** The number of common friends is also an important factor to estimate the relationships between users in social network analysis. For online networks, we define that users are friends if they are in the same group or followed by each other. For offline networks, users participated in the same event may be friends in the real world. Then, we adopt the numbers of co-friends in the online and offline social networks as two features, whose detailed formulas are provided in Table 3, respectively.
- **Network distance** Some participants do not have direct relationships to the host, but the structure of social networks may reveal additional information. The distance on social networks may represent the closeness between users. Hence, we compute the network distances between the user and the host on both of online and offline social networks as two features, as elaborated in Table 3, respectively.

4.3.5 Location features

The geographical information is one of the directest ways to link online users to the real world. The results in [9] also exhibit that the geographical information may affect users' event participation. Here, we derive four location features to incorporate these information.

- **Host distance** The distance to the host may directly influence users' willingness to attend the events. Users may tend to participate events which are hold in closer places. Hence, we use the *great-circle distance* [15] between the user and the host as a numerical feature. We adopt the great-circle distance because it calculates the real distance on the sphere of Earth and is a popular geographical distance utilized in previous studies [22].
- **Event distance** Some events may not be held nearby the homes of hosts. The distance to the event location may be more precisely estimate the user intent for participating the

event. The potential effectiveness of the event distance for predicting event participants is also shown in [9].

- **Preferred location** Users may have their preference about the event locations. To describe the user preference, we compute the centroid of events in which the user participated as the preferred location $L_{\text{pre}}(u)$. With the preferred location, we calculate the distances to the host and the event as two features.

5 Experiments

In this section, we conduct extensive experiments on three different datasets with totally ten subsets to verify the performance of our approach in predicting event participants.

5.1 Datasets and experimental settings

Our experimental data comprise three main datasets, including Meetup.com dataset [11], Plancast.com dataset [11] and Twitter Meme dataset [24]. All of three datasets are publicly available. More detailed descriptions of each dataset are provided as follows:

– **Meetup data** Meetup is an event-based social service. Users can express their interests to offline events by sending RSVPs. In other words, each RSVP represents that a user participated in a certain event. Users and events are usually associated with the geographical information of registered locations and event places. To represent the personal interests, users can apply some tags to describe themselves. Moreover, users can join online social groups and share comments and photographs with other members in the same groups. For some events without geographical information, we adopt the locations of their event hosts as their own locations. Besides, to eliminate the location bias, we separate users and events into the subsets of seven cities, including New York City (NYC), Log Angeles (LA), San Diego (SD), San Jose (SJ), Phoenix (PHX), London (LDN), and Paris (PA), by the user locations. Note that we split this dataset into subsets by geolocation information because other users from other cities are less likely to participate events. Such splitting will lead to high AUC scores for methods relying on geolocation information. Besides, in fact, we follow the experimental setting of existing studies [17,18], which also split the datasets based on geolocation information since users tend to participate in neighboring events.
– **Plancast data** Plancast.com is another EBSN service. In Plancast, users can follow other users or claim that they will participate in some events. The following behaviors can be used to construct the online social network. For the convenience of implementation, the original directed edges in the network are treated as undirected edges. To map to the settings of other datasets, we define that users are in the same group if all of them follow a certain user. Note that this dataset has no location and tag information, so the location and semantic features are unavailable here.
– **Twitter Meme data** To some extent, the hashtags in tweets of microblogs can be treated as a kind of *diffusion events*. Hence, we adopt the Twitter Meme dataset as one of experimental data. If users add a certain hashtag in their tweets, it is equivalent that they participate in the diffusion event of that hashtag. For a hashtag, the user who first utilized that hashtag is considered as the host of the corresponding event. Moreover, users can follow other users, so we can construct the social networks as the setting in the Plancast data. Note that this dataset also contains no geographical and tag information.

