Information Sciences 422 (2018) 432–446

Contents lists available at ScienceDirect

Information Sciences

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

# Forecasting participants of information diffusion on social networks with its applications

Cheng-Te Li[a], Yu-Jen Lin[b], Mi-Yen Yeh[b,*]

[a] *Department of Statistics, National Cheng Kung University, Tainan, Taiwan*
[b] *Institute of Information Science, Academia Sinica, Taipei, Taiwan*

ARTICLE INFO

*Article history:*
Received 11 September 2016
Revised 30 May 2017
Accepted 11 September 2017
Available online 12 September 2017

*Keywords:*
Diffusion forecasting
Information diffusion
Social network
Diffusion participation
Social influence
Popularity prediction

ABSTRACT

Social networking services allow users to adopt and spread information via diffusion actions, e.g., share, retweet, and reply. Real applications such as viral marketing and trending topic detection rely on information diffusion. Given past items with diffusion records on a social network, this paper aims at forecasting who will participate in the diffusion of a new item *c* (we use hashtags in the paper) with its *k* earliest adopters, without using content and profile information, i.e., finding which users will adopt *c* in the future. We define the Diffusion Participation Forecasting (DPF) problem, which is challenging since all users except for early adopters can be the candidates, comparing to existing studies that predict which one-layer followers will adopt a new hashtag given past diffusion observations with content and profile info. To solve the DPF problem, we propose an Adoption-based Participation Ranking (APR) model, which aims to rank the actual participants in reality at higher positions. The first is to estimate the adoption probability of a new hashtag for each user while the second is a random walk-based model that incorporates nodes with higher adoption probability values and early adopters to generate the forecasted participants. Experiments conducted on Twitter exhibit that our model can significantly outperform several competing methods in terms of Precision and Recall. Moreover, we demonstrate that an accurate DPF can be applied for effective targeted marketing using influence maximization and boosting the accuracy of popularity prediction in social media.

## 1. Introduction

Information diffusion is the major production of social interactions among users in online social networking services such as Twitter and Facebook. With various diffusion actions, such as retweet, share, reply and rating, different items (e.g., hashtags, short texts, images, and URLs) can be propagated from one user to another, which we call *diffusion participants*, in a social network. Understanding information diffusion from online social data can enable many real-world applications, such as viral marketing [12], trending topic detection [32], and information summarization [26]. Predicting the future participants who will spread a certain item is one of the most essential steps to uncover the mechanism of information diffusion. Some of existing studies predict whether or not a user will adopt the same item that his/her followees had adopted earlier [8,21,36] given the item content and/or user profiles. However, we think it is insufficient to find the diffusion participants

[*] Corresponding author.
*E-mail addresses:* chengte@mail.ncku.edu.tw, reliefli@gmail.com (C.-T. Li), maxwellbiga@gmail.com (Y.-J. Lin), miyen@iis.sinica.edu.tw (M.-Y. Yeh).

http://dx.doi.org/10.1016/j.ins.2017.09.034
0020-0255/© 2017 Elsevier Inc. All rights reserved.

from only the followers of the *early adopters* who are the first users participating in the diffusion. In addition, in real scenarios, it is unrealistic to assume that the item content and user profiles are always accessible due to the privacy issue and the storage constraint. Therefore, we propose to forecast the diffusion participants of an item among followers and followers of followers (i.e., multi-hop away) from early adopters without using item content and user profile information.

In this work, using Twitter as an example, we study the diffusion participation of hashtags among users. Note that participating in the diffusion of a hashtag is equivalent to adopting such hashtag. We will use *participants* and *adopters* interchangeably throughout this paper. Given a newly emerging hashtag in a social network along with its early adopters, we aim at forecasting the future diffusion participants of such hashtag among users who can be directly or indirectly connected with early adopters, using no textual content and user profiles. The diffusion participation forecasting problem is challenging due to the following three reasons. The first is *data sparsity*. From the retweeting records of users in Twitter, we find that most users participate in very limited diffusions, i.e., they adopt very few hashtags. Only a small set of users frequently and actively participate in various kinds of diffusions. Such limited diffusion records lead to severe data sparsity, and thus make us difficult to learn the behaviors of diffusion participation for most users. The second is forecasting the participation of multi-hop followers from the early adopters. Some of existing studies [2,25] divide the diffusion records of a certain hashtag into training and testing sets, use the training set to learn the predictive model, and predict users' adoption on the testing set for the same hashtag. Other studies [3,36] train the predictive models using the diffusion records of past hashtags, and predict the adoption of a new hashtag for "one-hop" followers of early adopters. Our goal is to learn the behaviors of diffusion participation from users' adoption of past hashtags, and forecast the users' adoption for a new hashtag. It is apparently difficult since we cannot directly learn the users' habits on the new hashtag. The third is forecasting without using item content or user profiles. We assume the content information (e.g., texts and images) is not always accessible. Therefore, the users' behaviors of diffusion participation can be learned based on only historical (e.g., the participation of past diffusions) and social information (e.g., social connections, and temporal clues). Such setting is much challenging compared with past models [8,37] that highly depend on content and profile info.

We develop an Adoption-based Participation Ranking (APR) model to find the diffusion participants based on the early adopters. The technical goal is to obtain a *participation probability* for each user, and use such probability to generate a ranking list of users such that the actual participants can be placed at highest positions. Our main idea is that a user $v$ has higher probability to adopt hashtag $c$ if (a) $v$ prefers hashtag $c$ or has higher willingness to adopt $c$, and (b) the diffusion of hashtag $c$ can easily *reach* $v$. For (a), we construct a predictive model to estimate an *adoption probability* for each user $v$ that models the preference or willingness of $v$ to adopt hashtag $c$. To have an effective predictive model, six categories of features are proposed to characterize the hidden correlation between the early adopters and each target user. For (b), we devise a random-walk ranking model to generate a *participation probability* for each target $v$, based on both the derived adoption probability of $v$ and the early adopters. Users with higher participation probability values are considered as the most potential participants who will adopt $c$.

Experiments conducted on a Twitter dataset disclose three insights. First, the proposed APR model can significantly outperform other competing methods in terms of Precision, Recall, and F1-measure. Second, the estimation of users' adoption probabilities, which model the behaviors and willingness of diffusion participation, plays a significant role for having accurate forecasting results. Third, jointly using early adopters and other users with higher estimated adoption probability values can lead to satisfying performance. Furthermore, we aim at applying the forecasted diffusion participants to the tasks of influence maximization and popularity prediction. The applications' evaluation exhibits two promising results. First, considering the forecasted diffusion participants as the seed candidates of influence propagation can lead to effective targeted marketing. This is achieved by identifying influential seeds and maximizing the influence spread with respect to the given hashtag. Second, treating the forecasted participants as additional adopters of information diffusion can significantly boost the performance of tweet popularity prediction. With such results, the power of the proposed diffusion participation forecasting is further proven.

Here we summarize the contributions of this paper as follows.

- We propose a novel information diffusion research problem, Diffusion Participation Forecasting (DPF), whose setting is more realistic and essentially different from existing studies in three folds: (1) forecasting the participants in multiple hops away from the early adopters, (2) making the prediction of new items from past diffusions, and (3) setting on no diffusion content and no user profiles.
- We develop an Adoption-based Participation Ranking (APR) model. Based on only the early adopters and the social network structure, APR jointly learns the adoption probability values of nodes and estimates the reachability scores of nodes to generate a ranking list as the forecasting result. In addition, we develop a robust collection of predictive feature sets (i.e., Diffusive, Social, Fringe, Temporal, First-Adopter, and Target) to learn the information adoption.
- Experiments conducted on a real Twitter dataset demonstrates APR can outperform baseline and state-of-the-art competitors, and demonstrate the effectiveness of the developed feature sets. The learned adoption probability of each node is also proven to be useful in boosting the performance.
- The proposed APR enables two applications, Forecasting-enhanced Influence Maximization and Adopter-expanded Popularity Prediction, in which the early adopters of information diffusion can provide more realistic settings for targeted marketing and tweet popularity prediction. Empirical studies show the forecasted diffusion participants generated by APR can boost the quality of such two applications.

**Table 1**
Comparison of existing studies and our work.

| | One-L | Multi-L | Sep-D | Cross-D | Content | Profile |
|---|---|---|---|---|---|---|
| [8] | ✓ | | | ✓ | ✓ | |
| [6] | ✓ | ✓ | ✓ | | | |
| [2] | ✓ | ✓ | | ✓ | | ✓ |
| [18] | ✓ | | | ✓ | ✓ | |
| [21] | ✓ | | | ✓ | | ✓ |
| [36] | ✓ | | ✓ | | | |
| [25] | ✓ | | ✓ | | | ✓ |
| [3] | ✓ | | | ✓ | ✓ | ✓ |
| [4] | ✓ | ✓ | | ✓ | ✓ | |
| [37] | ✓ | | | ✓ | ✓ | ✓ |
| [15] | ✓ | | ✓ | | ✓ | ✓ |
| **Our work** | ✓ | ✓ | | ✓ | | |

In the following, we first review related work in Section 2, followed by the problem formulation in Section 3. Then we present the details about the proposed method in Section 4. We will provide experimental results in Section 5, present the applications to influence maximization and popularity prediction in Section 6, and conclude this work in Section 7.

