Predicting Trending Messages and Diffusion Participants in Microblogging Network

Jingwen Bian[†], Yang Yang[†,*], and Tat-Seng Chua[†]
[†]National University of Singapore, Singapore
[*]The University of Queensland, Australia
{bian_jingwen, chuats}@comp.nus.edu.sg, dlyyang@gmail.com

ABSTRACT

Microblogging services have emerged as an essential way to strengthen the communications among individuals. One of the most important features of microblog over traditional social networks is the extensive proliferation in information diffusion. As the outbreak of information diffusion often brings in valuable opportunities or devastating effects, it will be beneficial if a mechanism can be provided to predict whether a piece of information will become viral, and which part of the network will participate in propagating this information. In this work, we define three types of influences, namely, interest-oriented influence, social-oriented influence, and epidemic-oriented influence, that will affect a user's decision on whether to perform a diffusion action. We propose a diffusion-targeted influence model to differentiate and quantify various types of influence. Further we model the problem of diffusion prediction by factorizing a user's intention to transmit a microblog into these influences. The learned prediction model is then used to predict the future diffusion state of any new microblog. We conduct experiments on a real-world microblogging dataset to evaluate our method, and the results demonstrate the superiority of the proposed framework as compared to the state-of-the-art approaches.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences; H.1 [**Information Systems**]: Models and Principles—*Human information processiong*

General Terms

Algorithms, Experimentation

Keywords

Social network; Social influence analysis; Diffusion prediction

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609616 .

1. INTRODUCTION

Recent years have witnessed the development of microblogging services and the important role they played in changing the way people live and communicate. As a novel type of social network, microblog integrates the specialities of various other traditional social media. Similar to Facebook, people can use it to manage the interpersonal relationships with their friends and post updates about their daily activities to keep themselves socially active. Microblog also operates like news media, where bountiful official accounts (by ordinary users in some cases, such as the witness of an accident) keep publishing the latest and miscellaneous news about the society. Additionally, microblog also serves as an interest discovery tool, assisting people in exploring and disseminating the content conforming to their personal interests.

The extent to which a social network spreads information is a key measurement that impacts the degree of user engagement as well as its revenue. Unlike friendship-based networks, in which the content spread is confined to close group of friends, content-centric microblogging network promotes the spread of information with no restriction: it allows users to connect with any people sharing common interests with them, and repost any content they deem interesting. Therefore, we could observe plenty of diffusion actions (i.e., the action that a user disseminates a microblog from his friend instead of posting original content) in microblogging platforms, and the number of diffusion actions could in turn reflects the distinction of the associated microblog, in terms of its novelty, popularity and importance. It has been reported [9] that the size of information cascades fits power-law, which means that of all the microblogs posted everyday, only a tiny proportion of them will break out (i.e., a large number of users participate in propagating these microblogs), while most will diminish. It will be of great benefit if a trending microblog could be predicted early before the outbreak actually happens. For instance, if a user predicts that a piece of information will emerge in the near future and propagates this information in the early stage, then his ideas or comments will have more chances to be exposed to other people, which could help to increase his reputation and influence. In addition to predicting trending microblogs, the ability to predict which users may participate in the diffusion process of a particular microblog is also desirable and valuable in many cases, e.g., analysing the predicted future participants to choose better advertisement strategy, recommending a microblog to potential participants to accelerate the information propagation process, etc.

In recent years, information diffusion has drawn considerable research interest in computer science, and a variety of techniques and models have been developed to capture the information diffusion in online social networks [14]. Some researchers focus on building standard models to explain the general information diffusion process, such as the two seminal models, namely Independent Cascades (IC) model [10] and Linear Threshold (LT) model [16]. These models are useful for simulating the information flow in social networks. However, they cannot be directly applied to predict the diffusion process. Another research direction lies in detecting the outbreak of information cascades [20], which focuses on the cascades that have already broken out. In this work, we target at a different problem: given a new microblog, we intend to predict whether it will become trending in the near future, and we also try to predict which users will participate in the future diffusion process of this microblog.

A user's action of propagating a microblog from a friend may be affected by various factors. Generally, there are three factors that contribute to a diffusion action: 1) the content of the microblog is in accordance with this user's interest; 2) the microblog is posted by this user's close friend, and his repost action is due to social needs; and 3) the information is epidemic (e.g., a piece of breaking news), and his propagation action is a result of conformity behavior (i.e., the act of matching attitudes, beliefs, and behaviors to group norms [23]). These factors exhibit different types of influences exerted on a user from different sources: his friends, his interests, and the information content. It is difficult to quantify these influences due to several challenges. The first challenge is how to differentiate these influences. In our scenario, only the diffusion action is observable, while the underlying influences that trigger this action is implicit. Therefore, it is impractical to directly infer the degree of different influences based on the performed diffusion actions. Second, in order to obtain the interest-oriented influence, we need to generate the user's interest profile, i.e., what kind of content he is interested in, from his microblogging history. Nowadays, a growing proportion of microblogs contain multimedia information [3, 8], e.g., both texts and images, and images could provide more information than the short texts contained in a microblog. How to discover the interest profile of a user from these multimedia contents remains a problem. A third challenge is in constructing a unified model that can jointly leverage various types of influences to model and predict the information diffusion process.

To solve the proposed problem and tackle the above challenges, we propose a novel scheme to quantify the three types of influences and adopt the learned influences to model and predict users' diffusion actions. Specifically, the proposed framework comprises three essential stages: user interest profile learning, diffusion-targeted influence learning, and microblog diffusion modeling and prediction. First, in order to learn a user's interest profile, we need to map the user's microblogs into the corresponding interest categories. We devise a classification approach, termed *Multi-Task Transfer Learning*, to jointly classify the multimedia microblogs posted by a user into various interest categories. In order to address the deficiency of labeled training data, we bring in external knowledge domain where labeled samples are easy to acquire, and the transfer learning technique is adopted to project data samples from different domains into the same embedding space. Meanwhile, a multi-task co-learning process is integrated for the classification task, which will benefit from the joint information shared by different media contents. In the second part, we propose a diffusion-targeted influence model to quantify various influences a user receives. Three types of influences are formally defined, and a factor graph model is elaborated to categorize and analyse these influences. Finally, given the history diffusion action set, we learn how the various influences could affect a user's decision of whether to perform a particular diffusion action. The learned weighting configuration for various influences will eventually contribute to the prediction of future diffusion status with regard to a new microblog.

