Modeling Information Diffusion over Social Networks

Dong Li, Shengping Zhang, Xin Sun, Huiyu Zhou, Sheng Li, and Xuelong Li, *Fellow, IEEE*

Abstract—Modeling the process of information diffusion is a challenging problem. Although numerous attempts have been made in order to solve this problem, very few studies are actually able to simulate and predict temporal dynamics of the diffusion process. In this paper, we propose a novel information diffusion model, namely GT model, which treats the nodes of a network as intelligent and rational agents and then calculates their corresponding payoffs, given different choices to make strategic decisions. By introducing time-related payoffs based on the diffusion data, the proposed GT model can be used to predict whether or not the user's behaviors will occur in a specific time interval. The user's payoff can be divided into two parts: social payoff from the user's social contacts and preference payoff from the user's idiosyncratic preference. We here exploit the global influence of the user and the social influence between any two users to accurately calculate the social payoff. In addition, we develop a new method of presenting social influence that can fully capture the temporal dynamics of social influence. Experimental results from two different datasets, Sina Weibo and Flickr, demonstrate the rationality and effectiveness of the proposed prediction method with different evaluation metrics.

Index Terms—Information diffusion, social network, modeling, prediction

1 Introduction

In recent years, with the rapid development and demanding requirements of online social networks [1], [2] (*e.g.*, Twitter, Facebook, Flickr), tremendous interests have arisen from the study of information diffusion. An example of information diffusion is: When someone adopts a piece of information, his or her neighbors may be influenced and then consider adopting the same information. Usually, information diffusion is caused by user actions, for example, users perform re-tweeting actions to diffuse tweets on Twitter. Therefore, information diffusion also can be regarded as user action diffusion.

Diffusion models have been used to explain and simulate how information is diffused over social networks. They have a wide range of applications, including viral marketing and breaking news detection [3], [4], [5], [6]. According to the use of data, we can categorise the existing studies on information diffusion modeling into two groups: theory-centric models and data-centric models. Theory-centric models are mainly used in epidemiology, sociology and economics. Widely studied diffusion models include epidemic model, independent cascade model, linear threshold model and the variants of these models. Using these models, some researchers investigated the challenging influence maximization problem [7], [8], [9], [10], which is related to viral marketing, and some explored the relationship between information diffusion and social network structures [11], [12]. Although theory-centric models provide partial solutions to the information diffusion problem, they also have inevitable shortcomings. For example, theory-centric models usually make use of randomly distributed parameters that are not learned from actual diffusion data. In addition, these models cannot work in real-time. These shortcomings prevent theory-centric models from being able to correctly predict the process of information diffusion.

In contrast to theory-centric models, data-centric models are usually learned from actual information diffusion data and can be categorised into macro- and micro-models. Macro-models, also called cascade generation models [13], [14], [15], [16], can generate diffusion cascades whose macro properties are similar to those of actual diffusion cascades. From the perspective of prediction, they also have the same limitation as theory-centric models. This limitation may be overcome by micro-models, which can predict whether or not a certain user in a social network can be activated by the information. The information

- *D. Li is with the School of Electrical and Information Engineering, Shandong University, Weihai, China and with Harbin Institute of Technology, China. Email: studyhibernate@163.com.*
- *S. Zhang is with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China. Email: s.zhang@hit.edu.cn. S. Zhang is the corresponding author.*
- *X. Sun is with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China. Email: sunxintyc@163.com.*
- *H. Zhou is with the Centre for Secure Information Technologies (CSIT), Queen's University Belfast, United Kingdom. Email: h.zhou@ecit.qub.ac.uk.*
- *S. Li is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. Email: lisheng@hit.edu.cn.*
- *X. Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China. Email: xuelong_li@opt.ac.cn.*

diffusion process is triggered by user actions, and therefore information diffusion prediction is actually for user behavior prediction. Micro-models hold the ability to predict diffusion processes, unfortunately, most of these models ignore the important fact that information diffusion fundamentally is a temporal dynamic process. Diffusion models can predict not only whether or not the user will perform the behavior but also when the user will perform the behavior.

In this paper, we propose a novel model (*i.e.* GT model) for information diffusion prediction based on actual diffusion data. Comparing to previous work, the advantages of our model are mainly in two aspects: (1) The proposed GT model considers the nodes of a network as autonomous, intelligent and rational agents. By introducing the time-related payoffs, the proposed GT model can simultaneously capture the user's own preference and his/her neighbors' behaviors, and realize the prediction that whether or not a users behaviors will occur in a specific time interval. (2) The proposed model can be learned efficiently using real diffusion data, therefore the proposed GT model is a highly scalable diffusion prediction system.

In summary, we make the following contributions in this research:

- We propose a new information diffusion model named as the GT model. In this model, users calculate the corresponding payoffs of different choices to make an appropriate decision of choosing which behavior to perform. This user payoff is composed of time-related social payoff from the user's social contacts, and preference payoff from his idiosyncratic preferences which are not time-related. Our diffusion model considers both two types of payoffs simultaneously in the process.
- We develop a method that jointly exploits the global influence of users and the social influence between users to calculate time-related social payoffs in the GT model based on actual diffusion data. This social payoffs consider both the individual properties and strategic interactions of the interacting users in a social network. Moreover, we use the similarity between information contents and user profiles to determine the user's preference payoff.
- We propose to use a non-negative vector with length K to represent social influence between users. This method not only takes the time series into account but also calculates a more accurate influence strength based on the statistics of the past diffusion.
- We evaluate the proposed predicting model against two real-life data sets from Sina WeiBo and Flickr. By testing the two different global influence computing methods as model parameters, pagerank and diffusion cascade, we illustrate how the model parameters affect the systematic performance, which is used to rationalise the proposed GT model. Finally, the comparison results demonstrate the superiority of the proposed GT model over other state of the art models.