Table 4 Data statistics of nine datasets

Dataset	Meetup							Plancast	Twitter
	NYC	LA	SD	SJ	PHX	LDN	PA		
# Users	16168	2846	3308	1636	1360	5890	1006	6141	3064
# Events	19558	4207	6028	2140	2828	7268	1250	5134	2180
# RSVPs	450407	80547	108797	42388	48273	159696	27866	201763	42932
# Located events	11707	2853	3195	1232	2070	3913	789	–	–
# Groups	9912	4580	2488	3124	1176	3732	699	–	–
# Users in groups	16167	2846	3308	1636	1360	5890	1006	–	–
# Tags	7702	3235	3160	2455	1956	4731	1541	–	–
# Users with tags	8645	1623	1909	951	849	3492	657	–	–
# Social links	–	–	–	–	–	–	–	278514	95440
# Users with links	–	–	–	–	–	–	–	6141	3064

Note that a RSVP indicates an entry representing a user participated in an event

For each dataset, we drop users who participate in < 10 events and events with < 10 participants. For the Twitter Meme data, the user who is the first/earliest to use the hashtag is treated as the host. Since the Meetup and Plancast datasets [11] do not provide the event host or the time that a user participates in an event, we randomly select a user as the host for each event. Note that although the randomly selected users are not the real hosts, such setting can be considered as the real-world case: Every user can serve as the event host, rather than only those users who frequently organizes social events. We then randomly select 50% of events as the training data; and the remaining events are utilized for evaluation. Finally, Table 4 shows the statistics of datasets after preprocessing. Besides, we set the latent dimension $k = 50$ (k is described in Sect. 4.1) because a larger k leads to similar performance.

5.2 Competitive methods

We compare our approach (PRMF) with the following seven competitive methods, in which the first five are simple baselines that employ feature values for the prediction, while the last two are state-of-the-art competitors.

- **Common tags (CT)** Tag information can be the useful semantic information for predicting event participants. Hence, the first compared method is the number of common tags to the event host. Users with more common tags will have high ranks.
- **Distances to host (DH) and event (DE)** The geographical information may be also useful. Therefore, there are two compared methods ranking users by the distances to the home of hosts and event locations. Users who are more likely to be closer to the host will be ranked higher.
- **Online weight (OnW)** To show the effectiveness of the online social networks, we apply the online weights [16] as one of the baseline methods. It is also one of features in our approach. The detailed formula is included in Table 3.
- **Offline weight (OffW)** This method is representative of the offline social networks. Note that we construct the offline networks with only the training events. The formula of offline weights is also listed in Table 3.
- **Bayesian personalized ranking matrix factorization (BPRMF)** BPRMF proposed in [20] is one of the state-of-the-art approaches for optimizing the ranking performance

of matrix factorization models. We treat the event hosts of users and the participants as items. We also set the latent dimension k of BPRMF as 50 for comparison.

– **Event-Centric Diffusion (ECD)** Combining online and offline weights, [11] proposed to conduct diffusion over EBSNs. Consider that the vector $\boldsymbol{v}^k$ represents the probabilities that users participate in the event after k diffusion steps, ECD can be expressed as $\boldsymbol{v}^{k+1} = \mathcal{D} \cdot \boldsymbol{v}^k$, where $\mathcal{D}$ is the transition matrix. With online and offline social networks, the transition matrix $\mathcal{D}$ can be $\mathcal{D} = \gamma \mathcal{D}^{\text{on}} + (1-\gamma) \cdot \mathcal{D}^{\text{off}}$, where $\mathcal{D}^{\text{on}}$ and $\mathcal{D}^{\text{off}}$ are constructed by normalizing online and offline weights. Note that we set γ as 0.5 in the experiments.

5.3 Evaluation metrics

With the ground truth $R(e)$, we can evaluate the quality of participant rankings by two metrics, including the *Area Under the ROC Curve* (AUC) and the *Precision at K* (P@K). AUC is an important measure to illustrate the performance of binary ranking systems. It specifies the probability that the predicted pairwise ranking is correct for two randomly sampled instances. In other words, AUC represents the pairwise accuracy between all pairs of instances. Given $R(e)$ and the predictions $\hat{r}_{hu}$, AUC can be defined as follows:

$$\text{AUC}(e) = \frac{\sum_{u_i \in R(e)} \sum_{u_j \in U - R(e)} \mathbb{1}(\hat{r}_{hu_i} > \hat{r}_{hu_j})}{|R(e)||U - R(e)|},$$

where $\mathbb{1}(\cdot)$ is the indicator function.