## 2. Related work

Existing studies on predicting participation of information diffusion for users can be categorized using three aspects. The first is the position of the target node $v$ (to be predicted): $v$ can belong to either the one-layer followers (*One-L*) or the multi-layer followers (*Multi-L*) away from the early adopters. The One-L's targets are users directly follow the early adopters. The diffusion participation of One-L users can be usually predicted by the percentage of early adopters among their followees. The Multi-L's targets are the indirect followers with two, three or more steps away from the early adopters. An important clue to predict the diffusion participation of Multi-L targets is the structural proximity from early adopters to them.

The second is the settings of training and prediction. There are two common settings: (a) learning from *partial* observations of *separate* diffusions and predicting its future participants (denoted by *Sep-D*), and (b) learning from *past* observations of *different* diffusions and predicting the participants of new diffusions (denoted by *Cross-D*). Sep-D divides all of the diffusion participants of a certain topic $c$ into training and testing sets, and learns how $c$ can be propagated from one node to another. Cross-D constructs the predictive model using past topics with different diffusion participants to learn users' habits of participation. Then Cross-D forecasts which users will participate in the diffusion of a new given item. It is worthwhile noting that the concept of Cross-D is similar to the learning from multiple social networks [23,28], whose main idea is to inferring the missing user-generated information by learning a latent space shared by different social networks. Although the topology of each information diffusion can be regarded as a particular network, the essential differences lie in that information diffusion may be participated by a small set of users (compared to the entire network) and it highly depends on user preferences on the diffused items. Hence the learning from multiple social networks is infeasible in this work.

The third is the information used for training and prediction. In addition to the social network and the information propagation, which are commonly used in existing work, some studies further consider either the *Content* information of topics or the *Profile* information of users. Content information usually refers to the posting texts or the visual semantics of images while profile information may include the age, gender, locations, endorsing log, and interests of users.

We create Table 1 to summarize the past relevant studies using such three dimensions, and to distinguish our work from them in the research line of diffusion participation prediction. It can be observed that our work is the first attempt to predict the diffusion participation for one and multi-layer followers through learning from past diffusion records without using content and profile information. In the experiments, we will compare the performance of our proposed model with the most similar work [2]. Note that it can be found that the study [4] is also relevant to our work. However, their method highly depends on content information of tweets while our setting assumes no content information is accessible due to either privacy or storage constraints. To avoid an unfair comparison, we did not compare with the method [4] in the experiment.

In addition to the aforementioned diffusion participation forecasting, some recent advances provide complementary investigation (not to predict the diffusion participants, but to characterize the diffusion process) in the research line of predicting information diffusion. Guille and Hacid [13] aim at predicting the temporal dynamics (e.g., the volume of participants over time) by learning the time-dependent diffusion probabilities via the behavioral features of individuals. Zhu et al. [40] propose to discover the most significant paths of information diffusion over multiple spreading events so that a more effective diffusion mechanism can be designed. Yoo et al. [35] alternatively deliver a field study, along with both the Independent Cascade and Linear Threshold models, to estimate how effectively the information diffuses in a social network.

## 3. Problem formulation

First, the universe set of hashtags is denoted by $C = \{c_1, c_2, \cdots, c_m\}$. A social network is a graph $G = (V, E)$ consisting of a node set $V$ and an edge set $E$. Since our used Twitter data contains follower–followee relationships between users, a

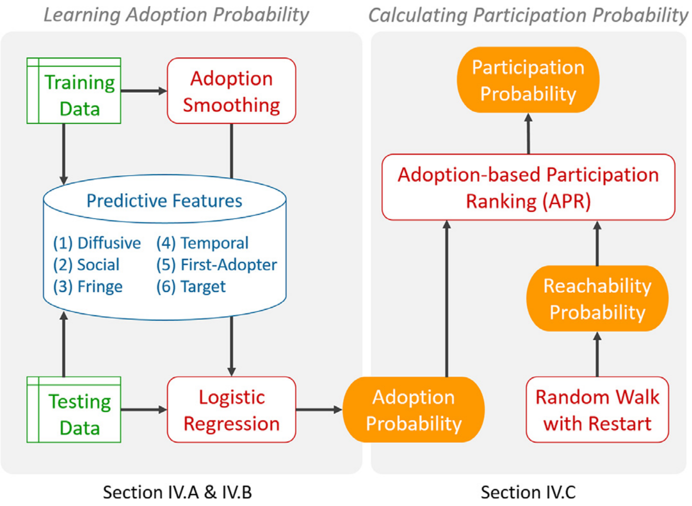

**Fig. 1.** The overview of the proposed method.

directed edge $e = (u \to v)$ can be created if node $v$ follows node $u$. Each node $v \in V$ has two states for each hashtag $c \in C$: $s_v(c) = 1$ indicates node $v$ had ever participated in the diffusion of hashtag $c$ (i.e., had ever adopted $c$) while $s_v(c) = 0$ refers to never adopting $c$. In addition, for each node $v$ with $s_v(c) = 1$, it is associated with a positive integer $t_v(c)$ that represents the adopting order. The order value $t_v(c)$ gets low if $v$ is an early adopter. Second, let $A(c)$ be the set of nodes that adopt hashtag $c$. Based the adopting order $t_v(c)$ for hashtag $c$, we can obtain the earliest $k$ nodes who adopt $c$, denoted by $A(c) = \{v \in V | s_v(c) = 1, t_v(c) \le k\}$. Third, let $(u, v, c, t)$ to be a diffusion action representing that node $v$ had ever adopted hashtag $c$ from node $u$ at time $t$. It can be also treated as node $v$ had ever reshared/retweeted $u$'s post containing hashtag $c$. Fourth, the diffusion graph of the $k$ earliest nodes adopting $c$ can be constructed using the diffusion actions involving $A(c)$. The diffusion graph of $A(c)$ is denoted by $H^k(c) = (A(c), R^k(c))$, where each $r = (u, v) \in R^k(c)$ is a directed edge constructed from the corresponding diffusion action $(u, v, c, t)$. Fifth, a social subgraph behind $A(c)$ can be extracted by finding its induced subgraph in the social network $G$. We denote the social subgraph of $A(c)$ as $G[A(c)] = (A(c), E[A(c)])$.

**Problem Definition: Diffusion Participation Forecasting (DPF).** Given a social network $G$, a hashtag $c$, and a $k$-node set $A(c)$ of early adopters for hashtag $c$ (i.e., early participants of the diffusion of hashtag $c$), together with the corresponding diffusion graph $H^k(c)$ and social subgraph $G[A(c)]$, the DPF problem is: among nodes in $V \backslash A(c)$, find the future diffusion participants who will adopt hashtag $c$ after the early adoption of $A(c)$.

We aim to solve the DPF problem by decomposing it into two stages. Specifically, a collection of hashtags $C$ is divided into a training set $C^{\bullet}$ and a testing set $C^{\circ}$. For each $c^{\bullet} \in C^{\bullet}$, we have several training instances, and each instance consists of $A(c)$ and a target node $v \in V \backslash A(c)$. The first stage is to construct a predictive model $\mathcal{M}$ based on $A(c^{\bullet})$, and use $\mathcal{M}$ to estimate the adoption probability $p_{c^{\circ}}(v)$ that a node $v$ will adopt a testing hashtag $c^{\circ}$ given $A(c^{\circ})$. The second stage is to generate a ranking list for all nodes in $V \backslash A(c^{\circ})$ such that the true diffusion participants can be placed at higher positions in the list. The ranking mechanism will consider the social network $G$, the $k$ earliest adopters $A(c^{\circ})$ for hashtag $c^{\circ}$, and the learned adoption probability $p_{c^{\circ}}(v)$.

## 4. Proposed method

We propose an adoption-based ranking model to solve the DPF problem. The rationale of our model is that the possibility of a user $v$ to participate the diffusion of hashtag $c$ is affected by two deterministic factors: (a) whether $v$ will be interested or willing to adopt $c$, and (b) whether $c$ can flow into $v$ by starting the propagation from early adopters. If the diffusion of $c$ has higher potential to reach user $v$ who also has high willingness to adopt it, $v$ tends to participate in such diffusion. On the other hand, the participation possibility of $v$ would be discounted if either the diffusion of $c$ is hard to reach $v$ or $v$ is less interested in $c$.

Based on these ideas, we first propose to learn the *adoption probability* of a node, which quantifies the willingness/preference that user $v$ feels about hashtag $c$. The learning procedure is based on the diffusions of training hashtags. It estimates the adoption probability values for new (testing) hashtags. In addition, we compute the *participation probability* of a node, which jointly models the potential that hashtag $c$ can reach $v$ through diffusion and the possibility of being interested by user $v$. The estimation of participation probability for new hashtag $c$ will rely on $c$'s $k$ earliest adopters and the learned adoption probability.

The overview of the proposed method is shown in Fig. 1. The proposed method consists of two parts. The first part is to estimate the *adoption probability* of a target node whose influence is incurred by the set of early adopters (not "an" particular early adopter) by a trained *logistic regression* model (Section 4.1) using a robust set of predictive features (Section 4.2). In learning the adoption probabilities for target nodes, we develop a *smoothing* strategy to address the data sparsity problem of hashtag adoption among users. The second part is to compute the *participation probability* for the target node by a participation ranking model (Section 4.3). Technically speaking, with the trained model that can output the probability that each testing/target node adopts the hashtag, we also compute the *reachability probability* (i.e., the probability the hashtag flows through the target node) using the technique of *random walk with restart*. Eventually a ranking model will combine both the adoption probability and the reachability probability to obtain the participation probability of each target node. Then the model will output those nodes with the highest potential to participate in the diffusion of the hashtag.