The rest of this paper is organized as follows: Section 2 shows related work. Section 3 presents the formulation of the problem. Section 4 reveals the proposed GT model and Section 5 presents the algorithms for learning and how we evaluate this new model. Section 6 presents the experimental results that demonstrate the effectiveness of our proposed methodology. Conclusions and future work are made in Section 7. A preliminary version of this paper was published in the Proceedings of 22nd ACM International Conference on Information and Knowledge Management [17].

2 Related work

In this section, we review the related work in two different aspects: cascade generation model and information diffusion prediction

Cascade generation model. This type of models aim to generate information cascades. These generated cascades maintain several properties of real cascades. Leskovec *et al.* [13] proposed a conceptual model that was quite similar to the epidemic model [18], [19]. They compared the generated cascades against the real cascades extracted from the post network and observed that these generated and real cascades could be matched in terms of cascade size and degree distributions. Liben-Nowell *et al.* [14] studied information spread on a global scale based on internet chain letters, and discovered that the structures of the diffusion trees were narrow and deep. Golub and Jackson [15] attempted to explain the structures observed in [14] using the Galton-Watson branching model [20]. Wang *et al.* [16] developed a stochastic branching model to demonstrate that the macroscopic structures of information propagating processes are largely independent of contextual information and can be well explained by a simple mechanism. Although these models are helpful for deeply understanding information diffusion, they are not capable of predicting.

Information diffusion prediction. Rodriguez *et al.* [21] attempted to infer diffusion process that has happened in the past to explain the observed data. Differently, in this paper, we aim to predict the future diffusion process based on current diffusion cascades and network structures. Yang *et al.* [22] explored how the repost behavior was impacted by such factors as users, messages and time. They also proposed a factor graph model to predict the retweet behaviors of the users based on the above important observations. R. Zaman *et al.* [23] made use of retweets as positive feedback and lack of retweets by followers in the retweet network as negative feedback to forecast retweet behaviors in Twitter. The relevant features for prediction were the tweeters and the retweeters. Fei [24] presented a multi-task learning algorithm with heterogeneous task relationships to address the problem of forecasting users' behaviors to their friends' postings in social networks. Liu [25] proposed a generative graphical model to estimate topic-level influence between nodes in the network, which utilized both the textual content related to each node and the heterogeneous link information. Furthermore, they studied how to leverage the mined topic-level influence to help the user behaviors prediction. Du *et al.* [26] estimated the influence between users in continuous-time diffusion networks via a randomized algorithm. Yang *et al.* [27] presented a Role-Aware information diffusion model which integrates social role recognition and diffusion mechanism into a unified framework. Chang *et al.* [28] attempted to predict the popularity of online serials with autoregressive models. Cheng *et al.* [29] found that the cascade growth becomes more predictable when we observe more of its reshares, and the structural and temporal features are key prediction factors. Hung *et al.* [30] generalize the diffusion prediction on novel topic problems to predict both cross-topic-observed and unobserved diffusions.

Tan *et al.* [31] proposed a noise tolerant time-varying factor graph model (NTT-FGM) to formalize the problem of social action tracking. They defined three factors to capture the intuitions discovered in observations and presented an efficient algorithm to learn the tracking model. Saito *et al.* [32] attempted to learn the diffusion probabilities of the independent cascade model [33] and linear threshold model [34] based on the real diffusion data. They first defined the above problem as a likelihood maximization problem, and then used an Expectation-Maximization (EM) algorithm to address the problem. Although these methods take into account time factors, they need substantial calculation time and cannot handle massive data. Goyal *et al* [35] proposed two time-dependent models, the CT model and DT model, for social influence calculation and applied them together with general threshold model to predict time-dependent information diffusion process. However, the prediction of information diffusion strongly relied on the users' activation thresholds, which are difficult to set in practice. Moreover, the approximate simulation mechanisms of the CT and DT models for social influence presentation also caused systematic performance to be degraded. In contrast to these systems, by fully considering all the interacting users to measure the time-related payoffs of different choices, our model can make better prediction and thus improve the performance dramatically.

3 Problem Formulation

In this section, we first give the essential definitions referred to in this work and then formalize the problem that we are going to address. A social network can be represented as $G = (V, E, T)$, where V is a set of $|V| = N$ users; and E is the set of edges: A directed/undirected edge $(u, v) \in E$ represents a social tie between user u and user v. Furthermore, T is a function labeling each edge with the time at which the social tie was created. Based on the network, we give the following definitions.

Definition 1 (Activate action)**.** *An activate action can be represented as a triple (u, a, t_u), which can be interpreted as that, user u is activated by a piece of information a at time t_u, or user u performs the action of adopting a piece of information a at time t_u.*

Let S_u be the set of information that user u adopts at all time. We denote the activate actions of all the users as the action log $\Omega = \{(u, a, t_u)\}$. Such an action log (also called information diffusion log) is available in many online systems. For example, on Twitter, the activate action (u, a, t_u) can be perceived as user u retweeting tweet a at time t_u.