The rest of the paper is organized as follows. Related works are briefly summarized and discussed in Section 2. In Section 3, we describe the formal definition of the problem addressed in this paper, and give a brief overview of the whole framework. In Sections 4, 5 and 6, respectively, we introduce the three main stages of our framework. Section 7 presents the experimental design and evaluation results. Finally, we conclude our work in the last section.

2. RELATED WORK

Influence analysis. Social influence is the behavioral change of a person because of the perceived relationship with other people, organizations and society in general. It has been a widely accepted phenomenon for decades, and many works have been done to demonstrate the existence of social influence in online social networks [1, 19, 13]. One important research direction is the problem of influence maximization. Given a network with influence estimates, influence maximization tries to select an initial set of users such that they will eventually influence the largest number of users. Kempe *et al.* introduced a fundamental work [16]. Following this, many other methods [6, 12, 5, 27] have been proposed to improve the efficiency. All the related works discussed above assume the influence probability on the edges are given as input, which is impractical for real-world problems. Some works have been done to infer the degree of influence from a given social network [11, 24, 2]. A probabilistic model was proposed in [11] to learn influence probabilities by mining past influence cascades. Tang *et al.* studied the topic-level social influence in [24], and a Topical Affinity Propagation (TAP) method was proposed to model this problem. Other works include the detection of influential users [26], influence measurement in Twitter [4], etc.

Outbreak detection. The target of outbreak detection is to select a set of nodes from a social network in order to detect the spread of a virus as fast as possible. Leskovec *dt al.* presented a general methodology for near-optimal sensor placement in [20]. By exploiting submodularity they developed an efficient algorithm much faster than the greedy algorithm. The work in [18] conducted evolutionary analysis in blog networks, and showed that the blogspace had been expanding in metrics of community structure and connectedness. The goal of the above works is to detect existing outbreaks, which is different from our target of predicting the outbreak of a microblog diffusion process before it happens. Recently, Cui *et al.* [9] raised the question of cascading outbreak prediction. Based on the historical cascade data, a data driven approach was proposed to select important nodes as sensors. The prediction is based on the cascading behaviors of these sensors. Although the problem is similar to ours, the above method could only predict whether a

cascade will breakout, but could not provide more detailed information about the scale of the cascade or which of the users will participate in the future diffusion process. Besides, the limitation with the small number of sensors results in low recall in prediction performance. The models described in [22] and [23] aim at modeling and predicting users' social actions based on the past action history. However, since a model needs to be trained for each information diffusion process, and the training process requires a considerable number of actions, these models could not be adopted for the diffusion prediction of a relatively new microblog. Unlike these methods, our proposed framework could quantify general influence degree. The prediction model is trained with regards to the behavior of users while is not constrained to any specific diffusion process. Therefore, our model can handle new incoming microblogs without the need to train a new model every time.

3. PROBLEM DEFINITION AND FRAMEWORK OVERVIEW

We denote the social network as a directed graph $G = (V, E)$, where V is the user set and $E \in V \times V$ represents the social relationships between users. We denote a user u' as a friend of u if there is a edge $(u, u') \in E$, i.e., u is a follower of u'. The basic action of a user is to post microblogs, which can either be original or reposted from friends. A microblog m contains two multimedia components: the textual part m^t and the visual part m^v (either m^t or m^v could be empty). We denote M_u as the set of all microblogs the user u has posted, and the overall microblog set as $M = \cup_{u \in V} M_u$. Next, we present the formal definitions of some terms used in this paper.

Definition 1. **Interest Profile.** The microblogs could be related to various interest categories. We denote $\mathcal{C} = \{c_1, ..., c_{|\mathcal{C}|}\}$ as the collection of all interest categories. The interest profile of user u, $I(u)$, is a $|\mathcal{C}|$-dimensional vector which represents the user's interest distribution over all interest categories $\mathcal{C}$.

Definition 2. **Diffusion action.** The users in the social network interact with each other through reposting microblogs published by their friends. A diffusion action is defined as a triple $a = (u, u', m)$, representing the user u reposts a microblog with content m from his friend u'. Here, m can either be original microblog sent by u' or a microblog reposted by u' from his friend.

The input to our problem is the social network G, the past microblogging history of all users M, as well as the diffusion action set A which contains the past diffusion actions of all users. For a new incoming microblog m_{new}, we intend to predict: (1) whether it will become trending in the near future, and (2) which of the users will participate in the diffusion process of this microblog.

Figure 1 presents the overall flowchart of our proposed framework, which comprises three main stages: user interest profile learning, diffusion-targeted influence learning and microblog diffusion modeling. The first component aims to discover a user's interest profile based on his microblogging history. Specifically, each microblog is first classified into a interest category, then the interest profile is generated based on the aggregation of these classification results. In the second stage, we define three types of influences, namely, interest-oriented influence, social-oriented influence

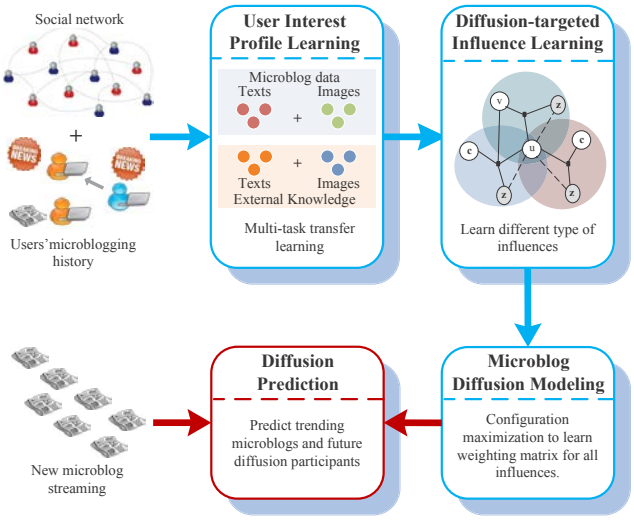

Figure 1: Overview of our framework.

and epidemic-oriented influence, which will affect a user's decision of whether to perform a diffusion action. We propose the diffusion-targeted social influence model to distinguish and quantify the degree of these three types of influences. Finally, in the third part, we take in the learned influences as factors, and analyze the weighting parameters of each influence in affecting a user's diffusion decision based on the diffusion action history. The learned weighting parameters are then adopted for predicting the trending microblog and the future diffusion activity of a new microblog.

4. USER INTEREST PROFILE LEARNING

Generally, a user will show different level of interest and possess different level of expertise for various interest categories, e.g., sports, music, history, etc. Therefore, the degree of influence he exerts to his friends, as well as his interest in propagating further information, will be different for various categories. As a result, learning the interest profile will be crucial for detailed analysis of a user's influence on his friends, his diffusion actions, and the prediction of his future action.