### 4.1. Estimating the adoption probability

Given a new hashtag $c^{\circ}$ and its $k$ earliest adopters $A(c^{\circ})$, we develop a predictive model to estimate the adoption probability $p_{c^{\circ}}(v)$ that a user $v$ will adopt $c^{\circ}$. In order to train the model, we employ the past hashtags $c^{\bullet} \in C^{\bullet}$, the corresponding $k$ earliest adopters $A(c^{\bullet})$, and each node $u \in V \backslash A(c^{\bullet})$, where $u$ could ever adopted $c^{\bullet}$ or not (i.e., $s_u(c^{\circ})$ can be 1 or 0). We consider a pair of early-adopter set $A(c^{\bullet})$ and a target node $u \notin A(c^{\bullet})$ to be an instance in the model. Six categories of features are extracted and fed into a learning model to capture the hidden correlation between $A(c^{\bullet})$ and target node $u$. With a certain model training, we can build a predictive model $\mathcal{M}$ so that given the testing instance (i.e., an early-adopter set $A(c^{\circ})$ and a node $v \notin A(c^{\circ})$), we can estimate $v$'s adoption probability $p_{c^{\circ}}(v)$: i.e., $\mathcal{M}(A(c^{\circ}), v) \rightarrow p_{c^{\circ}}(v)$, where $p_{c^{\circ}}(v) \in [0, 1]$.

To construct the predictive model $\mathcal{M}$, given the set of early adopters $A(c^{\bullet})$, we need to have the adoption probability $p_{c^{\bullet}}(v)$ for $v \in V \backslash A(c^{\bullet})$ and the list of features that characterize $A(c^{\bullet})$. If node $v$ had adopted $c^{\bullet}$, it is natural that $p_{c^{\bullet}}(v) = 1$. However, the number of nodes who adopted $c^{\bullet}$ is very small, compared with the entire node set $V$ in the network. Such fact leads to extreme imbalance for $p_{c^{\bullet}}(v)$ having values or not. Too many nodes with zero adoption probability values ($p_{c^{\bullet}}(v) = 0$) may destroy the quality of $\mathcal{M}$. To alleviate the imbalance problem, we propose to *smooth* the adoption probability values for nodes. The goal of the proposed smoothing is to let the predictive model of adoption probability have sufficient training data. While the adoption probability is to estimate the preference or willingness that a user will adopt a hashtag, we assume that if two users have more overlapping hashtags in the past and are close enough (e.g. friends) in the social network, the preference/willingness will be higher. In other words, if two nodes have more commonly used hashtags in the past, they tend to influence each other, i.e., have higher potential to adopt each other's hashtags. Also if two nodes have a lower shortest path length in the network, their hashtags have higher possibility to be seen and adopted by one another. Based on such ideas, we compute the adoption probability value of a node $v$ for a training hashtag $c^{\bullet}$ by:

$$p_{c^{\bullet}}(v) = \begin{cases} \beta^{l(v,v^{\star})} \cdot \dfrac{C(v) \cap C(v^{\star})}{C(v) \cup C(v^{\star})} & \text{if } c^{\bullet} \notin C(v) \\ 1 & \text{if } c^{\bullet} \in C(v) \end{cases}, \tag{1}$$

where $v^{\star}$ is the early-adopter node with the lowest shortest path length $l(v, v^{\star})$, i.e., $v^{\star} = \arg\max_{u \in A(c^{\bullet})} l(v, u)$, $C(v)$ is the set of hashtags used by node $v$ in the past, and $\beta$ is a damping factor and empirically set as 0.05. Based on the early adopters of training hashtags, along with their features and smoothed adoption probability values, we employ *Logistic Regression* to build the predictive model to estimate the adoption probability values of users for each testing hashtag.

### 4.2. Predictive features

We develop six categories of features to represent the early adopters $A(c)$: (a) diffusive, (b) social, (c) fringe, (d) temporal, (e) first-adopter, and (f) target features. The features are extracted from the diffusion graph of $H^k(c)$, the social subgraph $G[A(c)]$, the original social network $G$, the set $C(v)$ of previously used hashtags of each node $v \in A(c)$, and the time $T_v(c)$ that node $v$ adopts hashtag $c$. We elaborate the main idea for each category of features, and the detailed features and their description are listed in Table 2.

**Diffusive Features.** The structure of information diffusion at the early stage can significantly affect the willingness that other nodes in the network to participate the diffusion participants [5]. We develop a series of topological features to characterize the diffusion structure $H^k(c)$ of hashtag $c$ among its early adopters $A(c)$. The general idea is that if the diffusion structure of $c$ has higher *virality* [11], i.e., more concentrated interactions on adopting hashtag $c$, other nodes may have stronger intent to adopt $c$. We consider that hashtag $c$ tends to be viral if $H^k(c)$ has more components (i.e., the diffusion starts from more nodes), shorter path length and diameter, higher clustering coefficient, more edges (i.e., more interactions), and higher outdegree values on nodes.

**Social Features.** The social behaviors of early adopters in the network have positive impact on predicting the diffusion actions of other users in the network. User behaviors can be represented by various social roles, which had been shown to be effective on finding nodes with higher possibility to join a diffusion [19]. We devise a set of social features to capture the social behaviors of early adopters $A(c)$ in the social network $G$ and social subgraph $G[A(c)]$ induced by $A(c)$. These features generally aim to reflect that the early adopters with more followers and tightly connected to each other in the network can boost the visibility of hashtag $c$ and then increase the potential of adopting $c$.

**Fringe Features.** We assume that a user $v$ has higher potential to adopt hashtag $c$ if $v$ can either easily reach or be highly exposed to the early adopters of $c$. That says, the neighborhood (i.e., followers, denoted by $\Gamma(u)$, and followees, denoted by $\Omega(u)$) of early adopter $u \in A(c)$ in the network $G$, termed *fringe*, affects the visibility that other node $v$'s participation on the

Table 2
List of features to characterize the $k$ earliest adopters $A(c)$ for estimating adoption probability values.

| | |
|---|---|
| **Diffusive Features** | |
| $\#components$ | number of components in $H^k(c)$ |
| $virality$ | average path length ($apl$) values of components in $H^k(c)$, take the max, min, and average $apl$ values |
| $diameter$ | diameter ($dia$) values of components in $H^k(c)$, take the max, min, and average $dia$ values |
| $clus\_coeff^H$ | clustering coefficient ($cc$) values of components in $H^k(c)$, take the max, min, and average $cc$ values |
| $out\_degree^H$ | max, min, and average out-degree values for the largest and the smallest components in $H^k(c)$ |
| **out_degree** | vector of out-degree values for the $k$ early-adopter nodes $A(c)$ in $H^k(c)$ |
| $\#flows$ | total number of edges in $H^k(c)$, and numbers of edges in the largest and the smallest components |
| **Social Features** | |
| $out\_degree^G$ | out-degree ($outdeg$) (i.e., number of followers) of nodes $A(c)$ in $G$, take max, min, and average $outdeg$ values |
| $out\_degree^{G'}$ | out-degree ($outdeg$) for nodes in $G[A(c)]$, take max, min, and average $outdeg$ values |
| $out\_out\_degree^G$ | second-layer out-degree ($outoutdeg$) (i.e., followers of followers) for nodes $A(c)$ in $G$, take max, min, and average |
| $out\_out\_degree^{G'}$ | second-layer out-degree ($outoutdeg$) for nodes in $G[A(c)]$, take max, min, and average $outoutdeg$ values |
| $clus\_coeff^G$ | clustering coefficient ($cc$) values of for nodes $A(c)$ in $G$, take the max, min, and average $cc$ values |
| $clus\_coeff^{G'}$ | clustering coefficient ($cc$) values of for nodes in $G[A(c)]$, take the max, min, and average $cc$ values |
| $tightness^{G'}$ | number of edges in $G[A(c)]$ |
| **Fringe Features** | |
| $1st\_can\_adopters$ | number of first-layer future candidate adopters of $A(c)$, i.e., $\lvert\{u \mid u \in \cup_{v\in A(c)}\Gamma(v), \text{and } u \notin A(c)\}\rvert$ |
| $2nd\_can\_adopters$ | number of second-layer future candidate adopters of $A(c)$, i.e., $\lvert\{u \mid u \in \cup_{v\in A(c)}\Gamma^2(v)\backslash\Gamma(v), \text{and } u \notin A(c)\}\rvert$ |
| $exposure\_deg\_1st$ | number of exposure times that first-layer future candidates confront their early-adopters $A(c)$, i.e., $\Sigma_{v\in A(c)}\lvert\Gamma(v)\rvert$ |
| $exposure\_deg\_2nd$ | number of exposure times that second-layer future candidates confront $A(c)$, i.e., $\Sigma_{v\in A(c)}\lvert\Gamma^2(v)\backslash\Gamma(v)\rvert$ |
| $exposure\_ratio$ | percentage of early-adopters $A(c)$ among followees of candidate adopter, i.e., $\frac{1}{\lvert\cup_{v\in A(c)}\rvert}\sum_{u\in\cup_{v\in A(c)}\Gamma(v)}\frac{\Omega(u)\cap A(c)}{\Omega(u)}$ |
| **Temporal Features** | |
| $tot\_time$ | time difference between the first and the $k$th adoption for $c$, i.e., $T_{v_1}(c) - T_{v_k}(c)$ |
| $avg\Delta_{time}$ | average time difference between consecutive adoptions for $c$, i.e., $\frac{1}{k-1}\sum_{i=1}^{k-1}(T_{v_{i+1}}(c) - T_{v_i}(c))$ |
| $avg\Delta_{time}^{1..k/2}$ | average time difference between for the first $k/2$ consecutive adoptions, i.e., $\frac{1}{k/2-1}\sum_{i=1}^{k/2-1}(T_{v_{i+1}}(c) - T_{v_i}(c))$ |
| $avg\Delta_{time}^{k/2..k}$ | average time difference between for the last $k/2$ consecutive adoptions, i.e., $\frac{1}{k/2-1}\sum_{i=k/2}^{k-1}(T_{v_{i+1}}(c) - T_{v_i}(c))$ |
| $adoption\_speed$ | number of early adopters per time unit, i.e., $\frac{\lvert A(c)\rvert}{T_{v_1}(c) - T_{v_k}(c)}$ |
| $exposure\_speed$ | number of exposed first-layer candidate adopters per time unit, i.e., $\frac{\lvert\cup_{v\in A(c)}\Gamma(v)\backslash A(c)\rvert}{T_{v_1}(c) - T_{v_k}(c)}$ |
| **First-Adopter** $v_1(c)$**Features** | |
| $\#tags$ | number of past hashtags used by $v_1(c)$, i.e., $\lvert C(v_1(c))\rvert$ |
| $avg\_adopters$ | average number of final participants for past hashtags that $v_1(c)$ had ever adopted, i.e., $\frac{1}{\lvert C(v_1(c))\rvert}\sum_{c'\in C(v_1(c))}A(c')$ |
| $avg\_outdeg^H$ | average out-degree values $outdeg^{H^k(c')}$ of $v_1(c)$ in the past diffusion graphs of $v_1(c)$: $H^k(c')$ $(\forall c' \in C(v_1(c)))$ |
| $loyalty$ | average percentage of adopters among $v_1(c)$'s followers in its past diffusion graphs, i.e., $\frac{1}{\lvert C(v_1(c))\rvert}\frac{outdeg^{H^k(c')}(v_1(c))}{outdeg^G(v_1(c))}$ |
| $cur\_diffu\_ratio$ | percentage of adopters among $v_1(c)$'s followers in the current diffusion graph $H^k(c)$, i.e., $\frac{outdeg^{H^k(c)}(v_1(c))}{outdeg^G(v_1(c))}$ |
| **Target Features** | |
| $rwr$ | probability of random walk with restarting [30] from early adopters $A(c)$ to reach a candidate target $v$ |
| $comm\_hashtags$ | number of common tags between early adopters $u \in A(c)$ and the target $v$, i.e., $\frac{\lvert C(v)\cap C(u)\rvert}{\lvert C(v)\cup C(u)\rvert}$, take max, min, average |
| $1st\_nbr\_adopters$ | number of the target $v$'s followees who had adopted hashtag $c$, i.e., $\lvert\Omega(v)\cap A(c)\rvert$ |
| $2nd\_nbr\_adopters$ | number of the target $v$'s followees of followees who had adopted hashtag $c$, i.e., $\lvert\Omega^2(v)\cap A(c)\rvert$ |
| $nbr\_mediators$ | number of the target $v$'s followees whose followees had adopted hashtag $c$, i.e., $\lvert\{u \in \Omega(v) \mid u \in \cup_{w\in A(c)}\Gamma(w)\}\rvert$ |
| $past\_adoption$ | times that the target $v$ adopted past hashtags adopted by $A(c)$, i.e., $\lvert C(v)\cap C(u)\rvert$ $(\forall u \in A(c))$, take max, min, average |
| $past\_adoption\_bin$ | whether or not the target $v$ had ever adopted hashtags in the past |
| $past\_adoption\_all$ | number of hashtags that the target $v$ adopted in the past, i.e., $\lvert C(v)\backslash\{c\}\rvert$ |