Definition 2 (Information diffusion)**.** *We state that a piece of information a diffuses from user u to user v iff: (i) $(u, v) \in E$; (ii) $\exists (u, a, t_u), (v, a, t_v) \in \Omega$ with $t_u < t_v$; and (iii) $T(u, v) < t_u$. Once this is satisfied, we denote $diff(a, u, v, \Delta t)$, where $\Delta t = t_v - t_u$.*

Obviously, when we claim that a piece of information diffuses from user u to v, there must be a social tie between these users before they are activated by this information. Information diffusion over the social network leads to the natural notion of a diffusion cascade, defined as follows.

Definition 3 (Diffusion cascade)**.** *For each piece of information a, the diffusion cascade can be defined as $DC(a) = (V(a), E(a))$, where $V(a) = \{v | \exists t_v : (v, a, t_v) \in \Omega\}$ and $E(a) = \{\texttt{directed edge}(v_1, v_2) | diff(a, v_1, v_2, \Delta t)\}$.*

The diffusion cascade consists of users who are activated by a certain piece of information and edges connecting these users along the direction of propagation. When a user adopts a piece of information, s/he is activated or influenced. Once the user is activated, it becomes contagious and cannot be de-activated. The diffusion cascades in Definition 3 are with tree-like structures. For a node u, because it is possible that more than one parent of u may be activated and it is difficult to determine which parent really influences u to perform the action, we create links between all of the user's activated parents and u. Therefore, a node may have more than one parent. In addition, we can also develop other strategies to define a diffusion cascade. For example, we can create links only between a user's first or last parent and the current node u. In these strategies, a diffusion cascade is referred to as a real diffusion tree (i.e. each node has one parent). Next, we define the influence strength of a single user and that of two users.

Definition 4 (Global influence)**.** *Given a social network, $global_v$ is defined as the global influence of user v, which represents the influence strength of v over the whole network.*

Definition 5 (Social influence)**.** *Given two users u and v in a social network, we denote $social_{uv}(t)$ as the influence strength of user u on user v at time t.*

Note that $social_{uv}$ is not equal to $social_{vu}$ if the edge of the social network is directed. As the influence strength of user u on user v varies over time, introducing time variable t may lead to accurate descriptions of the influence strength between the users.

Global influence and social influence are fundamentally different. Global influence shows the authority, profession and popularity of a user in a social network, while social influence focuses on two interrelated users. If user a has two parents, users b and c, and even though the global influence of user b is stronger than that of user c, the social influence from b to a may still be smaller than that from c to a.

"Payoff" is a concept from game theory, which is a number that denotes the decision-making motive of a player in a game. In different scenarios, payoff can be in any quantifiable forms such as money or reputation of a player. Game theory is designed to address situations in which a player's decision depends not only on his/her personal preference, but also on the choices made by other players s/he is interacting with. In the information diffusion process, the behavior of a user is also related to both his/her preference and his/her neighbors' behaviors. Thus it can be seen that, the information diffusion process is very similar to the situations game theory simulates. Here, we bring the concept of "payoff" from game theory to model user behaviors in the diffusion process in social networks. In the diffusion process, "payoff" can be social relationships or followers, etc. In this paper, we propose a novel model (GT model) to predict information diffusion based on user payoff. One of the critical tasks is to calculate the user's time-related payoff resulting from her/his multiple choices, with which we can then predict the user's behavior as well as his acting time.

The payoff of a user can be divided into two parts: social payoff from the user's social contacts and individual payoff from her/his idiosyncratic preferences. In this paper, social payoff varies with time since a user may be bound with different payoffs when s/he adopts her/his friend's actions at different time, while preference payoff is not time-related. Based on the concepts discussed above, we present the following problems:

Problem 1 (Social payoff learning)**.** *Given a social network G and an action log Ω, the goal of our work is to learn the user's time-related payoff as a result of her/his various choices.*

Problem 2 (Preference payoff learning)**.** *Given a social network G and information set S_u released by the user, we intend to learn the user's preference payoff.*

4 The Proposed Model

We propose a novel method for information diffusion modeling through social networks. By introducing time series into the payoff calculation, the proposed model has the capability to predict the temporal dynamics of the information diffusion process. In our model, the diffusion process unfolds in discrete timesteps t, and begins from an initial active user set. When a user v observes a piece of information at time t, s/he calculates her/his payoffs depending on the neighbors' status to decide whether or not to adopt the information. If s/he adopts the information, her/his status becomes active at time $t+1$. We now describe the proposed model in more detail. For better illustration, Table 1 lists some mathematical symbols used in this paper.

When a user v spreads the same information as her/his neighbors do, s/he will get social payoff from the social contacts. The information itself also satisfies the idiosyncratic preference of user v, which brings v certain preference payoffs.