Given a user's historical microblogging data, the target of interest profile learning is to map the microblogs to the interest categories $\mathcal{C}$, which in essence is a microblog classification problem. Under the circumstance of microblogging, we are confronted with the following two problems:

- It is difficult to collect labeled training microblog data, as the data labeling task is tedious and expensive. Besides, as the users tend to talk about the latest trends, the contents of microblogs are highly dynamic and the data vocabulary changes continuously. Consequently, even if we can get some labeled data, they will quickly become out-dated over time.

- Microblog contains multimedia data which contains both text and image. Therefore, instead of traditional short-text classification task, the problem becomes cross-media classification where both textual and visual in-

formation should be incorporated to better capture the various aspects of a microblog.

In order to address the first problem, we propose to include external knowledge into the training process to assist the classification of microblogs into the related interest categories. The well-edited articles from portal website (such as Sina.com[1]) are chosen as the external knowledge with the following reasons: 1) the articles in portal website are well-categorized, which means we can directly get the interest category labels for these articles; 2) the contents of these articles cover nearly all aspects; and 3) these articles contain rich multimedia information, which is appropriate for our cross-media classification problem. We denote the external knowledge as $E = \{(e^t, e^v)\}$, where each data sample includes the textual content e^t and the visual content e^v. Besides, we also have the $|E| \times |\mathcal{C}|$ label matrix Y of the external knowledge, with each element $y_{ij} \in \{-1, 1\}$ indicating whether the i-th data sample belongs to the j-th interest category.

Although the microblog domain and the external knowledge domain are relevant, their data distribution are different, which makes it infeasible to directly use external data as training samples. Domain adaption is a solution to this problem [29]. Domain adaption aims at solving a learning problem in the target domain by utilizing training data in the source domain, allowing data from the both domains to be transferred to the same embedded space. Traditional domain adaption problems usually target at a single media type, while the problem in our scenario contains two modalities. One naive solution is to apply domain adaption techniques on each media type separately, and then train two unrelated classifiers for text and image. However, the contents of the two media types are not isolated and there is interrelationship between them. For example, the text and image contained in the same microblog data are usually related to the same topic. By applying the classification separately, these beneficial relationship will be ignored. With the above consideration, we propose the *Multi-Task Transfer Learning (MTTL)* model, which targets at the cross-media, domain-adaptive classification task.

4.1 The MTTL Model

Given the unlabeled microblog data M and the labeled external knowledge E, we target at jointly handling both the textual and visual classification tasks in microblogs. In each task, the external knowledge and microblog data need to be transferred to the same embedded space.

We first delineate two desirable properties for the transfer learning task, namely: 1) maximal alignment of distribution between the source and target domain data in the embedded space; and 2) preservation of the local geometry.

1) Objective 1: Distribution Matching. We employ transfer component analysis (TCA) [21] for transfer learning. Specifically, let the kernel matrices defined on the microblog domain, external knowledge domain, and cross-domain data in the embedded space be $K_{M,M}$, $K_{E,E}$ and $K_{M,E}$, respectively, and the kernel matrix defined on all the data be

$$K = \begin{bmatrix} K_{M,M} & K_{M,E} \\ K_{E,M} & K_{E,E} \end{bmatrix} \in \mathbb{R}^{(|M|+|E|)\times(|M|+|E|)} , \quad (1)$$

then TCA tackles the domain adaptation problem by minimizing the MMD distance between the two domains:

$$\min_{Q} tr(Q^T K L K Q) , \quad (2)$$

where $Q \in \mathbb{R}^{(|M|+|E|)\times d}$ is the embedding matrix; d is the dimensionality of the embedding space; and $L_{ij} = 1/|M|^2$ if both x_i and $x_j \in M$, $L_{ij} = 1/|E|^2$ if both x_i and $x_j \in E$, and $L_{ij} = -1/(|M| \times |E|)$ otherwise.

2) Objective 2: Locality Preserving. We would like to preserve the local structures of both the microblog and external knowledge data, i.e., if two data samples are close to each other in the original domain, this relationship should be preserved in the embedded space [28]. Let $\mathcal{G}$ be the k nearest neighbors graph of the original data with $g_{ij} = exp(-d_{ij}^2/\sigma^2)$ for $x_i, x_j \in M \cup E$ if x_i and x_j are in the same data domain and x_i belongs to the k nearest neighbor set of x_j, or vice versa, and $g_{ij} = 0$ otherwise. Let d_{ij} represents the distance of x_i and x_j, and the graph Laplacian matrix of $\mathcal{G}$ be $\mathcal{L}$. Note that after domain adaption using TCA, the data projection in the embedded space is $Q^T K$, where the i-th column $[Q^T K]_i$ provides the embedding coordinates of x_i. Hence, we minimize the following objective function for locality preserving:

$$\sum_{i,j} g_{ij} \left\| [Q^T K]_i - [Q^T K]_j \right\|^2 = tr(Q^T K \mathcal{L} K Q). \quad (3)$$

With the above two objectives, we are able to map the unlabeled microblog and the labeled external knowledge data into the same embedded space. In order to jointly learn both the textual and visual classifiers, we propose to utilize the following multi-task model to explores the intrinsic correlation:

$$\min_{\{Q_t, W_t, b_t\}|_{t=1}^2} \sum_{t=1}^{2} \|K_t Q_t W_t + \mathbf{1} b_t^T - Y_t\|_F^2 + \rho \|W\|_{2,1}$$
$$\text{s.t. } Q_t Q_t^T = I, \ t = 1, 2 \quad (4)$$

where $t = 1$ indicates the text classification task and $t = 2$ indicates the image classification task. $\{W_t, b_t\}$ are classification regression parameters. Q_t is the embedding matrix of the t-th task. The cross-media consistency is preserved by the $\ell_{2,1}$ regulation term $\|W\|_{2,1}$, where $W = [W_1, W_2]$ and $\|W\|_{2,1} = \sum_{j=1}^{d} \|w_j\|_2$ with w_j representing the j-th row of W.

Combining the three objectives in Eq.(2), (3) and (4), the final optimization problem for MTTL can be written as:

$$\min_{\{Q_t, W_t, b_t\}|_{t=1}^2} \sum_{t=1}^{2} (tr(Q_t^T A_t Q_t) + \mu \|K_t Q_t W_t + \mathbf{1} b_t^T - Y_t\|_F^2)$$
$$+ \rho \|W\|_{2,1}, \quad \text{s.t. } Q_t Q_t^T = I, \ t = 1, 2 \quad (5)$$

where $A_t = K_t L_t K_t + \delta K_t \mathcal{L}_t K_t$, and δ, μ and ρ are the balance parameters.