diffusion of $c$. Fringe features are developed to capture that a larger neighborhood of $A(c)$ or higher degree of exposure to $A(c)$ can lead to higher probability of other node $v$ adopting $c$ in the future.

**Temporal Features.** The speed or efficiency of performing adoption actions is an important property for modeling information propagation [14]. An adoption outbreak is characterized by a large number of early adopters within a short period [5]. While each user $v$'s adoption on a hashtag $c$ is associated with time information $T_v(c)$, we use the time difference between early adopters $A(c)$ to model the diffusion efficiency of $c$. In addition, we measure the adoption speed by quantify the number of early adopters and nodes exposed to $c$ per time unit.

**First-Adopter Features.** The features of the first participant of information diffusion can be more deterministic for predicting the future participant size than the content information [24]. While more participants can boost the possibility of hashtag $c$ to be notified by other users, we propose to extract features for $c$'s first adopter, denoted by $v_1(c)$. The features are developed to capture $v_1(c)$'s willingness of adopting different hashtags (measured by her past used hashtags), final size of diffusion after $v_1(c)$'s adoption, adoption loyalty among $v_1(c)$'s followers, and potential audience in the current hashtag $c$.

**Target Features.** There is strong evidence that the probability of the target node $v$ (to be predicted) adopting hashtag $c$ is positively correlated with the early adopters of $c$ [36]. More $c$'s early adopters directly or indirectly connecting to the target $v$ can lead to higher potential of participating in the diffusion of $c$. In addition to using the number of early adopters in node $v$'s one-step followees [36], we further consider the structural proximity in the network $G$, two-step followees, the common hashtags (using Jaccard's Coefficient), the mediators, and the past diffusion interactions between the target $v$ and the early adopters $A(c)$. Note that only non-early adopters can be treated as the target users. Hence, those non-early adopters with higher participation probability values will be considered as being predicted to participate the diffusion.

It should be noted that the first five feature sets are based on only the early adopters, and solely the Target features are used to quantify the possibility that the target user will adopt the hashtag. We need to point out that the willingness for a user to adopt a hashtag highly relies on the *popularity* of such hashtag based on several existing studies [27,31]. This fact is characterized by the first five feature sets. In fact, most hashtags in Twitter are adopted by a very limited number of users. Only a small proportion of hashtags can widely spread. If a hashtag is essentially impossible to affect more users, the willingness/preference of adopting it would be lower [1]. In other words, users tend to appreciate or adopt popular hashtags. Therefore, we propose Diffusive, Social, Fringe, and Temporal features to characterize the *popularity* or diffusion capability of a hashtag from different aspects about its early diffusion of users.

We also want to point out that since our problem setting is given the earliest $k$ users who adopt a hashtag, the features are constructed based on such $k$ earliest adopters, instead of a fixed time. The late $n$th ($n > k$) adopters are not considered for feature extraction to avoid peeking into the future. We had ensured the Target features are constructed based on only the training set.

### *4.3. Ranking model*

We propose a ranking model to forecast future diffusion participants. Given the early adopters $A(c^\circ)$, and the derived adoption probability of new hashtag $c^\circ$ for each target node $v \in V \backslash A(c^\circ$, the model is expected to generate a *participation probability* for each target $v$, which is used to produce the ranking results. We devise the ranking model based on a general principle. A node $v$ has higher probability to participate in a diffusion of $c^\circ$ if $v$ is able to not only be easily *reach* by $c^\circ$'s diffusion, but also possesses a higher adoption probability (i.e., preferring hashtag $c^\circ$ or be willing to adopt $c^\circ$). While the adoption probability $p_{c^\circ}(v)$ can be derived based on the predictive model $\mathcal{M}$, we use Random Walk with Restart (RWR) [30] to estimate the reachability from early adopters $A(c^\circ)$ to each target $v$. In addition, we consider nodes highly exposed to those with higher adoption probabilities can also tend to participate the diffusion of $c^\circ$. Note that graph-based semi-supervised learning [39] seems to be a possible solution to forecast the diffusion participants. However, its propagation mechanism highly depends on the similarity scores between nodes, which is not available to be obtained in our problem setting (we assume the profile information cannot be accessed). Hence it cannot be applied here. Another plausible approach to forecast the diffusion participants is Center-Piece Subgraph [7], which aims at extracting the subgraph whose nodes are structurally proximate to the query nodes in a simultaneous manner. However, it cannot be applied here since information diffusion cannot be directly reflected by the concurrent proximity with respect to all of the query nodes.

Let $\mathbf{x}$ be the vector of the participation probability $x_{c^\circ}(v)$ of hashtag $c^\circ$ for all target nodes, and be initialized using RWR scores restarting from early adopters $A(c^\circ)$. Also let $\mathbf{p}$ be the vector of the learned adoption probability $p_{c^\circ}(v)$. We develop the random walk-based mechanism based on an iterative process:

$$\mathbf{x}^{(t+1)} = (1-\alpha)\tilde{W}\mathbf{x}^{(t)} + \alpha\mathbf{p}, \tag{2}$$

where $\tilde{W}$ is the normalized weighted matrix associated with $W$ (the adjacency matrix of $G$), $\alpha$ is a weighting parameter, $0 < \alpha < 1$, which controls the preference between participation probability vector $\mathbf{x}$ and adoption probability vector $\mathbf{p}$. We set $\alpha = 0.7$ empirically to have best results. The vector $\mathbf{x}$ can converge when $\alpha$ is not too close to 1. Specifically, based on the PageRank error theory [10], we can have that if $\mathbf{x}^{(0)} \neq 0$, then $\left\|\mathbf{x} - \mathbf{x}^{(t)}\right\|_1 \leq 2\alpha^t$, where $\mathbf{x}$ is the true solution. We set $\alpha$ to be between 0.1 and 0.99. In the worst case that $\alpha = 0.99$, it takes only at most 3000 iterations to make the error converge to be less than $10^{-8}$. Eventually we can generate a ranking list for all the nodes $v \in V \backslash A(c^\circ)$ based on the derived participation probability vector $\mathbf{x}$. Users with higher participation probability values are ranked at higher positions, and are considered as the forecasted participants for the diffusion of hashtag $c^\circ$.