P@K is also an usual metric in information retrieval. It represents the performance of the top predictions instead of all instances of ranked list. To put it differently, P@K is the accuracy of instances ranked in the top-K positions. We calculate P@K from $K = 1$ to $K = 100$ for evaluating different situations.

5.4 Overall performance

Table 5 shows the performance of AUC for seven competitive methods and our PRMF approach in nine datasets. Figure 5 illustrates the performance of P@K in the NYC dataset. Comparing the competitive methods, we find that CT, DH, and DE have much bad performance. For CT, the reason is that the coverage of tags is too low to obtain information of enough users. For example, only 53% of users in the NYC have at least one tag. For DH and DE, this is because the difference between the distances of participants and non-participants may not significant enough. The methods OnW and OffW have better performances than DH and DE. These results are consistent with the data analysis in [9]. However, OnW sometimes outperforms OffW in the AUC of some datasets because user participation history sometimes is limited so that many participants will not be linked on the offline social networks. In contrast, if users were more active in the training data, such as the users in the Twitter dataset, OffW will obtain better performance. As a matrix factorization model without any demographic and semantic information, BPRMF reaches a good performance. Moreover, the results also prove that our approach based on latent factor models is reasonable. Among all compared methods, ECD is the best in both of AUC and P@K metric. This is because ECD can well propagate the probabilities over social networks and predict the participants more accurately.

For all testing datasets, our approach outperforms all of baseline methods. This is because we incorporate lots of information and the advantages of latent factor models. Figure 5 shows that the most significant improvement is the predictions at top positions. It means that our features can well gather the useful knowledge of each source information and rank the

Table 5 The performance of AUC for seven compared methods and our approach in nine datasets

Dataset	Meetup							Plancast	Twitter
	NYC	LA	SD	SJ	PHX	LDN	PA		
Common tags (CT)	0.5422	0.5204	0.5354	0.5426	0.5315	0.5362	0.5141	–	–
Distance to host (DH)	0.5254	0.6037	0.5306	0.5345	0.5591	0.5560	0.6122	–	–
Distance to event (DE)	0.5330	0.6474	0.5396	0.5514	0.5899	0.5575	0.6122	–	–
Online weight (OnW)	0.8177	0.7821	0.8131	0.8318	0.7672	0.8307	0.6556	0.7038	0.6847
Offline weight (OffW)	0.7751	0.7614	0.8052	0.7918	0.8042	0.7795	0.6801	0.6795	0.7883
BPRMF	0.7413	0.7147	0.7561	0.7308	0.7630	0.7169	0.6730	0.7316	0.7633
Event-Centric Diffusion (ECD)	0.8036	0.7955	0.8299	0.8334	0.8315	0.8401	0.7362	0.8163	0.8493
Our approach	0.8725	0.8451	0.8668	0.8501	0.8644	0.8528	0.7936	0.8274	0.8534

All improvements of our approach against the best compared method are significant differences at 99% level in a paired t test

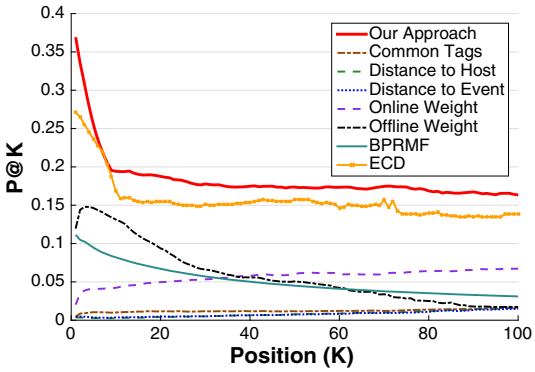


Fig. 5 The performance of P@K for seven compared methods and our approach in the NYC dataset

corresponding users into higher positions. We have also performed significant tests for the improvements of our approach against other competitive methods using a paired t test with a significant level of 99%.