4.2 Optimization

The problem in Eq.(5) can be reformulated as

$$\min_{\{Q_t, W_t, b_t\}|_{t=1}^2} \sum_{t=1}^{2} (tr(Q_t^T A_t Q_t) + \mu \|K_t Q_t W_t + \mathbf{1} b_t^T - Y_t\|_F^2)$$
$$+ \rho tr(W^T S W) \quad (6)$$

[1] http://www.sina.com.cn/

where S is a diagonal matrix with $S_{jj} = \frac{1}{2\|w_j\|_2}$. We design the following iteration strategy which includes two steps:

Step 1: Keep Q_t fixed, and update W_t and b_t. By setting the derivative of Eq.(6) w.r.t. b_t to zero, we obtain

$$b_t = \frac{1}{n_t}(Y_t - K_t Q_t W_t)^T \mathbf{1}, \tag{7}$$

where n_t is the number of training samples in the t-th task. Then substituting the derived b_t into Eq.(6) and setting the derivative w.r.t. W_t to zero, we get

$$W_t = (U_t^T U_t + \frac{\gamma}{\mu} S)^{-1} U_t^T V_t \tag{8}$$

where $U_t = H_t K_t Q_t$, $V_t = H_t Y_t$, and $H_t = I - \frac{1}{n_t}\mathbf{1}\mathbf{1}^T$ is the centering matrix.

Step 2: We update Q_t by fixing W_t and b_t. With W_t and b_t fixed, the objective function in Eq.(6) is reduced to

$$\begin{aligned} &\min_{Q_t} tr(Q_t^T A_t Q_t) + \mu \|K_t Q_t W_t + \mathbf{1} b_t^T - Y_t\|_F^2 \\ &\text{s.t. } Q_t Q_t^T = I, \ t = 1, 2. \end{aligned} \tag{9}$$

This optimization problem can be efficiently solved by the algorithm introduced in [25].

The whole algorithm is summarized in Algorithm 1.

Algorithm 1 Muti-Task-Transfer-Learning (MTTL)

Input:
Microblog data M, external knowledge data E and the label matrix for external data Y.
Output:
Transformation matrix Q_t and regression parameters W_t and b_t, for both text task (t=1) and image task (t=2).

```
Construct the kernel matrix K based on Eq.(1), the MMD matrix L, Laplacian matrix ℒ.
for t = 1 to 2 do
    Initialize Q_t, W_t and b_t;
end for
repeat
    Update the matrix S;
    for t = 1 to 2 do
        Update b_t according to Eq.(7);
        Update W_t according to Eq.(8);
        Solve Eq.(9) to update Q_t;
    end for
until convergence
```

The learned transformation matrix and regression parameters could be adopted to classify new microblogs. By denoting k^m as the kernel vector of microblog m, the classification output of m is $l_t(m) = W_t^T Q_t^T k_t^m + b_t$, $t = 1, 2$, then the corresponding interest category of m, $C(m)$, is the category with largest classification output in either textual or visual domain. The interest profile of the user u is then defined as:

$$I(u)_c = \frac{|\{m \in M_u | C(m) = c\}|}{|M_u|}, \ c = 1, 2, ..., |\mathcal{C}|.$$

5. DIFFUSION-TARGETED INFLUENCE LEARNING

Generally, the reposting action of a user may be affected by the following three types of influences:

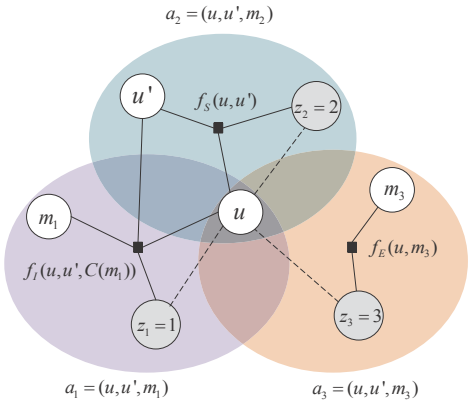

Figure 2: Graphical representation of the diffusion-targeted influence model. Three example diffusion actions of user u from the friend u' are shown as examples.

- **Interest-oriented Influence**: A user is likely to repost a microblog if the content is interesting to him. Consider a friend u' of u. If u' shares similar interests with u, or if u' is an expert for the interest category that u is interested in, then the friend u' is likely to have high influence on the user u.
- **Social-oriented Influence**: In this case, the repost action is based purely on the needs of social interaction. In other words, the action is triggered by the social influence exerted by his friend, instead of the information contained in this microblog.
- **Epidemic-oriented Influence**: If a microblog is epidemic (e.g., breaking news), it will be very probable to result in a repost action. In this situation, the influence is highly related to the epidemic-degree of this information rather than the content or the friend who post it.

Formally, let $F_I(u, u', c)$ denote the interest-oriented influence receive by user u from his friend u' related to the interest category c; $F_S(u, u')$ the social-oriented influence receives by u from u'; and $F_E(u, m)$ the epidemic-oriented influence of the microblog m. A user may receive different degree of influences from different friends according to either the closeness of their social connection, the friend's authority in the corresponding interest category, or simply the epidemic-degree of the disseminated information. In the following subsection, we elaborate a diffusion-targeted influence model, which efficiently differentiates and quantifies different types of influences.

5.1 Diffusion-targeted Influence Model

Consider a diffusion action $a = (u, u', m) \in A$. As aforementioned, this action could be triggered by three types of influences. We introduce a latent variable z, which indicates the source of the influence that leads to the diffusion action a. More precisely, $z = k, k \in \{1, 2, 3\}$ indicates that a is caused by the k-th type of influence. In order to infer the latent influence-indicator z for each diffusion action, as well as to quantify the degree of F_I, F_S and F_E, we propose the

Diffusion-targeted Influence Model. Figure 2 shows an illustrative example of three diffusion actions between user u and his friend u'. Our model comprises the following three feature functions which models the three types of influences F_I, F_S and F_E, respectively.