## 5. Experiments

We conduct experiments to evaluate our adoption-based participation ranking (APR). The evaluation goal is four-fold. First, we want to understand the performance of APR, comparing to existing approaches. Second, some information diffuse from central users (i.e., nodes with higher degree values) while some diffuse from peripheral users. We aim to investigate how do the positions of the early adopters influence the performance. Third, since some hashtags are more popular (i.e., adopted by more users) than others, we wonder the correlation between performance and the popularity of hashtags. Fourth, since our model is composed by estimating the adoption probability and the participation probability, and such two parts are controlled by a parameter $\alpha$, we should report the results by changing $\alpha$ to further understand our model.

### *5.1. Data and evaluation settings*

We use the Twitter data collected by Weng et al. [32], in which there are 121,807,378 tweets with 10,393,465 hashtags during Mar 24 and Apr 25, 2012. The social network is constructed based on their Follow relationship, i.e., a node $u$ can follow another node $v$ in Twitter. We use both Retweet and Mention actions to extract so-called information diffusion. That says, if user $u$ shares a tweet that is posted by another user $v$, or $u$ is mentioned in a tweet posted by $v$, we consider that

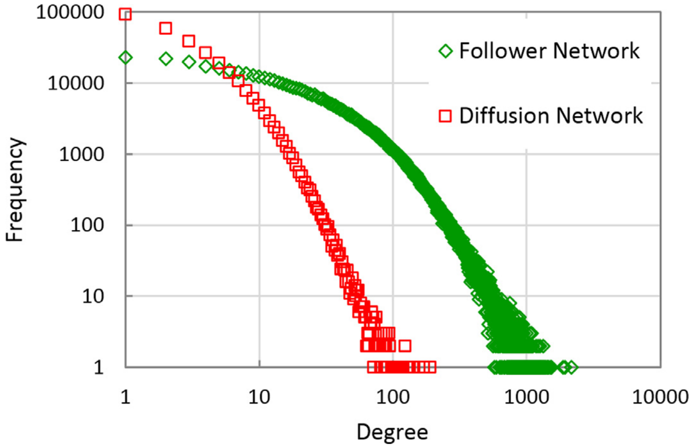

**Fig. 2.** Degree distributions of two networks.

**Table 3**
Statistics and properties for two networks.

| | #nodes | #edges | CluCoef | APL | alpha |
|---|---|---|---|---|---|
| Follower | 595,460 | 14,273,311 | 0.088 | 4.70 | 1.34 |
| Diffusion | 300,197 | 598,487 | 0.093 | 9.99 | 1.31 |

the hashtags used in such tweet are diffused from user $v$ to user $u$. The diffusions of hashtags can be also used to construct a diffusion network. The statistics with degree distributions of the follower social network and the diffusion network are shown in Table 3 and Fig. 2.

We compare our APR model with three competitors. The first is the state-of-the-art diffusion prediction model, *Supervised Random Walk* (SRW) [2], with their proposed features. The second and the third competitors are two typical regression-based methods, *Logistic Regression* (LR) and *Support Vector Regression* (SVR). Given our proposed features and the smoothed adoption probability values, we use LR and SVR to learn a predictive model between features of early adopters and the tendency of hashtag adoption of users. Users with higher predicted values are placed at higher positions in the ranking list. We use Precision, Recall, and F1-measure as the evaluation metrics. The number of early adopters for each hashtag is fixed as 25, i.e., $|A(c)| = 25$. In addition, all the hashtags are divided into a training set and a testing set. The ground truth of our evaluation is the set of users who really adopt a testing hashtag $c^{\circ}$ (i.e., participating the diffusion of $c^{\circ}$) after the early adopters. Note that we conduct the prediction at the time point that the diffusion had ended. The ending time of diffusions varies from one hashtag to another.

The evaluation consists of three parts, in which the first one is the general performance study while the last two belong to the sensitivity analysis. The first is *General Evaluation*, which presents the performance by varying $K$ highly ranked users ($K$ = 25, 50, 100, 200) to understand whether a method can find the actual participants at higher positions in the ranking list. The second is *Scenario Evaluation*, which exhibits the performance of our APR model by changing the neighborhood size of early adopters and the hashtag popularity, respectively (by fixing $K$ = 200). The neighborhood size is defined as the count of followers and their followers of the early adopters. Early adopters with a larger neighborhood mean that the hashtag is diffused from the central positions in the network while early adopters with a smaller neighborhood refer to diffusing from the peripheral positions. The hashtag popularity is defined as the times a hashtag had ever been mentioned or retweeted. The testing hashtags are equally divided into five levels (i.e., Level 1, 2, ... , 5) based on the neighborhood size of their early adopters and the hashtag popularity, respectively. A higher level value (e.g., 5) indicates a larger neighborhood and more popular hashtags. The third is *Parameter Evaluation*. We show the performance by varying the $\alpha$ value ($\alpha = 0.1, 0.3, 0.5, 0.7, 0.9$) to find the best setting of our model. Note that the damping factor $\beta$ in the adoption probability (in Eq. (1)) is set as 0.05 by default. The weighting parameter $\alpha$ in the mechanism of Random Walk with Restart (in Eq. (2)) is set as 0.7 by default. Besides, in the remaining evaluation except for the General Evaluation, the number of early adopters is set as 25 (i.e., $K = 25$).

### 5.2. Experimental results

**General Evaluation.** The results of different methods by varying $K$ highly ranked users were shown in Fig. 3. The proposed APR generally outperformed other methods, especially the state-of-the-art method SRW, in terms of Precision, Recall, and F1. Even SVR and LR with our proposed features can beat SRW. Such results verified the effectiveness of our model and features. We further find that our APR model can obtain higher Precision scores when fewer users (e.g., $K$ = 25) were