5.5 Analysis of experimental results

In this section, we conduct several analyses on experimental results in different situations.

We first observe how the size of training data affect the performance of our approach. Figure 6 shows the AUC performance with different size of training data in four datasets. The results show that with more training data, the model can reach better performance. It is also consistent in all four datasets. However, we also observe that when the training data increase, the improvements from smaller events decrease. This is, if training data are sufficient, our approach can achieve a good and stable performance without more training events.

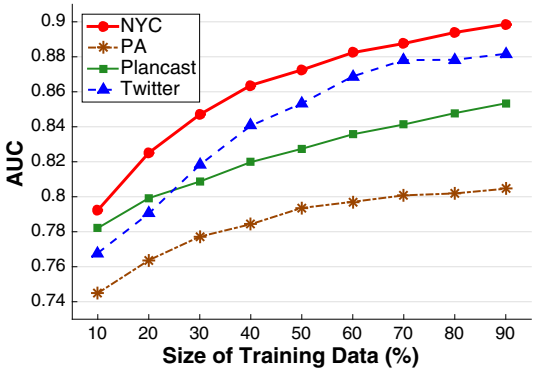


Fig. 6 The AUC performance of our approach with different sizes of training data in four datasets

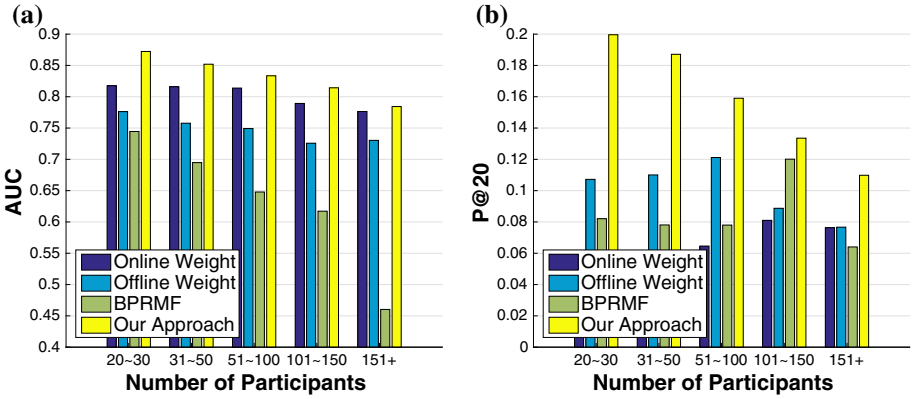


Fig. 7 The performance for different event sizes in the NYC dataset. **a** AUC and **b** P@20

We then analyze the effect of the event size, which is defined as the number of participants of an event. Figure 7 illustrates the performance of three baseline methods and our approach with different event sizes in the NYC dataset. In each size of events, our approach still outperforms other baselines. With smaller events, our approach can obtain better performance because the numbers of participants covered by our features are similar in events with different sizes. Therefore, other participants that cannot be ranked higher will reduce the overall performance.

The activeness of the event host may also affect the performance. Figure 8 shows the performance of three competitive methods and our approach with different host status of activeness in the NYC dataset. We simply define that a host is active if that user participated in more than nine events, which is the median number of event participation in the training data. It is obvious that all methods have better performance of both metrics with the hosts who are more active. The results are also reasonable because the active hosts may have much more records in the training so that the models can easily discover helpful information for event participant prediction.