- **Interest-related feature function** $f_I(u, u', c)$, which contains two factors. The first factor is defined as the ratio between the number of microblogs that u reposted from u' in the interest category c and the total number of microblogs that belongs to interest category c posted by u'. The second factor refers to the weight of the interest category c in the interest profile of user u'. Intuitively, a higher weight represents a higher authority level in the corresponding interest category, which should cause larger influence. This function is defined as follows:

$$f_I(u, u', c) = \frac{|\{a = (u, u', m)|C(m) = c \wedge z_a = 1\}|}{|\{m \in M_{u'}|C(m) = c\}|} \times I(u')_c \ .$$

- **Social-related feature function** $f_S(u, u')$, which is defined as the ratio between the number of actions that u diffuses a microblog from u', over the total number of microblogs belonging to u':

$$f_S(u, u') = \frac{|\{a = (u, u', \cdot)|z_a = 2\}|}{|M_{u'}|} \ .$$

- **Epidemic-related feature function** $f_E(m)$, which is defined as the ratio between the number of friends who repost microblog m, over the total number of friends of user u:

$$f_E(m) = \frac{|\{a = (\cdot, u', m)|u' \in (N)(u)\}|}{|\mathcal{N}(u)|} \ ,$$

  where $\mathcal{N}(u)$ denotes the friends of user u.

Typically, the target of this influence model is to best fit (reconstruct) the observation data, which is usually achieved by maximizing the likelihood function. With these feature functions, we define the objective likelihood function as:

$$\begin{aligned} P(Z) = & \frac{1}{R} \prod_{(u,u',m) \in A, z=1} f_I(u, u', C(m)) \\ & \times \prod_{(u,u',\cdot) \in A, z=2} f_S(u, u') \times \prod_{(\cdot,\cdot,m) \in A, z=3} f_E(m) \end{aligned} \tag{10}$$

where $Z = \{z_1, z_2, ..., z_{|A|}\}$ represents the hidden variables corresponding to all the actions in A, and R is a normalization factor. Figure 2 describes an illustration of this factorization. Each feature function (denoted in black box) is connected to the corresponding variables.

5.2 Model Learning

We intend to find the optimal parameter configuration that maximizes the objective function in Eq.(10). We propose to use the sum-product algorithm [17] to infer the latent variables. Two update rules are defined, one for message sent from variable node to factor node:

$$\mu_{z \to f}(z) = \prod_{f' \sim z \backslash f} \mu_{f' \to z}(z) \ ,$$

and one for message sent from factor node to variable node:

$$\mu_{f \to z}(z) = \sum_{\sim \{z\}} \left(f(Z) \prod_{z' \sim f \backslash z} \mu_{z' \to f}(z') \right) \ ,$$

where μ is the passed message; $f' \sim z \backslash f$ represents that f' is a neighbor node of the variable z on the factor graph except factor f; $z' \sim f \backslash z$ indicates that z' is a neighbor node of the factor f on the factor graph except variable z, and $\sim \{z\}$ represents all the variables in Z except z.

After the learning process, the interest-oriented influence $F_I(u, u', c)$, social-oriented influence $F_S(u, u')$ and epidemic-oriented influence $F_E(u, m)$ could be achieved by calculating $f_I(u, u', c)$, $f_S(u, u')$ and $f_E(u, m)$, respectively.

6. MICROBLOG DIFFUSION MODELING AND PREDICTION

After learning the influence between the users and their friends, our next target is to utilize these influences for microblog diffusion analysis and prediction. Let $h \in \{-1, 1\}$ indicates whether an action (u, u', m) is actually performed, i.e., whether user u reposts the microblog m of his friend u'. We maximize the conditional probability of user actions given the input social network G and history action set $\mathcal{A}$, i.e., $P_\theta(H|G, \mathcal{A})$. More precisely, for each action in $\mathcal{A}$, we construct a training instance. We target at finding the optimal parameter θ^* to maximize the joint conditional probability for all the actions.

Note that the diffusion action set A used for training in Section 5 contains only those performed actions. As the task here is to predict whether a user will perform a diffusion action, we also include unperformed diffusion actions as negative samples into the training set $\mathcal{A}$. Suppose u' post a microblog m at time $t_{u'}$, and u does not repost this microblog. Then we add the unperformed actions (u, u', m) into $\mathcal{A}$ if only $t_u - t_{u'} < \Delta$, where Δ is the threshold time interval, and t_u is the activation time stamp of the user u, i.e., u performs certain activity at the time. The underlying reasons for choosing these unperformed actions is as follows. If the interval between the posting time of a microblog post and the activation time is too large, then unperformed diffusion action may probably because the user misses the corresponding microblog. On the contrary, if the microblog is presented within the time duration $(t_u - \Delta, t_u)$ and this user does not repost this microblog, then we have good reason to believe that the unperformed diffusion action is actually caused by the lack of influence, and thus is suitable to be included to the training set.

In order to maximize the probability $P(H|G, \mathcal{A})$, we factorize the global probability as the product of several local factor functions. We adopt the influences learned in the previous stage as the input factors, and learn the weighting parameters. Integrating all the factors together, we obtain the following log-likelihood objective function:

$$\begin{aligned} \mathcal{O}(\theta) &= \log P_\theta(H|G, \mathcal{A}) \\ &= \sum_{i,j,d} \left(\sum_{a_k = (u_i, u_j, m) \in \mathcal{A} \wedge C(m) = c} \alpha_{ijc} g(h_k, F_I(u_i, u_j, c)) \right) \\ &\quad + \sum_{ij} \left(\sum_{a_k = (u_i, u_j, \cdot) \in \mathcal{A}} \beta_{ij} g(h_k, F_S(u_i, u_j)) \right) \\ &\quad + \sum_{i} \left(\sum_{a_k = (u_i, \cdot, m) \in \mathcal{A}} \gamma_i g(h_k, F_E(u_i, m)) \right) - \log R \end{aligned} \tag{11}$$

where $g(h_k, F(\cdot))$ acts as the feature functions to link the factors to the corresponding variables, which is defined as

$$g(h_k, F(\cdot)) = \begin{cases} F(\cdot) \text{ , if } h_k = 1, \\ 1 - F(\cdot) \text{ , if } h_k = 0. \end{cases} \quad (12)$$

α, β and γ are the factor weights, and R is a normalization factor which ensures that the distribution is normalized with the sum of the probabilities equals to 1. With the function defined in Eq.(11), the objective of the training process is to estimate an optimial parameter configuration of $\theta^* = \{\alpha^*, \beta^*, \gamma^*\}$ from the training set $\mathcal{A}$ that maximizes $\mathcal{O}(\theta)$. The learning process contains two steps: 1) compute the gradient for each parameter; and 2) optimize all parameters with gradient descents. Specifically, we first approximate the marginal distribution $P_\theta(h_k|G, \mathcal{A})$. With the marginal probabilities, the gradient of a parameter can be obtained by summing over all the corresponding factor functions. Next, we use a gradient descent method to solve the above problem.