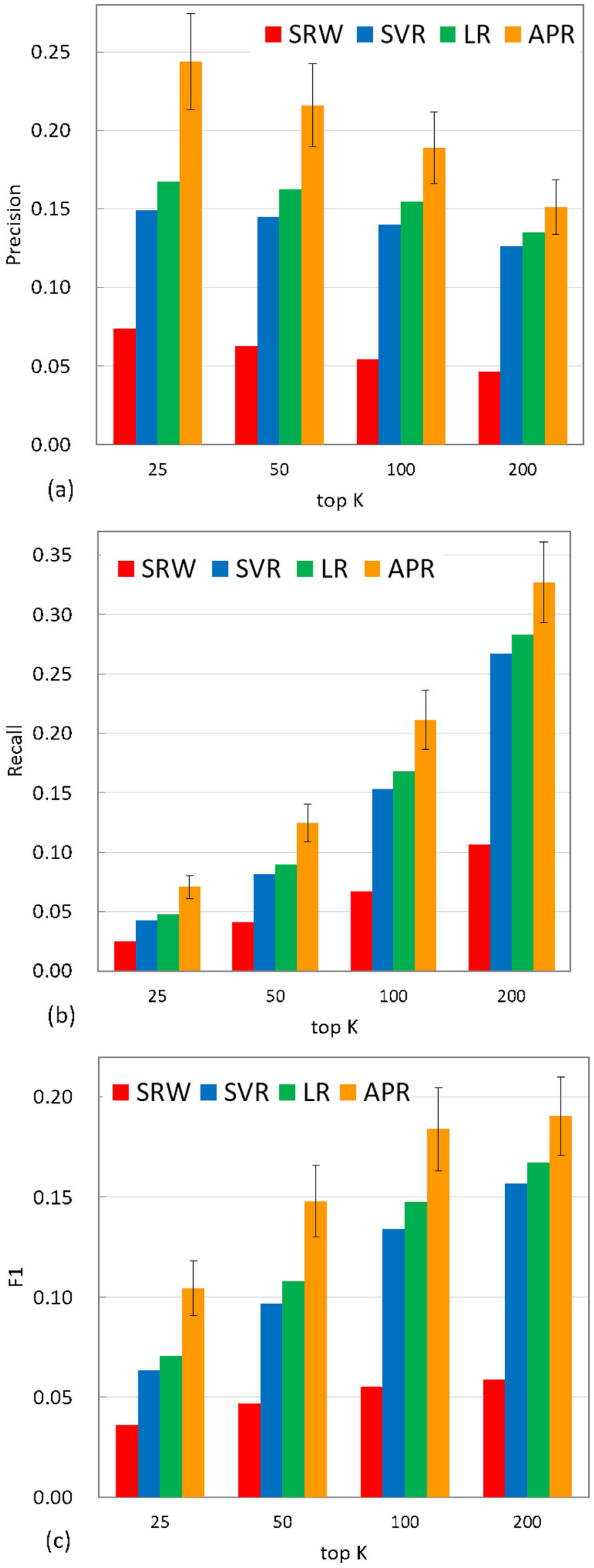

**Fig. 3.** Performance by varying *K* highly ranked users.

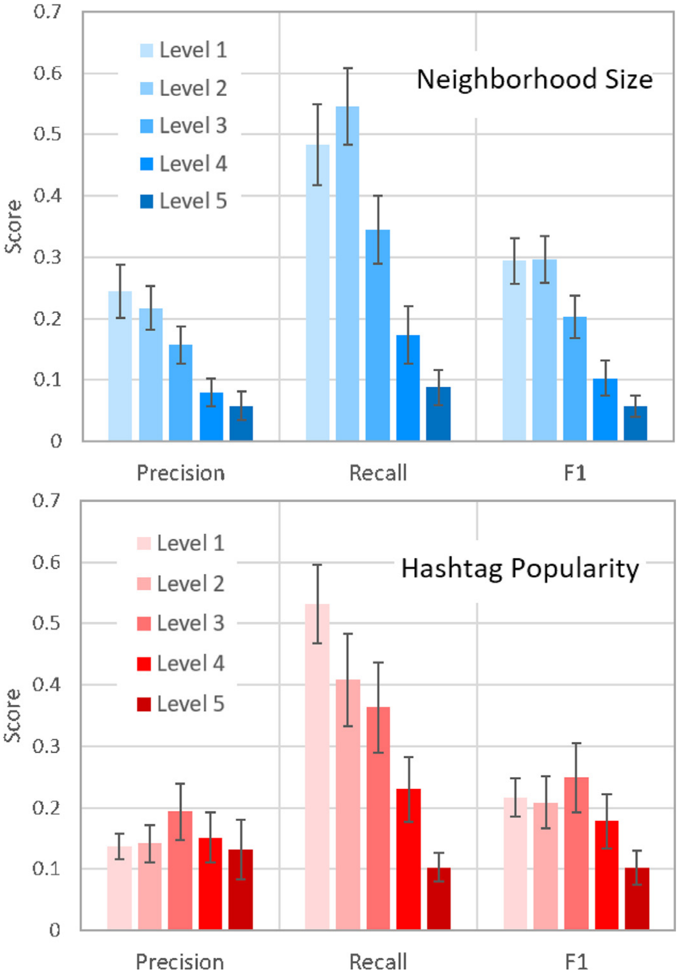

**Fig. 4.** Performance of our APR model for different Levels of neighborhood size and hashtag popularity.

reported. In addition, as $K$ increases, APR can steadily outperform other methods. Moreover, we learn that it is important to first estimate the adoption probability values for users, and then collectively use the early adopters and nodes with higher adoption probabilities to forecast future participants, rather than directly predicting the participation like the three competitors.

**Scenario Evaluation.** The results of ARP for different settings on neighborhood size and hashtag popularity were shown in Fig. 4(a) and (b). It was apparent that larger neighborhoods (e.g., Level 5) damage the performance. We think the reason is larger neighborhoods can bring more candidate users, which could weaken the predictability derived from estimating the adoption probability. We also find that the diffusion participants of more popular hashtags (e.g., Level 5) are more difficult to be accurately predicted. We think that people essentially tend to pursue popular topics. Such fact makes users with different features indistinguishable, and thus mitigates the effect of learning uses intents of participating diffusions.

**Parameter Evaluation.** We presented the results of different $\alpha$ values in Fig. 5. It is apparent that $\alpha = 0.7$ leads to the best performance. Since a higher $\alpha$ value in Eq. (2) gives the adoption probability a higher weight, this result reflected the importance of estimating the adoption probability. In addition, as $\alpha$ went too much higher (i.e., $\alpha = 0.9$), the performance dropped down. We think such decrement informs us the importance of the diffusion reachability from early adopters to other users cannot be over underemphasized.

## 6. Applications of diffusion participation forecasting

We further aim at validating the usefulness of forecasting diffusion participants through some applications, in addition to empirically verify the proposed method. In the research line of social information diffusion, two of the most essential topics are *Influence Maximization* [17] and *Popularity Prediction* [16]. Influence maximization is to find the most influential $k$ seed users in a social network while popularity prediction aims at estimating the number of users who will re-share a certain message or adopt a hashtag in the future. However, these two topics have some insufficient settings (will explain in

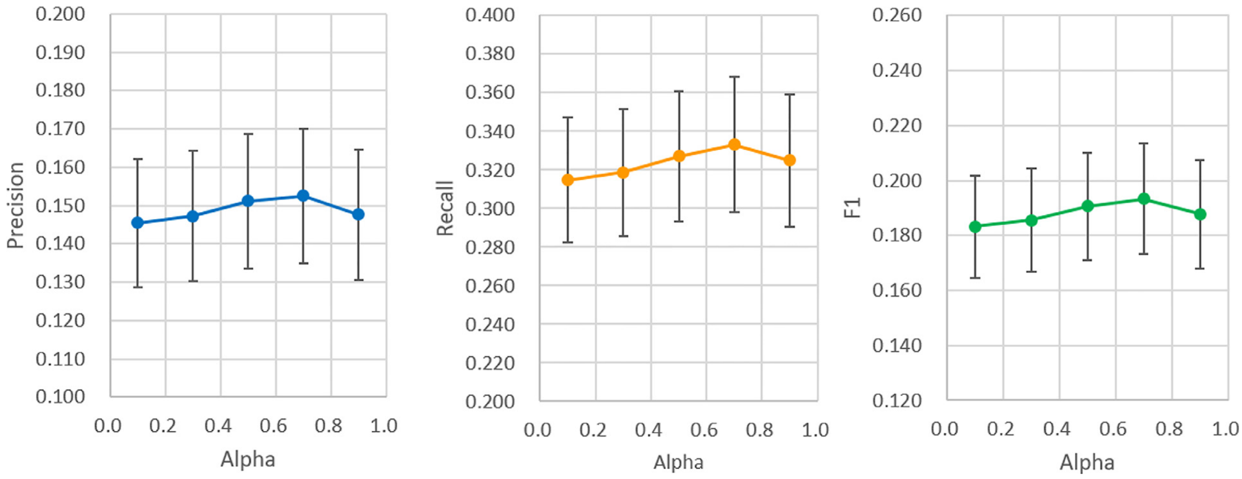

**Fig. 5.** Performance by varying $\alpha$ used in Eq. (2).

the following). We propose to smooth away the insufficient settings by imposing the forecasted diffusion participants in an realistic manner.

### 6.1. Influence maximization via diffusion forecasting

The original influence maximization problem is based on some simulation-based influence propagation model, e.g. the *Independent Cascade* (IC) model [17]. Although IC model can simulate which users will adopt the influence, the generated seed users may not be those users who truly adopt the information in the future. Besides, in real-world *targeted marketing*, the advertisers may target at some hashtags or products, and expect the identified influential users can have higher influence on those who will adopt the targeted hashtags or products. However, the IC-based simulation approach is unable to accurately find those influential users who will adopt the hashtag and successfully activate more people adopting the hashtag.

We propose to exploit the proposed diffusion participation forecasting to enhance the influence maximization. By extending the setting of the typical influence maximization, the goal is refined to find influential seed users who will not only adopt the given hashtag, but also successfully activate more people who will adopt the given hashtag. Our main idea is utilizing the forecasted diffusion participants of the given hashtag to distill the set of seed candidates. Therefore, we term the refined problem as *Forecasting-enhanced Influence Maximization* (FeIM), which is formally defined in the following.

**Problem Definition: Forecasting-enhanced Influence Maximization** (FeIM). Given a social network $G = (V, E)$, a hashtag $c$, and a $k$-node set $A(c)$ of early adopters for hashtag $c$, FeIM is to find a set $S$ of $n$ ($n > k$) seed nodes from $V$ such that (a) the number of seed nodes (in $S$) who will adopt hashtag $c$ is maximized, and (b) the number of successfully activated nodes who will adopt hashtag $c$ is maximized as well.

The intuitive approach to FeIM is selecting the $n$ seeds from the $k$-node set of early adopters $A(c)$. However, it is infeasible since $n$ is usually greater than $k$. In addition, not all the nodes in $A(c)$ are truly influential although they use the hashtag $c$. Thus selecting seeds from $A(c)$ may not lead to maximum influence spread. We devise a naive extension by expanding the seed candidates from the early adopters $A(c)$ to their neighborhood. That says, the $n$ influential seeds will be selected from $A(c) \cup \mathcal{N}^h(A(c))$, where $\mathcal{N}^h(A(c))$ is the neighboring nodes up to $h$ hops. We term this strategy as *Neighborhood Expansion*. This design has two rationales. First, the neighboring nodes of the early adopters have higher possibility to obtain the information about hashtag $c$ and further adopt it. Second, in addition to the set of early adopters, considering more nodes as the seed candidates can boost the potential of finding nodes that are more influential. We set $h = 2$ because more hops may include nodes that have less possibility to adopt hashtag $c$.

Since the goal is to find those who will adopt the hashtag $c$ and are influential with respect to $c$, we can alternatively exploit the forecasted diffusion participants for $c$ to expand the seed candidates, in addition to the set of early adopters $A(c)$. In other words, the seeds will be selected from $A(c) \cup \mathcal{F}(A(c))$, where $\mathcal{F}(A(c))$ is the set of forecasted diffusion participants based on hashtag $c$'s early adopters $A(c)$. We term this strategy as *Forecasting Expansion*. This approach is based on the homophily effect in influence propagation [22,34]: those with similar interests tend to gather together and influence one another. Therefore, if we can accurately identify more users who will adopt hashtag $c$, it is possible to lead to higher influence spread for $c$.