Last but not least, different tags may lead to difference of performance. We simply define the tag popularity of an event as the average popularity of the event host's tags. Then, the tag popularity of the event e can be simply expressed as follows:

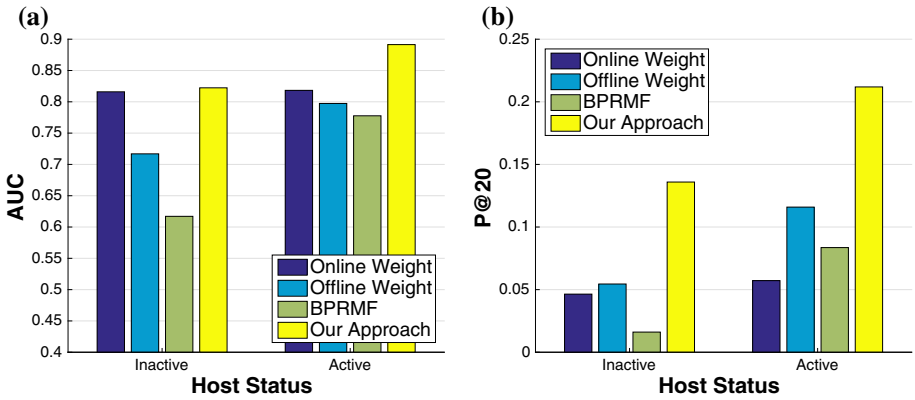


Fig. 8 The performance for different host status in the NYC dataset. Note that each active user participated in more than nine events, which is the median number of user event participation in the training data. **a** AUC and **b** P@20

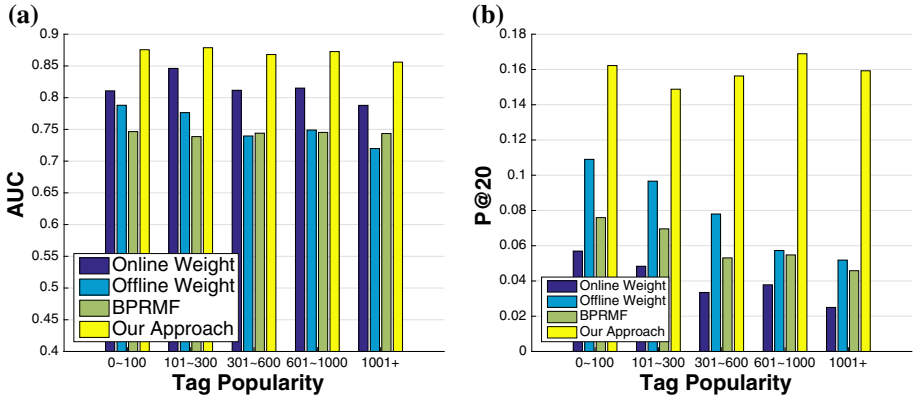


Fig. 9 Performance for different tag popularity of events in the NYC dataset. **a** AUC and **b** P@20

$$\frac{1}{|T(H(e))|}\sum_{t\in T(H(e))}\text{pop}(t),$$

where pop(t) means the number of users that have the tag t in their profiles. Figure 9 exhibits the performance of three competitive methods and our approach with different tag popularity values of events in the NYC dataset. The trends of BPRMF and OffW methods, which do not utilize the tag information, indicate in that dataset events with higher tag popularity are more tough. However, both of OnW and our approach can keep similar performance. It means that methods incorporating the tag information can achieve better performance in events with higher tag popularity.

5.6 Analysis of feature effectiveness

To verify the effectiveness of the features, in this section, we conduct the leave-one-out evaluation to analyze six categories of features. Figure 10 shows the performance loss of each features set in leave-one-out evaluation. All feature sets have positive performance loss.

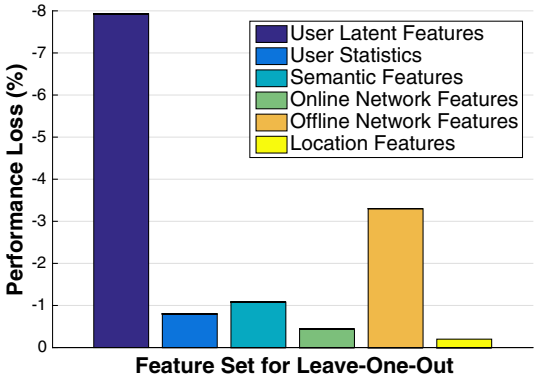


Fig. 10 Performance loss of each feature set in leave-one-out feature evaluation

Hence, all of them are useful in our approach. The most significant feature is the category of user latent features. It is reasonable because the performance of BPRMF has already proven that latent factors are powerful in this task. Among extracted features, the category of offline network features outperforms other features. This result is also consistent with the data analysis in [9]. The location features obtain the least performance loss. The reason is as same as the low performance of DE and DH methods as shown in Table 5 and Fig. 5.