Diffusion Prediction. Given a new microblog m_{new}, and the action set A_{new} consisting of all existing diffusion actions related to m_{new}, the learned influences and weighting parameters can be used to predict the future participants in disseminating this new microblog. In practice, it is meaningless to do prediction for every new microblog since only a small portion will finally break out according to the power-law of information cascades [9]. Therefore, we devise certain criteria for starting the monitoring and prediction, e.g., we delay the prediction until the number of existing diffusion actions $|A_{new}|$ exceeds some minimum number N_{delay}.

The diffusion of microblogs through the social network fits the Independent Cascade (IC) model [16]. IC starts with an initial set of active nodes, and the process unfolds in discrete steps according to the following rule: when a node becomes active, it is given a single chance to activate each currently inactive neighbor. If it succeeds, the corresponding neighbor will become active and follow this rule to activate more neighbors. But whether or not this node succeeds, it cannot make further attempts to active its neighbors in the subsequent rounds. This process runs until no more activations are possible. Following the IC model, for each active user (user that has already performed the diffusion action) u and each of his friend u', we predict whether the action $a_{new} = (u, u', m_{new})$ will be performed by predicting the corresponding indicator h according to:

$$\begin{aligned} h^* &= \arg\max_h \log P(h|G, \mathcal{A}) \\ &= \arg\max_h \Big(\alpha_{ijC(m_{new})} g(h, F_I(u_i, u_j, C(m_{new}))) \\ &\quad + \beta_{ij} g(h, F_S(u_i, u_j)) + \gamma_i g(h, F_E(u_i, m_{new})) \Big). \end{aligned}$$

The above prediction process simply assumes that the delay in receiving the information will not affect the diffusion action. In other words, no matter how long the message is received after the original posting time, the user will make the same diffusion decision. However, this assumption will not hold under the microblogging circumstance, where people have more intention in reposting fresh microblogs, and the outbreak of a microblog usually happens during a relatively short time period. To handle this problem, we propose to incorporate a time decay factor to the feature functions in Eq.(12) as:

$$g(h_k, F(\cdot)) = \begin{cases} \lambda^l F(\cdot) \text{ , if } h_k = 1 \\ 1 - \lambda^l F(\cdot) \text{ , if } h_k = 0 \end{cases}$$

where $0 < \lambda < 1$ is the decay parameter and l is the length of the diffusion path when information reaches the predicted user. This new feature function penalizes long pathes.

With the above prediction process, we are able to predict the future hotness of a microblog in terms of its estimated reposting number, as well as the users who will participate in the diffusion process.

7. EXPERIMENTS

In this section, we present the experimental results for evaluating our proposed approach.

7.1 Dataset and Experimental Settings

We conduct the experiments on a real-world dataset collected from Tencent Weibo[2], one of the largest microblogging platforms in China. We crawled a network with around 2.62 million users and all the microblogs posted by them from July 1st to August 31th in 2013, which gives rise to a total number of 192.3 million microblogs. We could observe a very high percentage of diffusion actions in the collected dataset, in which 63% of these microblogs are reposted from friends. The statistics of this dataset is shown in Table 1. In this experiment we focus on the original microblogs, and predict how they will be diffused through this social network. In order to estimate the diffusion lifetime of a microblog, we calculated the average duration between the time a microblog is originally posted to the time that the repost number of this microblog reaches 90% of the total repost number. The average time is less than 4 days, which means most of the diffusion actions are performed within 4 days after a microblog is posted. According to this observation, we divided our dataset into two parts, the training set with microblogs posted in the first 50 days, and the testing set with microblogs posted in the following 8 days, while we exclude the original microblogs posted in the last 4 days from the testing set, as the diffusion process of these microblogs may not have been finished and could not provide a valid groundtruth for our evaluation. For the external knowledge required in the MTTL classification model, we crawled 0.65 million articles (with 0.83 million images) from 20 categories[3] from Sina.com[4].

Before the evaluation, we first pre-processed the texts and images. Texts were firstly segmented, then stop words, low-frequency words, mentions and urls were removed from the text vocabulary. For visual feature extraction, scale-invariant feature transform (SIFT) descriptors were first extracted from each image. We then trained a code book of 1,000 visual words. With the trained codebook, each descriptor was quantized into a visual word. Each image was further represented as a 1,000-D bag-of-visual-words feature. The parameters in MTTL were empiricially set as follows: $\sigma = 1$, $\rho = 0.1$, $\mu = 0.1$, $\delta = 0.01$. The threshold time interval Δ is set to 10 minutes.

One important parameter which will influence the experiment performance is the decay factor λ, which reflects the

[2]http://t.qq.com/

[3]The 20 categories are listed in the Appendix.

[4]http://www.sina.com.cn/

Table 1: Statistics of our dataset.

	#Microblogs	#Original	#Repost	#Image	#Days
Whole dataset	192.3m	71.5m	120.9m	129.2m	62
Training set	154.7m	60.3m	94.5m	103.9m	50
Testing set	25.3m	9.4m	15.9m	9.7m	8

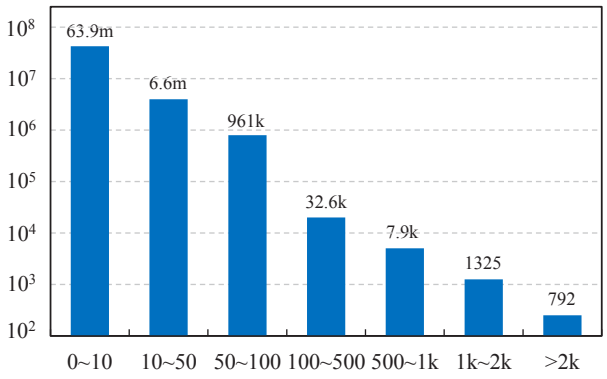

Figure 3: The distribution of repost number for all microblogs in our dataset.

simulation for time delay, i.e., the time duration between the initial posting of a microblog and the time that it reaches the user. Figure 4 shows the influence of λ in affecting the performance of predicting trending microblogs in terms of F1 value (refer to the following subsection for experiment details). As we can see, a large λ fails to provide enough decay ability, and a small λ causes too many potential diffusion actions to be rejected. Therefore, we adopt the optimal value of 0.95 for λ.

7.2 Predicting Trending Microblogs

Our first task is to predict whether a new coming microblog will become trending in the near future. Figure 3 shows the power-law distribution of the repost number in our dataset. We empirically define the trending microblog as one with repost number exceeding 1,000. This results in 168 trending microblogs in the testing set. We also randomly selected 832 non-trending microblogs with more than 200 repost number from the testing set. The prediction performance is measured for these 1,000 microblogs. We repeated the random selection 20 times, and the average result is reported. We compare our approach (denoted as TMP) to the following state-of-the-art methods:

- OSLOR [9]: The orthogonal sparse logistic regression model is a data-driven approach for cascading outbreak prediction. This method selects a set of nodes as sensors, and predicts the outbreaks based on the cascading behaviors of these sensors.