We conduct experiments to compare *Forecasting Expansion* with *Neighborhood Expansion* for solving the FeIM problem. The greedy algorithm [17] is used to select the set of $n$ seed nodes. There are two set of evaluation metrics. The first includes *Seed Precision* (SP) and *Seed Recall* (SR), defined as $SP(S) = \frac{|S \cap V(c)|}{n}$ and $SR(S) = \frac{|S \cap V(c)|}{|V(c)|}$, where $V(c)$ is the set of users who adopt hashtag $c$ among all the nodes in $V$. The second is *Influence Precision* (IP) and *Influence Recall* (IR) for hashtag $c$, defined

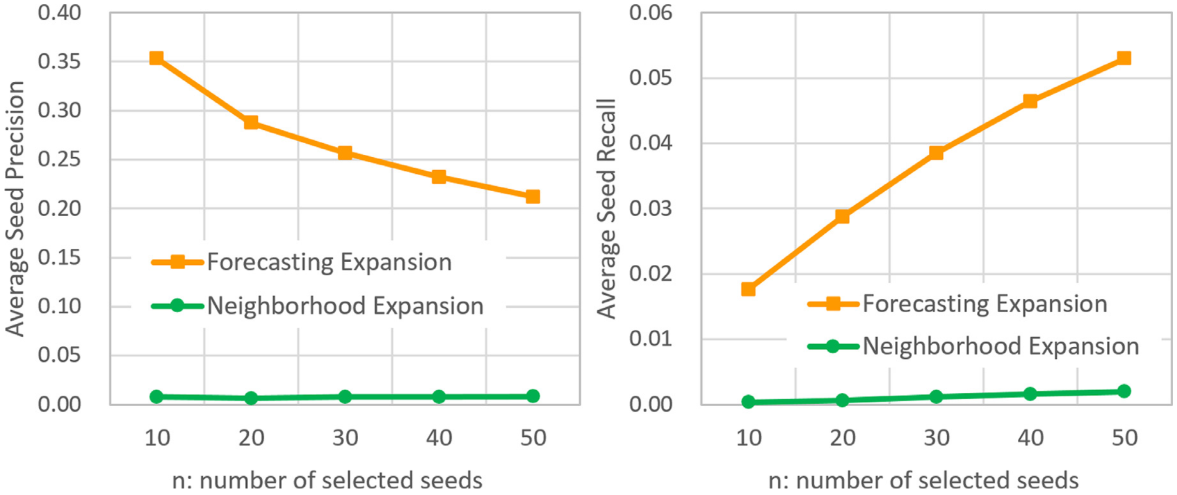

**Fig. 6.** Precision and Recall for the selected seeds.

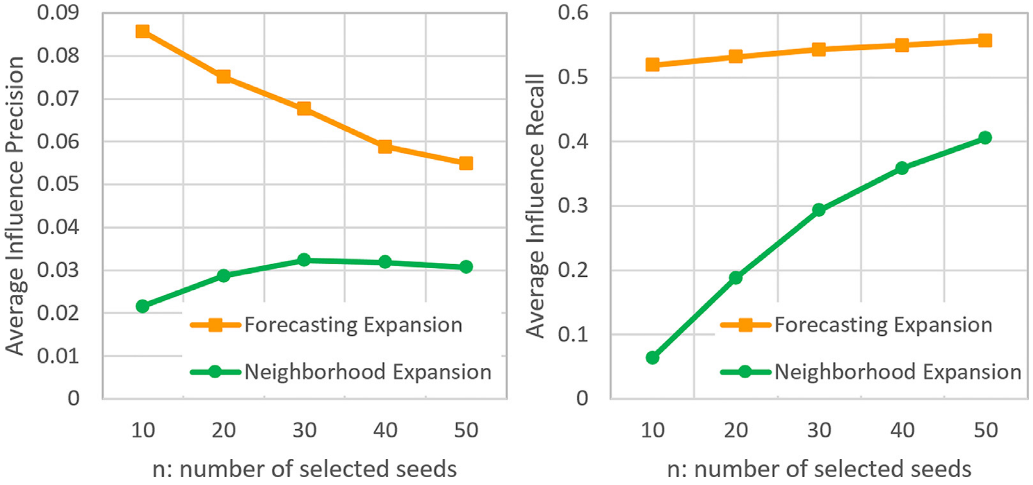

**Fig. 7.** Precision and Recall for the influence w.r.t. hashtag *c*.

as $IP(S) = \frac{|R(S) \cap V(c)|}{|R(S)|}$ and $IR(S) = \frac{|R(S) \cap V(c)|}{|V(c)|}$, where $R(S)$ is the set of users who are successfully activated by the selected seed set $S$. We compute the score of each evaluation metric for each hashtag, and report the average score over all the hashtags. We set $k = 25$ by default and vary $n = 10, 20, \ldots, 50$. To obtain $R(S)$, we choose IC model to spread the influence, in which the uniform propagation probability $p = 0.01$ is used [17]. The Monte-Carlo simulation is performed up to 1000 times, and nodes being successfully activated up to 10 times are added into $R(S)$.

The results are shown in Figs. 6 and 7. In general, we can find that Forecasting Expansion significantly outperforms Neighborhood Expansion in terms of the four evaluation metrics. Such result exhibits the effectiveness of forecasting diffusion participants on finding influential users who will adopt certain information. In finding influential seeds adopting hashtag $c$ in Fig. 6, we think the Precision would be more important because only those being willing to adopt $c$ are meaningful and practical to promote the hashtag/product. The average Seed Precision (SP) scores which are higher than 0.25 from 10 to 30 seeds could be relatively satisfying. It can be also observed that the scores of average Seed Recall are quite low (e.g. below 0.1). The reason is that the set of users who adopt hashtag $c$ (i.e., $V(c)$) is usually not small, and reporting the Seed Recall scores using the selected seeds up to 50 is not enough (but the value 50 is sufficient to find seeds for the marketing purpose).

As for the results of maximizing the number of successfully activated users adopting hashtag $c$ in Fig. 7, we think the Recall would be more meaningful. It is because the industry ideally expects that all the users who have higher potential to adopt the hashtag/product can be reach and activated. By varying the number of selected seeds to spread the influence, the average Influence Recall (IR) scores using Forecasting Expansion are higher than 0.5, which significantly outperforms Neighborhood Expansion. We believe the Neighborhood Expansion can truly lead to higher *general* influence spread (i.e., considering no hashtag $c$). However, in targeted marketing, considering whether activated users are interested in the hash-

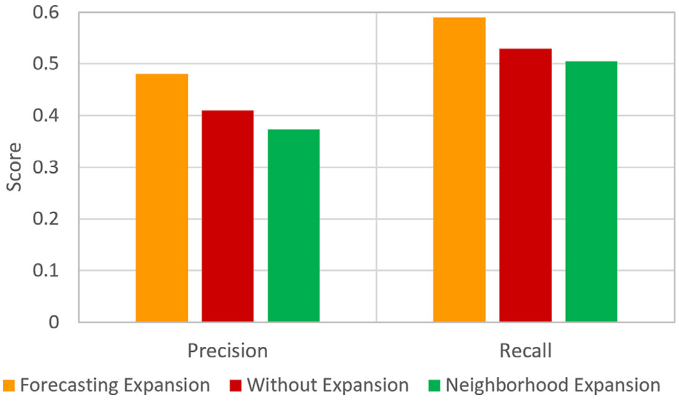

**Fig. 8.** Precision and Recall for popularity prediction.

tag/product would be more practical. We can also find that Influence Precision scores are quite low (less than 0.1). It may because there are a large number of activated users incurred by the influential seeds, which make $R(S)$ large. In short, these outcomes imply that the proposed method in this paper is able to not only accurately find the diffusion participants, but also bring effective seeding candidates and lead to higher influence spread of the hashtag/product for targeted marketing.

### 6.2. Popularity prediction with forecasted participants

The popularity prediction in a social network assumes that there are a set of early adopters for certain information (e.g. hashtag or message), and aims to predict the number of users who will adopt such information in the future (e.g. one month after). Since marketers want to establish the advertising strategies and understand the profit/cost of the campaign as early as possible, the set of early adopters is usually small due to the time limit. For example, they are given only 24 hours to estimate the potential of the product or the hashtag. In other words, one may have the early adopters to collect the early adopters and use them to estimate the potential. However, more early adopters can lead to higher prediction accuracy [5,20,38]. Therefore, in addition to the early adopters, it is highly demanded to have a certain method that can accurately find more possible adopters in the future.

Here we apply the diffusion participation forecasting technique to expand the set of early adopters. With the forecasted adopters, we aim at boosting the performance of popularity prediction for the given hashtag. We should point out that while popularity prediction is to estimate the "volume" of future adopters, several existing methods [9,29,38] make the prediction purely according to the evolution of volume change at early stages. Nevertheless, which users are the early adopters had been proven to have a positive effect on increasing the performance [5,20]. Hence we suppose that the prediction accuracy can get improved only if more adopters are accurately identified.

**Problem Definition: Adopter-expanded Popularity Prediction** (AePP). Given a social network $G = (V, E)$, a hashtag $c$, and a $k$-node set $A(c)$ of early adopters for hashtag $c$, AePP is to find a set $S$ of $n$ nodes from $V$ and considers them as additional adopters such that the performance of popularity prediction can be boosted.

Similar to the previous application to FeIM, we would like to know the effectiveness of treating the forecasted participants as additional adopters. The expanded set of adopters $A'(c)$ is $A'(c) = A(c) \cup \mathcal{F}(A(c))$. Hence we have the *Forecasting Expansion* for AePP. *Forecasting Expansion* will be empirically compared with *Neighborhood Expansion* and *Without Expansion*, which respectively represent making the prediction based on the expanded adopter set $A'(c) = A(c) \cup \mathcal{N}^m(A(c))$ $(m = 2)$ and the original early adopters. As for the method for popularity prediction, we employ Support Vector Regression along with the first five feature sets we propose (i.e., Diffusive, Social, Fringe, Temporal, and First-Adopter). In addition, we consider the popularity prediction as a multi-label classification task by following the setting [16,20,33]. The popularity values are divided into several classes based on the order of magnitude of the total popularity, i.e., $\lceil log_{10}|\hat{A}(c)| + 0.5\rceil$, where $|\hat{A}(c)|$ is the total number of users who adopt hashtag $c$ in the end. The number of early adopters $k$ is set as 25, and the number of forecasted diffusion participants who are used as additional adopters is set as 100. The evaluation metrics are Precision and Recall.

