Who Should I Invite for My Party? Combining User Preference and Influence Maximization for Social Events

Zhiwen Yu[†], Rong Du[†], Bin Guo[†], Huang Xu[†], Tao Gu[‡], Zhu Wang[†], Daqing Zhang[♭]
[†] Northwestern Polytechnical University, Xi'an 710129, Shaanxi, China
[‡] RMIT University, Melbourne, Australia
[♭] Institut Mines-TELECOM/TELECOM SudParis, France
zhiwenyu@nwpu.edu.cn

ABSTRACT
The newly emerging event-based social networks (EBSNs) extend social interaction from online to offline, providing an appealing platform for people to organize and participate real-world social events. In this paper, we investigate how to select potential participants in EBSNs from an event host's point of view. We formulate the problem as mining influential and preferable invitee set, considering from two complementary aspects. The first aspect concerns users' preference with respect to the event. The second aspect is influence maximization, which aims to influence the largest number of users to participate the event. In particular, we propose a novel Credit Distribution-User Influence Preference (CD-UIP) algorithm to find the most influential and preferable followers as the invitees. We collect a real-world dataset from a popular EBSNs called "Douban Events", and the experimental results on the dataset demonstrate the proposed algorithm outperforms the state-of-the-art prediction methods.

Author Keywords
event-based social networks; user preference; influence maximization; invitee set; CD-UIP.

ACM Classification Keywords
H.3.5 Online Information Services: Web-based services

General Terms
Algorithms, Experimentation, Performance

INTRODUCTION AND RELATED WORK
The popularity of Event Based Social Networks (EBSNs) such as *Facebook Events*[1], *Meetup*[2], and *Douban Events*[3], has created increasing opportunities for people to expand their online interactions to physical interactions and participate in real-world events. Leveraging on these platforms, users now can organize and participate a variety of real-world social events (e.g., watching movie together, hosting a party, and organizing a group travel) for socialization. Much work has been done in EBSNs, and they mainly focus on analyzing the characteristics of EBSNs when comparing to traditional social networks, such as network properties, community structures, and information flow [1, 2, 3]. There also have been studies in building prediction and recommendation systems for various events from a participant's perspective [1, 4].

Our previous work presents an algorithm to predict activity attendance in EBSNs from a users point of view [5]. In this paper, we are interested in how EBSNs facilitate organizing real-world social events from an event host's perspective. When organizing a social event, a host can simply invite all her/his followers as potential participants, leveraging on EBSNs. However, not all the followers have the same degree of interest in that event. If a follower continues receiving uninterested event invitations, it may result in disconnecting the follower's relationship with the host. Hence, it is important to take user preference and social influence into account when inviting followers to participate events. So we propose a novel approach to invite participants by discovering activity fans with strong social influence from large-scale online social networks. We formulate this problem formally as mining Influential Preferable Set (IPS). There have been a few studies on team or group formation [6, 7, 8]. Though, these works either focused on contact grouping in online communities or context-aware grouping in the physical world. None of them consider about the IPS problem. The usage of data that connects online/offline spaces for group formation is also not studied.

To address IPS problem, we design a novel algorithm which aims to find the most influential and preferable set of followers when organizing social events. It extends the credit distribution (CD) model [9] by combining both user preference and influence maximization, and thus named as CD-UIP. In CD-UIP, we introduce a method to assign influence credit, in which followers will give influencers the credits for influencing them. We integrate both user influence and user preference to achieve better assignment. To evaluate CD-UIP, we crawled a large dataset from the Douban Events and conduct extensive experiments to evaluate the possibility of attendance and an invitee set's influence spread. In summary, the contributions of this paper are as follows:

[1] www.facebook.com/events
[2] www.meetup.com
[3] beijing.douban.com

UbiComp '15, September 7-11, 2015, Osaka, Japan.