6 Discussions and conclusions

We aim to predict event participants for a single host. The contribution is fourfold. First, we formulate the host-aware event participants prediction, which is challenging due to data sparsity and cold-start problems. Second, based on some data analysis, we devise six diverse sets of features, which are verified to be effective in modeling the preferences and willingness of users to attend events. Third, we develop a novel feature-based Participant-Ranking Matrix Factorization (PRMF) model, which demonstrates the satisfying performance. Fourth, extensive analyses unveil how event size, host status, and topic popularity affect the prediction quality.

The experiment also delivers some findings. (a) Participants of larger events are more difficult to be predicted. (b) Those actors actively organizing events will lead to better performance of predicting participants. (c) If event hosts are associated with popular tags, the corresponding participants are more difficult to predict. (d) Among all of the defined features, the set of user latent features is the most influential in prediction performance. This also exhibits the significant usefulness of our feature-based factorization method in learning user features. In other words, our model can discover effective implicit features (i.e., latent features), rather than explicit features used in state-of-the-art methods.

This study is the first attempt to explore the host-aware event participant prediction, and the results encourage future studies to develop a more accurate participant recommender for event hosts. Nevertheless, our problem setting and solution have two major limitations and drawbacks. One is allowing only one host. However, it is often that real-world events have multiple organizers. The other is about the sequential effect of participants. Participants are sequentially added into the event. However, we recommend the participants at one time. Sequential prediction of event participants is the main target of our future work. The other potential future work is modeling the spreading effect of event invitation. The propagation

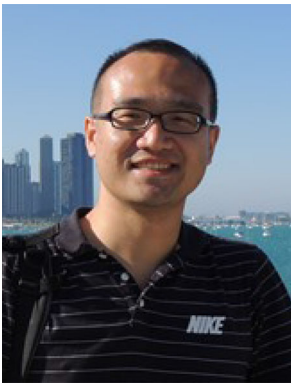
and exposure of invitations among users are believed to be an important clue for participant prediction.

Acknowledgements This work was sponsored by Ministry of Science and Technology (MOST) of Taiwan under Grants 104-2221-E-006-272-MY2, 106-2118-M-006-010-MY2, 106-2628-E-006-005-MY3, 106-3114-E-006-002, and 107-2636-E-006-002, and also by Academia Sinica under Grant AS-107-TP-A05.