The evaluation results are shown in Fig. 8. We can find Forecasting Expansion obviously outperforms Neighborhood Expansion and Without Expansion in terms of both Precision and Recall. Such result proves that including additional adopters who are accurately forecasted can boost the performance of popularity prediction. In contrast, adding more users who could not adopt the hashtag may damage the performance, as presented by Neighborhood Expansion whose results are worse than Without Expansion.

## 7. Conclusion

This paper proposes to forecast the participants of information diffusion on social networks given a set of early adopters. The properties of our problem, compared with existing studies, are predicting the diffusion participation multi-layer users away from the early adopters and using no posting content and user profiles. An adoption-based participation ranking model is developed and validated to be effective on solve the DPF problem. The most important insight is the estimation of users' adoption probabilities, which models the behaviors and willingness of diffusion participation by six categories of proposed features. In addition, jointly using early adopters and other users with higher estimated adoption probability values can lead to satisfying experimental results. Moreover, with the applications to influence maximization and popularity prediction using the forecasted diffusion participants, we have proven being able to provide a more effective targeted marketing and an more accurate estimation of the potential of a campaign.

## Acknowledgements

This work was sponsored by Ministry of Science and Technology (MOST) of Taiwan under grants: 105-2628-E-001-002-MY2, 106-2628-E-006-005-MY3, 106-2118-M-006-010-MY2, 106-3114-E-002-008, and 106-3114-E-006-002. This work is also sponsored by FY16-Microsoft: FY16-RES-THEME-021.

## References

[1] A. Anderson, D. Huttenlocher, J. Kleinberg, J. Leskovec, M. Tiwari, Global diffusion via cascading invitations: structure, growth, and homophily, in: Proceedings of ACM International Conference on World Wide Web (WWW), 2015, pp. 66–76.

[2] L. Backstrom, J. Leskovec, Supervised random walks: predicting and recommending links in social networks, in: Proceedings of ACM International Conference on Web Search and Data Mining (WSDM), 2011, pp. 635–644.

[3] J. Bian, Y. Yang, T.-S. Chua, Predicting trending messages and diffusion participants in microblogging network, in: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR), 2014, pp. 537–546.

[4] S. Bourigault, C. Lagnier, S. Lamprier, L. Denoyer, P. Gallinari, Learning social network embeddings for predicting information diffusion, in: Proceedings of ACM International Conference on Web Search and Data Mining (WSDM), 2014, pp. 393–402.

[5] J. Cheng, L. Adamic, P.A. Dow, J.M. Kleinberg, J. Leskovec, 2014, Can cascades be predicted? Proceedings of ACM International Conference on World Wide Web (WWW), 925–936.

[6] F.C.T. Chua, H. Lauw, E.-P. Lim, Predicting item adoption using social correlation, in: Proceedings of SIAM International Conference on Data Mining (SDM), 2011, pp. 367–378.

[7] C. Faloutsos, K.S. McCurley, A. Tomkins, Fast discovery of connection subgraphs, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2004, pp. 118–127.

[8] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, W. Kellerer, Outtweeting the twitterers - predicting information cascades in microblogs, in: Proceedings of the 3rd Wonference on Online Social Networks (WOSN), 2010.

[9] S. Gao, J. Ma, Z. Chen, Modeling and predicting retweeting dynamics on microblogging platforms, in: Proceedings of ACM International Conference on Web Search and Data Mining (WSDM), 2015, pp. 107–116.

[10] D.F. Gleich, PageRank beyond the Web., SIAM Rev. 57 (3) (2015) 321–363.

[11] S. Goel, A. Anderson, J. Hofman, D.J. Watts, The structural virality of online diffusion, Manage. Sci. 62 (1) (2015) 180–196.

[12] A. Goyal, F. Bonchi, L.V.S. Lakshmanan, A data-based approach to social influence maximization, Proc. VLDB Endowment 5 (1) (2011) 73–84.

[13] A. Guille, H. Hacid, A predictive model for the temporal dynamics of information diffusion in online social networks, in: Proceedings of ACM International Conference on World Wide Web, 2012, pp. 1145–1152.

[14] C.-T. Ho, C.-T. Li, S.-D. Lin, Modeling and visualizing information propagation in a micro-blogging platform, in: Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2011, pp. 328–335.

[15] B.-T. Hoang, K. Chelghoum, I. Kacem, A learning-based model for predicting information diffusion in social networks: case of twitter, in: Proceedings of International Conference on Control, Decision and Information Technologies (CoDIT), 2016, pp. 752–757.

[16] L. Hong, O. Dan, B.D. Davison, Predicting popular messages in twitter, in: Proceedings of ACM International Conference on World Wide Web (WWW), 2011, pp. 57–58.

[17] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2003, pp. 137–146.

[18] T.-T. Kuo, S.-C. Hung, W.-S. Lin, N. Peng, D. Lin S, W.-F. Lin, Exploiting latent information to predict diffusions of novel topics on social networks, in: Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), 2012, pp. 344–348.

[19] C.-T. Li, Y.-J. Lin, M.-Y. Yeh, The roles of network communities in social information diffusion, in: Proceedings of IEEE International Conference on Big Data (BigData), 2015, pp. 391–400.

[20] C.-T. Li, M.-K. Shan, S.-H. Jheng, K.-C. Chou, Exploiting concept drift to predict popularity of social multimedia in microblogs, Inf. Sci. 339 (2016) 310–331.

[21] Z. Luo, M. Osborne, J. Tang, T. Wang, Who will retweet me?: Finding retweeters in twitter., in: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR), 2013, pp. 869–872.

[22] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: homophily in social networks, Annu. Rev. Sociol. 27 (1) (2011) 415–444.

[23] L. Nie, X. Song, T.-S. Chua, Learning from Multiple Social Networks. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool, 2016.

[24] S. Petrovic, M. Osborne, V. Lavrenko, RT to win! predicting message propagation in twitter, in: Proceedings of AAAI International Conference on Web and Social Media (ICWSM), 2011, pp. 586–589.

[25] A. Sadilek, H. Kautz, V. Silenzio, Predicting disease transmission from geo-tagged micro-blog data, in: Proceedings of AAAI International Conference on Artificial Intelligence (AAAI), 2014, pp. 136–142.

[26] D. Shahaf, J. Yang, C. Suen, J. Jacobs, H. Wang, J. Leskovec, 2013, Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 1097–1105.

[27] B. Shulman, A. Sharma, D. Cosley, Predictability of popularity: gaps between prediction and understanding, in: Proceedings of AAAI International Conference on Web and Social Media (ICWSM), 2016, pp. 348–357.

[28] X. Song, L. Nie, L. Zhang, M. Akbari, T.-S. Chua, Multiple social network learning and its application in volunteerism tendency prediction, in: Proceedings of ACM International Conference on Research and Development in Information Retrieval (SIGIR), 2016, pp. 213–222.

[29] G. Szabo, B.A. Huberman, Predicting the popularity of online content, Commun. ACM 53 (8) (2010) 80–88.

[30] H. Tong, C. Faloutsos, J.-Y. Pan, Fast random walk with restart and its applications, in: Proceedings of IEEE International Conference on Data Mining (ICDM), 2006, pp. 613–622.

[31] O. Tsur, A. Rappoport, What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities, in: Proceedings of ACM International Conference on Web Search and Data Mining (WSDM), 2012, pp. 643–652.

[32] L. Weng, F. Menczer, Y.-Y. Ahn, Virality prediction and community structure in social networks, Sci. Rep. 3 (2522) (2013).

[33] L. Weng, F. Menczer, Y.-Y. Ahn, Predicting successful memes using network and community structure, in: Proceedings of AAAI International Conference on Web and Social Media (ICWSM), 2014, pp. 535–544.

[34] S.-H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, H. Zha, Like like alike: joint friendship and interest propagation in social networks, in: Proceedings of ACM International Conference on World Wide Web (WWW), 2011, pp. 537–546.

[35] E. Yoo, W. Rand, M. Efekhar, E. Rabinovich, Evaluating information diffusion speed and its determinants in social media networks during humanitarian crises, J. Oper. Manage. 45 (2016) 123–133.

[36] J. Zhang, B. Liu, J. Tang, T. Chen, J. Li, Social influence locality for modeling retweeting behaviors, in: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), 2013, pp. 2761–2767.

[37] Q. Zhang, Y. Gong, Y. Guo, X. Huang, Retweet behavior prediction using hierarchical dirichlet process, in: Proceedings of AAAI International Conference on Artificial Intelligence (AAAI), 2015, pp. 403–409.

[38] Q. Zhao, M.A. Erdogdu, H.Y. He, A. Rajaraman, J. Leskovec, SEISMIC: a self-exciting point process model for predicting tweet popularity, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2015, pp. 1513–1522.

[39] D. Zhou, O. Bousquet, T. NavinLal, J. Weston, B. Scholkopf, Learning with local and global consistency, in: Proceedings of International Conference on Neural Information Processing Systems (NIPS), 2003, pp. 321–328.

[40] H. Zhu, X. Yin, J. Ma, W. Hu, Identifying the main paths of information diffusion in online social networks, Physica A 452 (2016) 320–328.