- ETL [7]: In the hot emerging topic learner, features are proposed for outbreak training and prediction. Although this method was originally used for emerging topic prediction, the proposed features and learning methods can also be applied to our task.

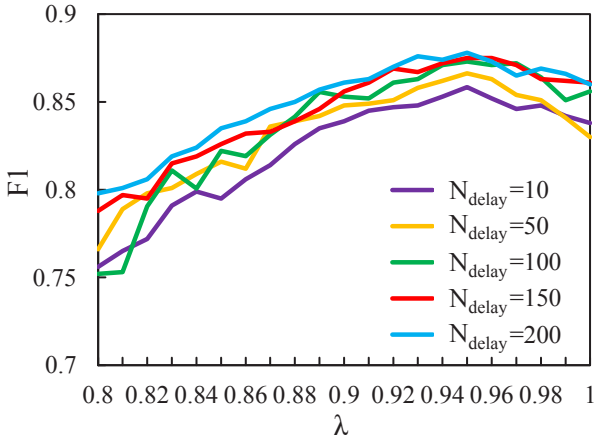

Figure 4: Influence of λ, in terms of F1 value of predicting trending microblogs.

- PMP [15]: This method is able to predict popular messages in Twitter. Similarly, several features are defined and a multi-class classification method is adopted to predict the volume of retweets.

Furthermore, we also design two comparing methods by replacing the diffusion-targeted influence model of our framework with other influence learning model, and build the prediction model based the new types of influence. Specifically, the following influence models are adopted:

- TFG [24]: The topical factor graph model targets at quantifying the topic-level social influence between each pair of users.

- IPL [11]: The influence probabilities learning method adopts a probabilistic approached to assign each pair of users an influence probability.

The delaying number N_{delay} (refer to Section 6) is set in the range of $\{10, 50, 100, 150, 200\}$. Precision, recall and F1 score are used as the evaluation measures. The results for our proposed TMP and the comparing methods are presented in Figure 5. From this figure, we have the following observations. 1) In terms of F1 measurement, TMP significantly outperforms the comparing methods. Larger delaying number will benefit the prediction performance as more information about the diffusion process is available. 2) OSLOR achieves slightly better performance on precision as compared to TMP, however, the recall performance of OSLOR is far worse than that of TMP. OSLOR is designed to only monitor the most influential users, who are probable to trigger many reposting actions while inevitably less likely to participate in many outbreaking diffusion processes. In contrary, instead of the global influential measurement, our method models the local influences for each user and takes more factors into consideration in the prediction procedure, leading to more comprehensive results. 3) In general, the influence based methods, i.e., TMP, TFG and IPL, perform better than feature based methods, i.e., ETL and PMP. This is because those simple features defined on the small number of early participating users do not possess sufficient prediction ability. 4) By comparing TMP with the other two influence based methods, i.e., TFG and IPL, the performance

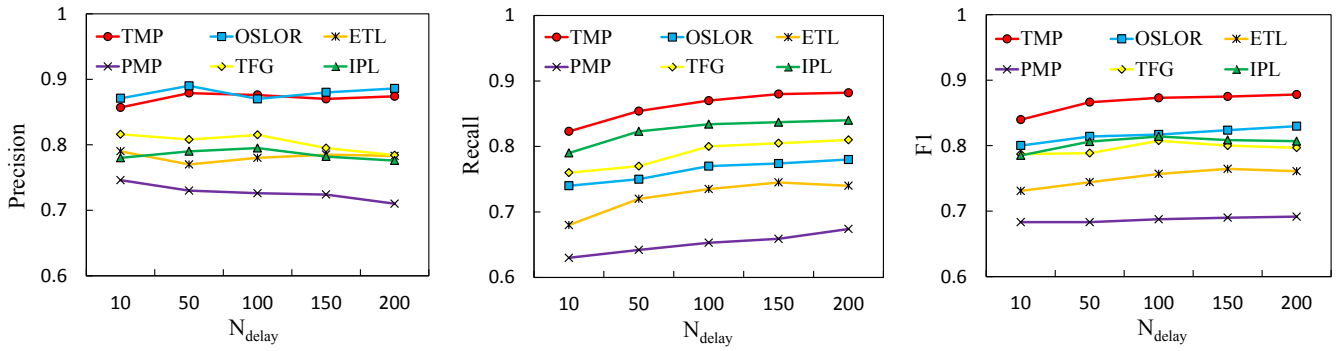

Figure 5: The results of predicting trending microblogs.

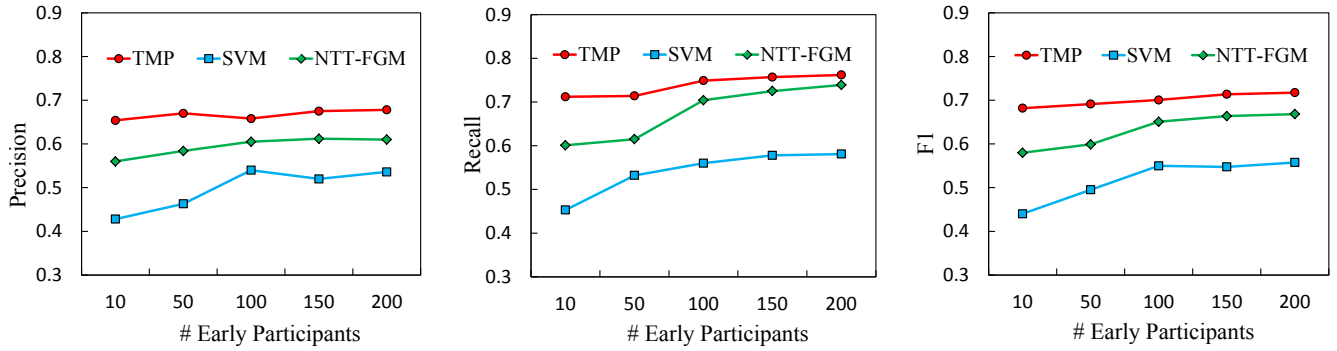

Figure 6: The results of predicting future diffusion participants.

improvement demonstrates the effectiveness of our proposed influence model. While our proposed TMP characterizes the network in a more comprehensive way, the other two influence models only model either the interest-related influence or pair-wise influence among users.