http://dx.doi.org/10.1145/2750858.2805839

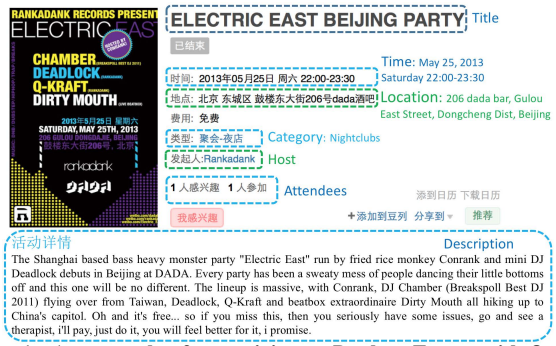

Figure 1. An example of an activity on Douban Events with five key elements: location, time, attendees, host, and content.

- To the best of our knowledge, this is the first work to discover the most influential and preferable set of followers in EBSNs as invitees participating in real-world events.
- We formulate the problem as the IPS mining problem, and develop a novel algorithm which combines user preference and influence maximization leveraging on the basic credit distribution model.
- By applying the proposed algorithm to a real-world dataset, we demonstrate its effectiveness in discovering the most influential and preferable invitee set. We compare our algorithm with the state-of-the-art approaches [10], and the results show CD-UIP achieves better performance.

PROBLEM STATEMENT AND SYSTEM OVERVIEW

We first give a brief description of the Douban dataset we used. The problem and an overview of our system is then presented.

The Douban dataset

The dataset is collected from Douban Events, which is a popular event-based social network in China. It allows people to advertise, search and participate in real-world events. Fig. 1 shows the main elements of an activity in Douban Events. A host user can post and share an activity in Douban Events. The details of a posted activity include its time, address, host name, and content. Other people can view the shared activity and the users who have already expressed their interest to attend the activity (e.g., existing attendees), based on which they can decide whether to participate in this activity. In Douban Events, users could follow each other and see the activities that their followers attend. On the other hand, users could influence their followers by recommending or showing the activities they attend.

Problem Statement

Assuming there is a new activity a. $H(a)$ represents the host and $F(a)$ denotes the followers of $H(a)$. The social graph related to a is denoted as $G(a) = (U, E)$, where U corresponding to users and directed edges E means social ties between users. The event attendance log is $L(User, Activity, Time)$, where a tuple $(u, a, t) \in L$ indicates that user u attended activity a at time t. To solve the IPS problem, we first consider the u's preference to activity a, $Pref_u(a)$, which can be extracted from u' historical attendance behavior (denoted by A_u, and $A_u = \{a_1, a_2, ...\}$) based on content, context, and social relationship. Then, we consider social influence in

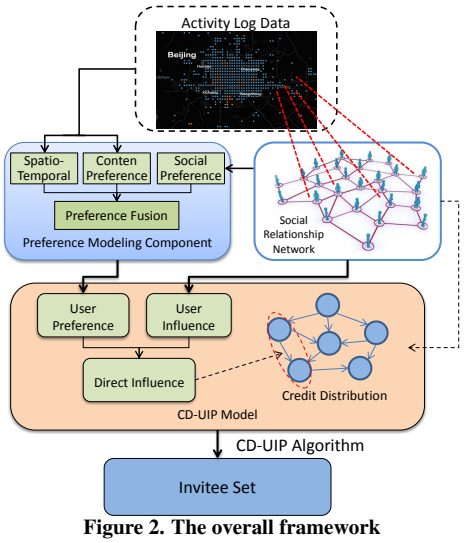

Figure 2. The overall framework

$F(a)$, which is related to the notion of *propagation*. A propagation exists from user u to v if they are socially linked, and u attends a before v. $N_{in}(u, a) = \{v|(v, u) \in E(a)\}$ denotes the set of potential influencers of u for participating a. So our problem is to select preferable and influential followers as the invitee set IS to maximize the expected number of participants, which is denoted by $\sigma_m(S)$.