References

1. Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: COMPSTAT '10. pp 177–186
2. Chen J, Geyer W, Dugan C, Muller M, Guy I (2009) Make new friends, but keep the old: recommending people on social networking sites. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM. pp 201–210
3. Du R, Yu Z, Mei T, Wang Z, Wang Z, Guo B (2014) Predicting activity attendance in event-based social networks: content, context and social influence. In: UbiComp '14. pp 425–434
4. Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. CRC Press, Boca Raton
5. Georgiev P, Noulas A, Mascolo C (2014) The call of the crowd: event participation in location-based social services. In: ICWSM '14. pp 341–350
6. Guy I, Zwerdling N, Ronen I, Carmel D, Uziel E (2010) Social media recommendation based on people and tags. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, ACM. pp 194–201
7. Hu Y, Farnham S, Talamadupula K (2015) Predicting user engagement on twitter with real-world events. In: ICWSM '15. pp 168–177
8. Jamali M, Ester M (2010) A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the fourth ACM conference on Recommender systems, ACM. pp 135–142
9. Jiang J-Y, Li C-T (2016) Analyzing social event participants for a single organizer. In: ICWSM '16
10. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. IEEE Comput 8:30–37
11. Liu X, He Q, Tian Y, Lee W-C, McPherson J, Han J (2012) Event-based social networks: linking the online and offline social worlds. In: KDD '12. pp 1032–1040
12. Ma H, King I, Lyu MR (2009) Learning to recommend with social trust ensemble. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, SIGIR '09. pp 203–210
13. Ma H, Zhou D, Liu C, Lyu MR, King I (2011) Recommender systems with social regularization. In: Proceedings of the fourth ACM international conference on web search and data mining, WSDM '11. pp 287–296
14. Macedo AQ, Marinho LB, Santos RL (2015) Context-aware event recommendation in event-based social networks. In: RecSys '15. pp 123–130
15. Moritz H (1980) Geodetic reference system 1980. J Geod 54(3):395–405
16. Newman ME (2001) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys Rev E 64(1):016132
17. Pham T-AN, Li X, Cong G, Zhang Z (2015) A general graph-based model for recommendation in event-based social networks. In: ICDE '15. pp 567–578
18. Qiao Z, Zhang P, Zhou C, Cao Y, Guo L, Zhang Y (2014) Event recommendation in event-based social networks. In: AAAI international conference on artificial intelligence, pp 567–578
19. Rendle S (2010) Factorization machines. In: ICDM '10. pp 995–1000
20. Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L (2009) Bpr: Bayesian personalized ranking from implicit feedback. In: UAI '09. pp 452–461
21. Saez-Trumper D, Quercia D, Crowcroft J (2012) Ads and the city: considering geographic distance goes a long way. In: Proceedings of the sixth ACM conference on recommender systems, RecSys '12. pp 187–194
22. Scellato S, Noulas A, Lambiotte R, Mascolo C (2011) Socio-spatial properties of online location-based social networks. ICWSM 11:329–336
23. Tu W, Cheung DW-L, Mamoulis N, Yang M, Lu Z (2015) Activity-partner recommendation. In: PAKDD '15. pp. 591–604

24. Weng L, Menczer F, Ahn Y-Y (2013) Virality prediction and community structure in social networks. Sci Rep 3:2522
25. Xu T, Zhong H, Zhu H, Xiong H, Chen E, Liu G (2015) Exploring the impact of dynamic mutual influence on social event participation. In: SDM '15. pp 262–270
26. Yu Z, Du R, Guo B, Xu H, Gu T, Wang Z, Zhang D (2015) Who should i invite for my party?: Combining user preference and influence maximization for social events. In: UbiComp '15. pp 879–883
27. Zhang W, Wang J, Feng W (2013) Combining latent factor model with location features for event-based group recommendation. In: KDD '13. pp 910–918
28. Zhang X, Zhao J, Cao G (2015) Who will attend? - predicting event attendance in event-based social network. In: MDM '15. pp 74–83



Cheng-Te Li is currently an Assistant Professor with Institute of Data Science and Department of Statistics, National Cheng Kung University, Tainan, Taiwan. He received the M.S. and Ph.D. degree from Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan, in 2009 and 2013, respectively. He was an Assistant Research Fellow at Research Center for Information Technology Innovation (2014-2016) in Academia Sinica, Tainan, Taiwan. His research interests include social and information networks, data mining, machine learning, and social media analytics. He was a recipient of the Facebook Fellowship 2012 Finalist Award, the ACM KDD Cup 2012 First Prize, the IEEE/ACM ASONAM 2011 Best Paper Award and the Microsoft Research Asia Fellowship 2010.

Jyun-Yu Jiang is a Ph.D. student in Department of Computer Science at University of California, Los Angeles. Before that, he received his M.S. and B.S. degree from Department of Computer Science and Information Engineering at National Taiwan University in 2015 and 2013. His research interests include Information Retrieval, Natural Language Processing, Data Mining and Machine Learning. More specifically, he is interested in developing effective machine learning techniques and efficient algorithms to solve real-world problems. In the Bachelor stage, he participated in the ACM International Collegiate Programming Contests (ACM-ICPC) as a contestant. He also participated in KDDCUP 2012 and won the champion with NTU team.