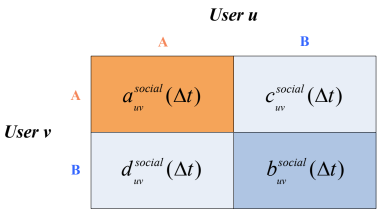
In a social network, we consider the simplest case in which each node has two possible choices, A and B, when the user observes a piece of information. To be concrete, the piece of information is a tweet, choice A is retweeting the tweet and choice B is not retweeting the tweet. We define $P^{soc}_A(v,t_v)$ as social payoff that user v may obtain when s/he chooses A at time t_v, and $P^{soc}_B(v,t_v)$ as social payoff that user v may obtain when s/he chooses B at time t_v. We also define $P^{pre}_{v,A}$ as the preference payoff that user v may hold when user v chooses A (adopting information i). User payoff is the combination of social payoff and preference payoff. $P_A(v,i,t_v)$ is defined as the user payoff of v when s/he chooses A at time t_v, and $P_B(v,i,t_v)$ is defined as the user payoff of user v when s/he chooses B at time t_v. The calculation of user payoffs of v in different scenarios is described as follows:

$$\begin{cases} P_A(v,i,t_v) = P^{soc}_A(v,t_v) + \beta * P^{pre}_{v,A} \\ P_B(v,i,t_v) = P^{soc}_B(v,t_v) \end{cases} \tag{1}$$

When user v adopts information i, we adopt a linear method (the first item in Eq. 1) to combine the computed social payoff and preference payoff in order to estimate the user payoff. β is a parameter for specifying the tradeoff between the two competitive payoffs. When user v does not adopt information i, s/he will not have any preference payoff. So user payoff is equivalent to social payoff in choice B. After sustaining payoffs in different choices, if $P_A(v,i,t_v) \geq P_B(v,i,t_v)$, user v will choose A at time t_v, otherwise

TABLE 1: Important mathematical symbols.

Symbols	Description
G, V, E	a social network G with node set V and edge set E
S_u	the set of information that user u adopts at all time
$diff(a, u, v, \Delta t)$	a item means information a spreads form user u to user v in time delay Δt
$P_A(v, i, t_v)$	the user payoff of v when s/he chooses A at time t_v
$P^{soc}_A(v, t_v)$	the social payoff that user v obtains when s/he chooses A at time t_v
$P^{pre}_{v,A}$	the preference payoff that user v obtains when user v chooses A
$a^{soc}_{uv}(\Delta t)$	the social influence user v gets from user u in time delay Δt
$global_u$	the global influence of user u
$social_{uv}(\Delta t)$	the social influence from user u to v in time delay Δt



		User u	
		A	B
User v	A	$a^{social}_{uv}(\Delta t)$	$c^{social}_{uv}(\Delta t)$
	B	$d^{social}_{uv}(\Delta t)$	$b^{social}_{uv}(\Delta t)$

Fig. 1: Payoff matrix of user v.

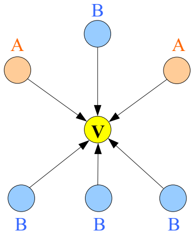


Fig. 2: User v makes choice between behaviors A and B, depending on the neighbors' actions.

choose B. In the next section, we will introduce how to calculate social payoff and preference payoff separately.

4.1 Social Payoff

For nodes u and v linked by an edge, we have several possibilities for them to combine together. The payoffs for user v are defined as follows:

If both u and v make choice A, v has payoff $a^{soc}_{uv}(\Delta t)$;

If both u and v make choice B, v has payoff $b^{soc}_{uv}(\Delta t)$;

If u makes choice A whilst v makes choice B, v gets payoff $c^{soc}_{uv}(\Delta t)$;

If u makes choice B whilst v makes choice A, v gets payoff $d^{soc}_{uv}(\Delta t)$.

Based on these choices made by user u and v, a payoff matrix of user v is generated and shown in Figure 1. $\Delta t = t_u - t_v$ denotes the time delay between the choices made by users u and v respectively. Here, we introduce time series into the calculation of a user's payoff for the first time. Therefore, in our proposed model, at different times, user v's responses to u's behaviors may result in different social payoffs.

Figure 1 illustrates a case of a single edge in the network. The total social payoffs are the sum of individual social payoffs generated when the user faces each player, as shown in Figure 2. Therefore, the choice of user v corresponds to all the choices made by all its neighbors.

A question arises: if some of the neighbors adopt choice A while others adopt B, how do we calculate social payoffs of user v in different choices? Obviously, this depends on the relative numbers of the neighbors with their choices as well as the social payoff matrix between v and each of its neighbors. Here, we denote $N_A(v)$ as the set of v's neighbors who adopt choice A and $N_B(v)$ as the set of neighbors who adopt B. If node v adopts choice A at time t_v, it will have the social payoff defined below:

$$\begin{aligned} P^{soc}_A(v, t_v) = & \Sigma_{u\in N_A(v)} a^{soc}_{uv}(t_v - t_u) + \\ & \Sigma_{u\in N_B(v)} c^{soc}_{uv}(t_v - t_u) \end{aligned} \tag{2}$$

Similarly, if node v adopts choice B at time t_v, the social payoff becomes:

$$\begin{aligned} P^{soc}_B(v, t_v) = & \Sigma_{u\in N_A(v)} d^{soc}_{uv}(t_v - t_u) + \\ & \Sigma_{u\in N_B(v)} b^{soc}_{uv}(t_v - t_u) \end{aligned} \tag{3}$$

Users performing behaviors (spreading different information) in different social networks will be of different types of social payoffs from their social relations. For example, if a user performs the behavior of joining in a community, the social payoff may include personal connections, and if a user performs the behavior of retweeting a tweet on Twitter, the social payoff will be additional followers. Therefore, it is very difficult to directly measure social payoffs of users from different behaviors due to the lack of a common ground. To solve this problem, we here present a novel method of calculating social payoffs, which is applicable to different diffusion information in different social networks. Specifically, we exploit both the global influence of individual users and the social influence between users to compute the payoffs of users' choices. Global influence shows the profession and authority of a user in particular fields while social influence reflects the degree of how one user has affected another one. Considering two users u and v in a social network with a social link between them, we anticipate that the more social payoffs that v has received following the choices of u in the history, the stronger intention that v will make the same choice as u at the present state.