System Overview

As Fig.2 shows, our framework has three components: the user preference component (left), the social influence component (right), and the CD-UIP model component (bottom).

Preference Modeling Component. We first extract features towards each activity from three aspects: content preference, spatio-temporal context, and social relationship with the host. We then propose a multi-factor (MF) model using logistic regression to evaluate the contribution of each feature. Based on the result, we can compute the similarity between each pair of events and get the complete preference of the target user to a specific event.

Social Influence Component. The second component models the followers' social influence, and it is the key component in our framework. This is related to influence maximization which aims for attracting a large number of participants. To achieve this, we introduce the influence credit distribution model, in which each follower gives an influence credit to the influencer for influencing them to participate in the activity.

CD-UIP Model. To incorporate user preference and social influence into our system, we propose the CD-UIP in the invitee set selection component. Specifically, we reassign the direct credit in the CD model from four main aspects: 1) influence decays over time in an exponential fashion; 2)some users are more influenceable than others; 3) users' preferences to the activity; 4) the number of potential influencers.

COMPUTING USER PREFERENCE

The previous work has proposed SVD-MFN to better predict a target user's attendance [5]. It is based on the extraction and usage of three macro features from Douban. (1) The *content*

preference feature describes user preference to a Douban activity through its category, title, and description. The Latent Dirichlet Allocation (LDA) method is used for content similarity measurement. (2) The *spatio-temporal* feature characterizes the spatial proximity and temporal context similarity of activities. (3) The *social relationship* feature defines two types of social relationships between a user and the host, including the following relationship and preferring relationship (how many times a user has attended the host's events). The three features are fused as $Sim(a, a_i)$ with a sum of their weighted values, which describes the similarity between an upcoming activity a and a past event a_i for target user u. Finally, the users preference value to an event can be formulated as: $Pref_u(a) = \sum_{a_i \in A_u} Sim_u(a, a_i)/|A_u|$.

INFLUENCE MAXIMIZATION

The influence maximization problem aims to find a set of users which maximize the expected spread of influence in the social network. Previous studies mainly focused on two propagation models: the Independent Cascade (IC) model and the Linear Threshold (LT) model. However, these two models may not work well in our system. First, the time complexity caused by the edge learning and Monte Carlo (MC) simulation can be quite high. Second, the accuracy could be very low as the existing models did not fully utilize a user's history attendance logs. So we use both the IC and LT models as the baseline for our comparison study.

To overcome the shortcomings, Goyal et al. [9] proposed a totally different model to directly mine the available log of past attendance propagation traces. The basic idea is: When a user v attends an activity a before u, the direct influence credit $\gamma_{v,u}(a)$, means the influence credit given to v for influencing u to participate in activity a. Influence credit also distribute transitively backwards in the propagation graph $G(a)$ such that not only u gives credit to the users $v \in N_{in}(u, a)$, but also they in turn pass on the credit to their predecessors in $G(a)$ and so on. The related definitions are as follows:

- The total credit given to v for influencing u on activity a :
$$\Gamma_{v,u}(a) = \sum_{w \in N_{in}(u,a)} \Gamma_{v,w}(a) \cdot \gamma_{w,u}(a) \quad (1)$$
- The total credit given to a set of nodes $S \subseteq U(a)$ for influencing user u on activity a :
$$\Gamma_{S,u}(a) = \begin{cases} 1 & if \quad v \in S; \\ \sum_{w \in N_{in}(u,a)} \Gamma_{S,w}(a) \cdot \gamma_{w,u}(a) & otherwise \end{cases} \quad (2)$$
- The total influence credit given to v by u for all activities in A_u, which obtained by taking the total credit over all the activities and normalizing it by the number of A_u:
$$\kappa_{\mathrm{v,u}} = \frac{1}{\mathrm{A_u}} \sum_{a \in A_u} \Gamma_{v,u}(a) \quad (3)$$
- The total influence credit of $S \subseteq U$ for activities in A_u:
$$\kappa_{\mathrm{S,u}} = \frac{1}{\mathrm{A_u}} \sum_{a \in A_u} \Gamma_{S,u}(a) \quad (4)$$
- The influence spread $\sigma_{cd}(S)$ as the total influence credit given to S from the entire social network:
$$\sigma_{cd}(S) = \sum_{u \in U} \kappa_{S,u} \quad (5)$$