7.3 Predicting Diffusion Participants

The second task targets at predicting which users will participate in propagating a particular microblog. We compare our proposed TMP with the following methods:

- SVM: It uses the associated interest profile of users as well as the states of their neighbors to train a microblog classifier, which is then employed to predict the user actions.
- NTT-FGM [22]: The noise tolerant time-varying factor graph approach simultaneously models social network structure, user attributes and user action history for predicting the users' future actions.

Both of the two comparing methods need to train a prediction model, which requires a sufficiently large number of positive training samples to achieve satisfactory performance. We randomly select 1,000 microblogs whose final repost numbers exceed 200, and we want to predict all the diffusion participants when the retweet number of this microblog (denoted as # early participants) reaches 10, 50, 100, 150 and 200, respectively. The average prediction results for these 1,000 microblogs are presented in Figure 6. The results demonstrate the superiority of our proposed method under all evaluation measurements. In addition, our method also shows more stable prediction performance over different early participant numbers. The underlying reason is that our training procedure does not depend on these early diffusion activities. On the other hand, the two comparing methods need to train a prediction model for each microblog diffusion process, and the performance heavily relies on the number of training samples, i.e., the early participants.

7.4 Component Contribution Analysis

In this part, we evaluate the effects of the three types of influences defined in our framework, namely, interest-oriented influence, social-oriented influence and epidemic-oriented influence; while the interest-oriented influence is closely related to the result of user interest profile learning. Specifically, the following cases are evaluated: 1) TMP-P, which replaces our interest profile learning component with LDA to generate topics and infer user interest profile. 2) TMP-I,TMP-S and TMP-E, which removes the interest-oriented influence, social-oriented influence and epidemic-oriented influence from our framework, respectively. The results of predicting trending microblogs in terms of F1 measurement are shown in Figure 7. From the figure, we can see that the performance drops significantly by removing any of the proposed components. Specially, epidemic-oriented influence has the strongest correlation with the diffusion process of trending microblogs, hence removing it will decrease the prediction performance the most. Besides, by replacing our proposed interest profile learning component with other profile learning method, the performance decreases, which demonstrates the effectiveness of our model.

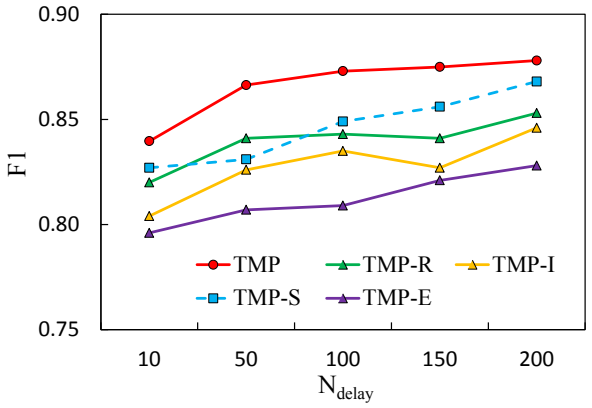

Figure 7: Effects of different components, in terms of F1 value of predicting trending microblogs.

8. CONCLUSION

In this work, we presented a novel approach for predicting the trending microblogs and the subsequent diffusion actions in microblogging network. Specifically, we defined three types of influence, namely, interest-oriented influence, social-oriented influence, and epidemic-oriented influence, which jointly determine the user diffusion action. We devised a multi-task transfer learning model for identifying the interest categories of microblogs. A diffusion-targeted influence model was proposed for quantifying different types of influences. We formulated the diffusion prediction as a factorization of the user intention of reposting a microblog. Extensive experiments have been conducted on a real-world microblogging dataset to show the superiority of our proposed approach as compared to the state-of-the-art methods.

9. ACKNOWLEDGMENTS

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

10. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *SIGKDD*, pages 7–15, 2008.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *WSDM*, pages 65–74, 2011.

[3] J. Bian, Y. Yang, and T.-S. Chua. Multimedia summarization for trending topics in microblogs. In *CIKM*, pages 1807–1812, 2013.

[4] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10:10–17, 2010.

[5] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *SIGKDD*, pages 1029–1038, 2010.

[6] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *SIGKDD*, pages 199–208, 2009.

[7] Y. Chen, H. Amiri, Z. Li, and T.-S. Chua. Emerging topic detection for organizations from microblogs. In *SIGIR*, pages 43–52, 2013.

[8] T.-S. Chua, H. Luan, M. Sun, and S. Yang. Next: Nus-tsinghua center for extreme search of user-generated content. *IEEE MultiMedia*, 19(3):81–87, 2012.

[9] P. Cui, S. Jin, L. Yu, F. Wang, W. Zhu, and S. Yang. Cascading outbreak prediction in networks: a data-driven approach. In *SIGKDD*, pages 901–909, 2013.

[10] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[11] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *WSDM*, pages 241–250, 2010.

[12] A. Goyal, F. Bonchi, and L. V. Lakshmanan. A data-based approach to social influence maximization. *PVLDB*, 5(1):73–84, 2011.

[13] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW*, pages 491–501, 2004.

[14] A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks: A survey. *SIGMOD Record*, 42(2):17, 2013.

[15] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *WWW*, pages 57–58, 2011.

[16] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *SIGKDD*, pages 137–146, 2003.

[17] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *TIT*, 47(2):498–519, 2001.

[18] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *WWW*, 8(2):159–178, 2005.

[19] T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *WWW*, pages 601–610, 2010.

[20] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *SIGKDD*, pages 420–429, 2007.

[21] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *TNN*, 22(2):199–210, 2011.

[22] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In *SIGKDD*, pages 1049–1058, 2010.

[23] J. Tang, W. Sen, and J. Sun. Confluence: conformity influence in large social networks. In *SIGKDD*, pages 347–355, 2013.

[24] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *SIGKDD*, pages 807–816, 2009.

[25] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, pages 1–38, 2013.

[26] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270, 2010.

[27] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *WWW*, pages 981–990, 2010.

[28] Y. Yang, Y. Yang, Z. Huang, H. T. Shen, and F. Nie. Tag localization with spatial correlations and joint group sparsity. In *CVPR*, pages 881–888, 2011.

[29] Y. Yang, Y. Yang, and H. T. Shen. Effective transfer tagging from image to video. *TOMCCAP*, 9(2):14, 2013.

APPENDIX

The 20 interest categories adopted in this work are: *Military, Society, Stocks, Sports, Constellation, Automobile, Variety, History, Fashion, Health, Entertainment, Travel, Apps, Phone, Technology, Household, Education, Amusement, Movie, Music.*