This idea can be formulated as follows: the greater global influence the user u has and the greater social influence is shown between u and v, the more payoffs the user v will obtain from u if the former makes the same choice as the latter. For example, if Olivia is good at shopping and Jessica have been greatly influenced by Olivia in the past, Jessica is much more likely to purchase high quality goods (*i.e.* social payoff) when she makes the same choices as Olivia. A user may have different preference payoffs if s/he adopts the same information at different times. In other words, social payoff is time-related. Based on the description above, we define the social payoff matrix as

$$\begin{cases} c_{uv}^{soc}(\Delta t) = d_{uv}^{soc}(\Delta t) = 0 \\ a_{uv}^{soc}(\Delta t) = b_{uv}^{soc}(\Delta t) = global_u * social_{uv}(\Delta t) \end{cases} \quad (4)$$

where $global_u$ denotes the global influence of user u, and $social_{uv}(\Delta t)$ denotes the social influence between users u and v. If user v adopts different behaviors from u, s/he gets no social payoff; however, if user v adopts the same behavior as u does, the global influence and social influence are jointly exploited to measure the user payoff.

We use two methods in this work to measure the global influence of a single user in the social network. The first is the calculation of a pagerank value, which is based on the network topology structure analysis. Pagerank was originally used to analyze hyperlink networks to measure the importance of web documents [36]. Here, we apply it to the social network for influence calculation. The second method is based on the average size of diffusion cascades triggered by the user's adoption of information as the user's global influence, whose measurement is more direct compared to the pagerank method. A detailed discussion of how these two methods affect the performance of the GT model will be presented in Section 6.

Much effort has been invested in the study of social influence. However, only the CT models and DT models proposed by Goyal *et al.* [35] considered the time factor. The CT models describes the social influence by an exponential decay function. Despite of its simplicity, it has a significant drawback in that it assumes that all the social influence functions follow the same parametric form. The DT models set the influence of an active user u on her/his neighbor v at a constant value of $p_{u,v}$ after u performs the action within a time window of $\tau_{u,v}$, which is the average time delay of information diffusion from user u to v. After the time window $\tau_{u,v}$, the influence value presumably drops to 0. Due to the approximate simulation mechanism, it is difficult for the CT and DT models to capture the complex dynamics of social influence between users. In our work, we propose a novel method for accurate social influence calculation. We deliberately represent the social influence function as a non-negative vector with length K, where the kth component $social_{uv}(k)$ represents the social influence of user u on its neighbor v at time k. $social_{uv}(k)$ is used in Eq. 4 for calculating the social payoff. More precisely, we define:

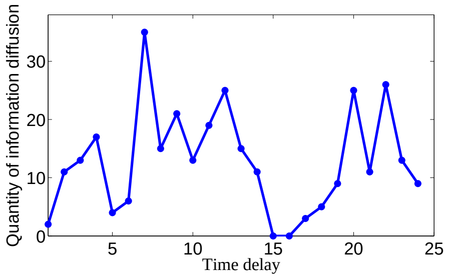


Fig. 3: An example of social influence between two users of Sina Weibo dataset.

$$social_{uv}(k) = \frac{|\{a|\exists \Delta t : diff(a,u,v,\Delta t) \bigwedge k-1 \le \Delta t \le k\}|}{|S_u|} \quad (5)$$

Figure 3 shows an example of social influence between two users in the Sina Weibo dataset, calculated by the proposed method. Compared to the CT and DT models, which use a simple exponential decay or a constant value for social influence, the proposed method can describe the temporal dynamics of social influence between two users. Here, we should highlight that, our GT model can use not only our proposed social influence calculation method but also other methods or improvements to our method.

4.2 Preference Payoff

Preference payoff results from the user's adopting information that satisfies her/his preferences. Thus, preference payoff is derived from the information itself and has nothing to do with the user's social relationships. Although the preferences of the user change over time, such changes are usually slow. In the current online social networks, information spreads quickly. Comparing the cycle of information diffusion with that of user preference's change, we can assume user preference to be static within the period of information diffusion. Thus, a user will hold the same preference payoff if s/he adopts the same information at different times. That is to say, preference payoff is not time-dependent.

In this paper, we consider the similarity between information content and user preference as the user's preference payoff. A major problem is how we profile information and the users. In this paper, we adopt a vector space model to construct individual profiles. Vector space model is simple but effective, and has been used widely in information retrieval and item recommendation.

For each user, we collect all the information adopted by each user as a document. Similarly, every piece of information also can be viewed as a document. For user u, it is profiled as a word vector, $u =< t_1 : w_{t_1}, t_2 : w_{t_2}, \ldots, t_n : w_{t_n} >$. Each weight w_{t_i} represents the degree of interest of user u over the period t_i. We use the standard TF-IDF method [37] to calculate the weights in the vector. For a piece of information i, it also can be profiled as $i =< t_1 : w_{t_1}^{'}, t_2 : w_{t_2}^{'}, \ldots, t_n : w_{t_n}^{'} >$. After having calculated the profiles of the user and her/his information, we compute the cosine similarity between these profiles, which is used as the preference payoff.