THE CD-UIP ALGORITHM

To combine user preference and influence maximization, we reassign the direct credit $\gamma_{v,u}(a)$ from four aspects: a) similar to human's memory, influence decays over time; b) each person's influenceability is different due to their personality; c) users' preferences to the activity has a strong influence on users' attendance possibility; d) the number of potential influencers, which means the more potential influencers u has, the less v influences u. Motivated by these ideas, we assign direct credit as Eq. 6.

$$\gamma_{v,u}(a) = \frac{infl(u)}{N_{in}(u,a)} \cdot \exp\left(-\frac{t(u,a)-t(v,a)}{\tau_{v,u}(a)}\right) \cdot Pref_u(a) \cdot Pref_v(a) \quad (6)$$

Here, $\tau_{v,u}(a)$ is the average time taken for activity a to propagate from v to u. The exponential term in the equation achieves the desired effect that influence decays over time. Moreover, $infl(u)$ denotes the user's influenceability, that is, how prone user u can be influenced by the social context [11]. Precisely, $infl(u)$ is defined as a fraction of activities that u has attended under the influence of at least one of its neighbors; The value of $infl(u)$ is normalized by $N_{in}(u, a)$ to ensure that the sum of direct credits assigned to neighbors of u for activity a is at most 1. Note that both $infl(u)$ and $\tau_{v,u}(a)$ are learnt from the training subset of L.

Algorithm 1 CD-UIP : Combine User Influence and Preference based on the CD model.

Input: G; $k\%$
Output: Invitee Set: $IS(a)$ for each $a \in A$
1: SC $\leftarrow \emptyset$; $IS(a) \leftarrow \emptyset$, $Q \leftarrow \emptyset$
2: **for** each $v \in Parents(u)$ **do**
3: compute $\gamma_{v,u}(a)$; $UC[v][u][a] \leftarrow UC[v][u][a] + \gamma_{v,u}(a)$
4: **end for**
5: **for** each $a \in A$ **do**
6: **for** each $u \in U$ **do**
7: $x.mg \leftarrow x.mg + UC[x][u][a]/|A_u|$; $x.it \leftarrow 0$; add x to Q
8: **while** $|IS(a)| < k\% \times |F(a)|$ **do**
9: $x \leftarrow pop(Q)$
10: **if** $x.it = |IS(a)|$ **then**
11: $IS(a) \leftarrow IS(a) \cup \{x\}$; $UC[v][u][a] \leftarrow UC[v][u][a] - UC[v][x][a] \times UC[x][u][a]$; $SC[u][a] \leftarrow SC[u][a] + UC[x][u][a] \times (1 - SC[x][a])$;
12: **else**
13: $x.mg \leftarrow x.mg + x.mg \times (1 - SC[x][a])$;
14: $x.it \leftarrow |IS(a)|$; Reinsert x into Q and heapify;
15: **end if**
16: **end while**
17: **end for**
18: **end for**

We record $\Gamma_{v,u}(a)$ into the data structure UC (User Credits). Another data structure is SC (Set Credits), where each $SC[x][a]$ refers to the total credit given to the current invitee set IS by a user x for an activity a, that is, $\Gamma_{\mathrm{IS,x}}(a)$. In total, we apply the CD-UIP to select the invitee seed $IS(a)$ by Algorithm 1. The size of $IS(a)$ is $k\%$ of $F(a)$. After obtaining $\gamma_{v,u}$, and UC(lines 2,3), we use greedy algorithm. The algorithm also maintains a queue Q, where the entry of a user x is stored in the form $< x, mg, it >$ and mg represents the marginal gain x and it represents the count of iteration(line 7). In each iteration, the top element x of Q is analyzed (lines 8 and 9). If x is analyzed in the current iteration, it will be the next seed user and the data structure UC and SC is updated. If $x.it < |IS(a)|$, the marginal gain of x will be recomputed(line 13). Then $x.it$ is reset and x is re-inserted into Q (line 14).

EXPERIMENTAL RESULTS

Dataset Statistics

We use the Douban Events APIs to collect all the valid activities for 21 months (i.e., from January 2012 to October 2013). Table 1 gives a brief summary about the dataset. Just as described in the second section, we divide the data into three types: the attendance history log, the details of activities, and the social information. The attendance history log is represented as User-Activity pairs in the table. Each activity contains various information including topic, title, time, address, host, etc. We obtain the following relationship with hosts as well as the relationship between users. The reason is that users may be influenced by hosts' invitation and their followers' recommendation.

Time Interval	2012-01-01 to 2013-10-01
Number of Hosts	430
Number of Activities	2,191
Number of Users	4,600
Number of Following Relationship with Hosts	11,831
Number of Following Relationship between Users	142,487
Number of User-Activity Pairs	90,892

Table 1. Statistics of the Douban Events dataset

Accuracy of Spread Prediction

The objective of our work is to find the most influential invitee set. It is similar to the goal of viral marketing [12]. To trigger a large cascade of adoption of a new product or innovation, the viral marketing approach tries to first identify a subset of most influenced people to adopt it [12]. We assume that if we could accurately predict the influence spread of a given seed set, then we can identify the most influential user set according to the result of prediction. For example, we could run our algorithm over the whole user set and make a top-k selection for the best user set. Therefore, the experiment purpose is transferred to test whether the spread prediction method is effective. To determine which model (i.e., LT, IC, CD and CD-UIP) achieves the best accuracy in predicting the expected spread of node sets, we compute their influence spread. For a given seed dataset S, we compute the expected spread $\sigma(S)$ predicted by each method and compare it with the actual spread of S according to the ground truth. The actual spread is the number of users who attend the activity. Fig.3 reports the root mean squared error (RMSE) between the predicted and actual spread for the three algorithms respectively. An interesting observation is that IC beats LT when the actual spread is large, and LT performs better when the actual spread is small. This is probably due to the fact that both IC and LT models always tend to predict the spread as very high [11]. The CD and CD-UIP models have similar performance and outperform IC and LT for the whole dataset. Because we determine the direct influence credit by combining user influence and preference in CD-UIP model, it reaches a lower RMSE than CD model, especially when the actual spread larger than 100.

Influence Spread

For an activity host, the best algorithm is the one influences the most number of participants (i.e., per activity influence). So we use user preference (UP) and user influence (UI) as the baseline to compare with CD-UIP algorithm (i.e., taking into account both user preference and influence) on influence of

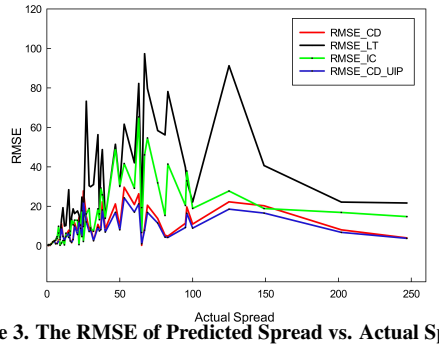

Figure 3. The RMSE of Predicted Spread vs. Actual Spread.