Finally, we highlight that in this section we only refer to two choices A and B for one single piece of information to introduce the proposed GT model for the purpose of simplicity. In fact, the GT model is not only applicable to the situation of two choices, but also can be used to deal with the situation of many choices for multiple pieces of information. For example, the information of *HTC phone* and *iPhone* are propagated in a social network. For an usual user, s/he may have three possible choices: One is *HTC phone*, one is *iPhone* and neither of the two options. The proposed GT model is also applicable to this more complicated case.

5 Algorithms

In this section, we present the algorithms for learning the parameters of the proposed GT model, and predicting information diffusion based on the GT model. Since we have presented our major concept in Section 4, we here focus on learning global and social influence, which are used to calculate social payoffs. To start with the depiction, we suppose that the inputs consist of a social network and an action log.

5.1 Learning Algorithms

5.1.1 Global Influence

In our work, we learn the global influence of individual users from two perspectives: social network's topology structure and properties of diffusion cascades. How to calculate the global influence of users is not the problem we intend to solve in this paper. Here, we adopt two popular methods to estimate user global influence. Any other similar method (*e.g.* outdegree, betweenness) can also be considered as the measure of global influences.

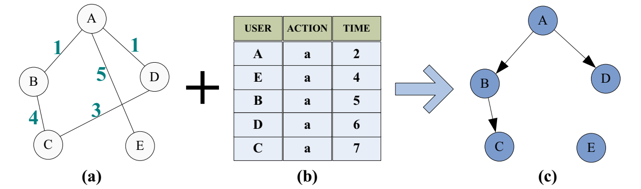
Pagerank Algorithm. Google uses the established pagerank algorithm [36] to calculate the importance of Web pages purely based on the link structure of the World Wide Web to improve its search results. Here, we apply pagerank to the social networks for influence calculation. In a social network, a node can be described as a user, and each directed/undirected edge can be associated with a relation map between two users. The standard pagerank algorithm has been exhaustively discussed in the literature, so we omit their technical details in this paper.

Diffusion Cascades. Diffusion cascades triggered by the information adopted by a user clearly indicate the user's global influence. Therefore, it is reasonable to link the properties of diffusion cascades with the measurements of the user's global influence. To do this, we first mine the information diffusion cascades based on the social network and the action log. In our work, we assume that a user is activated only once so each node in the diffusion cascades only has one parent and the diffusion cascades are treated as diffusion trees. Algorithm 1 illustrates how we mine diffusion cascades from a social network and the action log. The algorithm includes a triplet with information a in a data structure of $action\_table$.

Algorithm 1 Illustration of how diffusion cascades mine.

```
1:  For each piece of information $a$ do
2:    $action\_table = \emptyset$;
3:    For each user triplet $(v, a, t_v)$ in time order do
4:      For each user $u$: $(u, a, t_u) \in action\_table$
5:        If $(u, v) \in E$ && $T(u, v) < t_u$ do
6:          Add $diff(a, u, v, \Delta t)$ to diffusion cascade of $a$;
7:          Break;
8:        End
9:      End
10:     Add $(v, a, t_v)$ to $action\_table$;
11:   End
12: End
13: return Diffusion cascades;
```

For a particular piece of information a, we consider the users activated by this information individually in a chronological order. For the currently considered user v, Lines 4-9 are to find user u who diffuses the information to user v from the users who have been considered already and included in the action table. It is clear that these users in the action table have been activated before user v. The conditions of the diffusion action that occurs from users u to v are that there must be an edge from u to v in the social network, which should be created earlier than when user u is activated by information a. This is because if there is no social tie between u and v when u adopts the information, v would not be able to retrieve the information from u, and therefore the information is impossibly diffused from u to v. Once the conditions are satisfied, we save this information diffusion for user v and then "break" in order to consider the user next to v. The "break"



USER	ACTION	TIME
A	a	2
E	a	4
B	a	5
D	a	6
C	a	7

Fig. 4: Illustration of mining information diffusion cascades based on social networks and action logs: (a) shows a social network containing 5 users and 5 links labeled with corresponding time, (b) presents the action log of these five users, and (c) is the corresponding diffusion tree of information a.

operation in Line 7 ensures that for each user, we only search for one user who diffuses the information to the concerned user, so that each node in the diffusion cascade has at most one parent. If the $action\_table$ is defined as a stack, then when we search the table, the parent we seek will be the last activated user, while if the $action\_table$ is defined as a list, the parent we pursue will be the first activated user. Without this break operation, we would obtain a diffusion cascade for each action, unlike a diffusion tree, where one node may have more than one parent.

Figure 4 illustrates an example of mining the information diffusion cascades based on the social network and the action log. Figure. 4(a) shows a social network containing five users A, B, C, D and E with five links between them. The links are labeled with the times when two users established their relation. The action log of these five users is presented in Figure. 4(b). By implementing Algorithm 1, we can obtain the corresponding information diffusion tree as shown in Figure. 4(c). The edges in this tree are labeled with the time delay of the information diffused from the parent node to the child node. Note that even although both A and E adopted information a, there is no edge created from A to E in the diffusion tree because the link between A and E had not been created when A adopts the information at time 2 and it is impossible for E to observe A's actions at that time. On the other hand, although the edge between D and C is created before both of them adopt the same information, there is no edge from D to C in the diffusion tree. Thus, we obtain B as the parent of C and therefore quit the parent searching process for node C.

Here, we show a real diffusion cascade triggered by a microblog in Sina Weibo in Figure 5. The red spot in Figure 5 represents the user u who releases the information. The information is reposted by the followers of user u, and is reposted by the followers of the followers of user u, so on and so forth. Eventually, the diffusion cascade (*i.e.* Figure 5) is formed. Algorithm 1 enables us to obtain the diffusion trees of all the information adopted by a particular user. Then, Algorithm 2 applies these diffusion trees to the calculation of the user's global influence.

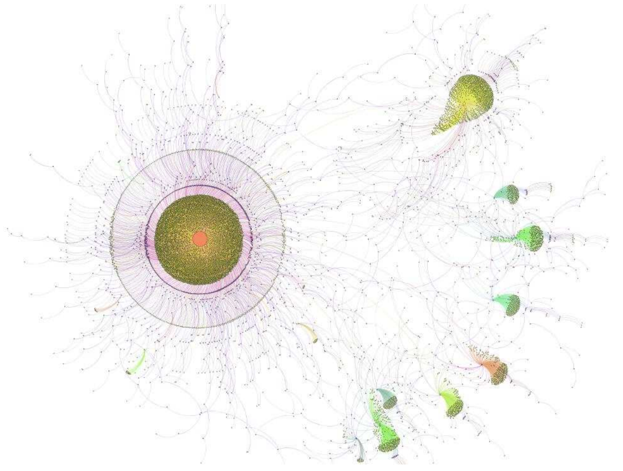


Fig. 5: Illustration of a diffusion cascade triggered by a microblog in Sina Weibo dataset.

Algorithm 2 Global influence computed using diffusion cascades.

```
For each user $u$ do
    $global_u = active\_person_u = 0$;
    For each information $a$ adopted by user $u$ do
        Add $a$ to $A_u$;
        For each user pair $(p, q) : \exists diff(a, p, q, \Delta t) \in$ diffusion cascades of $u$
            $active\_person_u + +$;
        End
    End
    If $A_u = \emptyset$ Then $global_u = 0$;
    Else $global_u = active\_person_u / |A_u|$;
End
```

In Algorithm 2, the average size of the diffusion cascades caused by a user's actions is used as a measurement of the user's global influence. $active\_person_u$ denotes the number of the users whose activation is derived from user u and is normalized by $|A_u|$ to obtain the average size of the diffusion trees caused by u.

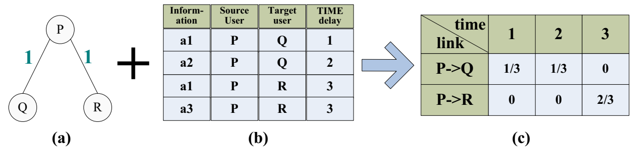
5.1.2 Social Influence

In this work, we present a novel method to calculate the social influence between two users. We consider the influence function as a non-negative vector with length K:

$$(social_{uv}(1), social_{uv}(2), \cdots, social_{uv}(K))$$

where the kth component $social_{uv}(k)$ presents the social influence of user u on its neighbor v at time k. K is the maximum diffusion time delay, and after this time, the influence value will drop to zero. For parameter K, different values are adopted for different datasets using statistical methods that will be described in detail in Section 6.

Algorithm 3 describes how we calculate $social_{uv}(k)$ based on the diffusion cascades mined from the in-



Inform-ation	Source User	Target user	TIME delay
a1	P	Q	1
a2	P	Q	2
a1	P	R	3
a3	P	R	3

time / link	1	2	3
P->Q	1/3	1/3	0
P->R	0	0	2/3

Fig. 6: Social influence calculation based on information diffusion statistics. (a) A social network containing 3 users and 2 links labeled with their time; (b) Information diffusion recorded in diffusion cascades; (c) Social influence between two users.

formation data using Algorithm 1. Lines 1-5 of the algorithm illustrate the initialization process, where $count_{uv}(k)$ denotes the amount of information diffused from user u to v at time k and is calculated according to the diffusion cascades in Lines 6-8. Finally, we use $|A_u|$ to normalize $count_{uv}(k)$ in order to obtain the social influence in Line 9-13. By considering the time series and implementing the calculation based on the past diffusion statistics, the proposed method can accurately capture the temporal dynamics of the social influence between users and thus produce more reliable prediction.

Algorithm 3 Social influence's calculation.

```
1:  For each social link $(u, v)$ do
2:      For $k = 1$ to $K$ do
3:          $social_{uv}(k) = count_{uv}(k) = 0$;
4:      End
5:  End
6:  For each $diff(a, u, v, \Delta t) \in$ diffusion cascades do
7:      If $k - 1 < \Delta t < k$ Then $count_{uv}(k) + +$;
8:  End
9:  For each social link $(u, v)$ do
10:     For $k = 1$ to $K$ do
11:         $social_{uv}(k) = count_{uv}(k)/|A_u|$;
12:     End
13: End
```

Figure 6 shows an example of social influence calculation using the proposed method. Figure. 6(a) shows a social network containing 3 users and 2 links labeled with their time of creation. The table shown in Figure. 6(b) records the information diffused from the source user to the target user with a time delay, for example, the first row can be explained as information a propagated from user P to Q with a time delay of 1. In this case, it is obvious that in total user P diffused three pieces of information ($a1$, $a2$, $a3$), resulting in $|A_P| = 3$. At the time delay of 1, there is one piece of information, $a1$, diffused from P to Q, leading to $count_{PQ}(1) = 1$. Meanwhile, the social influence between P and Q at time delay 1, $social_{PQ}(1) = 1/3$, is shown in Figure. 6(c). At the time delay of 2, there is one piece of information, $a2$, diffused from P to Q, and therefore $count_{PQ}(2) = 1$ while $social_{PQ}(2) = 1/3$. At the time delay of 3, there are two pieces of information, $a1$ and $a3$, diffused from P to R. Thus $count_{PR}(3) = 2$, while $social_{PR}(3) = 2/3$.