invitee sets. The influence is calculated as the actual number of participants influenced by the invitee set as shown in Fig.4. It also shows the performance per participant invitee and per invitee where per participant invitee represents the average influence of participants in the invitee set. They all means the influence of invitee we select, but invitee in the former are actually attend activities. Obviously, the UI algorithm performs the best for influence spread, and our CD-UIP algorithm comes next. The UP algorithm performs poorly as it does not consider social influence at all. Per invitee represents the average influence of the whole invitee set. The UI algorithm becomes worse even than UP, and CD-UIP performs the best (e.g., each invitee influences almost 10 participants). The most important metric is the total influence set for each activity (i.e., per activity). It represents how many users influenced by the invitees become participants. From the figure, we observe that there are 35.95 participants on average who join the activity, compare to 11.24 by UI, and 15.24 by UP. Overall, our CD-UIP algorithm outperforms the other two algorithms.

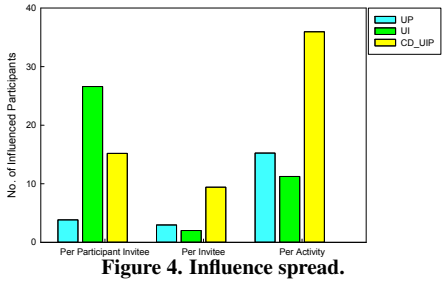

Figure 4. Influence spread.

CONCLUSION AND FUTURE WORK

This paper studies the problem of identifying the most influential and preferable set of invitees in EBSNs. Our work intends to lay a foundation for understanding emerging EBSNs and providing key insights to facilitate personalized event recommendation, marketing and targeted advertising. For our future work, we plan to build a comprehensive recommendation system for EBSNs, which is able to suggest appropriate events to users and recommend suitable invitee sets to the host.

ACKNOWLEDGMENTS

This work was supported in part by the National Basic Research Program of China (No. 2015CB352400), the National Natural Science Foundation of China (No. 61222209, 61373119, 61332005), and the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20126102110043).

REFERENCES

1. X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han, "Event-based social networks: linking the online and offline social worlds," in *Knowledge Discovery and Data Mining*, 2012, pp. 1032–1040.
2. J. Han, J. Niu, A. Chin, W. Wang, C. Tong, and X. Wang, "How online social network affects offline events: A case study on douban," in *Ubiquitous Intelligence & Computing and International Conference on Autonomic & Trusted Computing*, 2012, pp. 752–757.
3. B. Xu, A. Chin, and D. Cosley, "On how event size and interactivity affect social networks," in *CHI Extended Abstracts on Human Factors in Computing Systems*, 2013, pp. 865–870.
4. E. Minkov, B. Charrow, J. Ledlie, S. J. Teller, and T. Jaakkola, "Collaborative future event recommendation," in *International Conference on Information and Knowledge Management*, 2010, pp. 819–828.
5. R. Du, Z. Yu, T. Mei, Z. Wang, Z. Wang, and 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.
6. D. MacLean, S. Hangal, S. K. Teh, M. S. Lam, and J. Heer, "Groups without tears: mining social topologies from email," in *Proceedings of the 16th international conference on Intelligent user interfaces*. ACM, 2011, pp. 83–92.
7. E. G. Boix, A. L. Carreton, C. Scholliers, T. Van Cutsem, W. De Meuter, and T. D'Hondt, "Flocks: enabling dynamic group interactions in mobile social networking applications," in *Proceedings of the 2011 ACM Symposium on Applied Computing*. ACM, 2011, pp. 425–432.
8. B. Guo, Z. Yu, D. Zhang, H. He, J. Tian, and X. Zhou, "Toward a group-aware smartphone sensing system," *Pervasive Computing, IEEE*, vol. 13, no. 4, pp. 80–88, 2014.
9. A. Goyal, F. Bonchi, and L. V. Lakshmanan, "A data-based approach to social influence maximization," *Proceedings of the VLDB Endowment*, vol. 5, no. 1, pp. 73–84, 2011.
10. D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.
11. A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 241–250.
12. M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 61–70.