5.2 Information Prediction Algorithm

Based on the GT model proposed in the previous section, we here present an algorithm for predicting the information diffusion process. Our goal is to predict which node in the social network will be activated at time t. Then, we adopt three metrics to evaluate the prediction performance.

The prediction algorithm based on the GT model is presented in Algorithm 4. This algorithm focuses on the question of whether a user will perform an action at time t. For a user u, if s/he has performed the action, we claim that u is active; if s/he has not performed the action but at least one of his neighbors does, we claim that u is inactive. Here, we should note that if u does not perform the action and none of its neighbors does, u is not inactive. The *prediction part* of Algorithm 4 is to predict the user's actions at time t. We compare the payoff of node v assuming s/he will perform the action influenced by all the active neighbors (Lines 8-11) with certain payoff, assuming that v will not perform the action influenced by all the inactive neighbors (Lines 12-15). k in Lines 9 and 13 give the time delays. For an active user u, t^u_{active} denotes the time when u performs the action, and for an inactive user u, we define his time $t^u_{inactive}$ by the latest time when one of his neighbors performed the action. We construct the profile of information calculating preference payoff (lines 17-18). Line 19 is used to combine the social payoff and the preference payoff. If user v does not adopt information a, s/he will not get any preference payoff. Thus, we only add preference payoff to $payoff_A(v, i, t)$ but not to $payoff_B(v, i, t)$. If we omit lines 19-20 of Algorithm 4, this algorithm will become a prediction algorithm only based on social payoff. The following *results statistics part* in Algorithm 4 reveals the evaluation of this prediction method.

Finally, we claim that our prediction algorithm still works when the prediction time t is larger than the parameter K which is the size of social influence vector. For user u and its neighbor v, the condition $t > K$ can be divided into two situations: (1) $t - t^u_{active} \leq K$ or $t - t^u_{inactive} \leq K$. In this situation, based on Algorithm 4, the user u still will be counted to predict the behavior of user v. (2) $t - t^u_{active} > K$ or $t - t^u_{inactive} > K$. In this situation, the user u can not be used to predict the behavior of user v. However, the probability that this situation occurs is small because that the parameter K is decided based on real dataset analysis such that most of time delays between two users' behaviors are less than K. Therefore, our method still can predict one user's behavior based on most of its neighbors' behaviors when $t > K$.

Algorithm 4 Information diffusion prediction.

```
For each information $i$ in testing dataset (adoping information $i$ as choice $A$, otherwise as choice $B$) do
  Initialization: $TP = FN = FP = TN = 0$;
  For each inactive user $v$
    //prediction part
    $P_A^{soc}(v,t) = 0$;
    $P_B^{soc}(v,t) = 0$;
    For each link related with $v$, (u,v) do
      If $u$ is active do
        $k = [t - t^u_{active}]$;
        $P_A^{soc}(v,t) = P_A^{soc}(v,t) + social_{uv}(k) * personal_u$;
      End
      If $u$ is inactive do
        $k = [t - t^u_{inactive}]$;
        $P_B^{soc}(v,t) = P_B^{social}(v,t) + social_{uv}(k) * personal_u$;
      End
    End
    Construct profiles of information $i$ and user $v$;
    Get user $v$'s preference payoff $P^{pre}_{v,A}$;
    $P_A(v,i,t) = P_A^{soc}(v,t) + \beta * P^{pre}_{v,A}$;
    $P_B(v,i,t) = P_B^{soc}(v,t)$;
    If $P_A(v,i,t) \geq P_B(v,i,t)$
    Then $v$ is active;
    Else $v$ is inactive;
    //results statistics part
    If $v$'s real status is active && the prediction result is active Then $TP++$;
    If $v$'s real status is active && the prediction result is inactive Then $FN++$;
    If $v$'s real status is inactive && the prediction result is active Then $FP++$;
    If $v$'s real status is inactive && the prediction result is inactive Then $TN++$;
  End
End
```

Evaluation. We adopt three measurements to evaluate the proposed prediction method, which are Precision, Recall and F1-Measure. Precision is the ratio of the number of the predicted active users that are also activated actually to the total number of predicted active users. Recall is the ratio of the number of the predicted active users that are also activated actually to the total number of users who are activated actually. F1-Measure is the harmonic mean of precision and recall. The specific calculation formulas of the three measurements are as following:

$$\begin{aligned} &Precision = TP/(TP+FP) \\ &Recall = TP/(TP+FN) \\ &F1 = 2 * Precision * Recall/(Precision + Recall) \end{aligned} \tag{6}$$

where TP stands for true positive, FN false negatives, FP false positive and TN true negative. In our problem setting, we ignore the cases that none of the user's neighbors is active and we only consider the users for whom at least one neighbor is activated by the information before him. Under this presupposition, we denote TP as the number of users who are actually activated by the information and our prediction method gives the same predicting results, FP as the number of users who are in fact not activated by the information but are predicted to be activated, TN as the number of users who are actually not activated by the information and our model also predicts them to be not activated, and FN as the number of users who are activated by the information in reality but not in the model's prediction (see the *results statistics part* of Algorithm 4).

6 Experiments

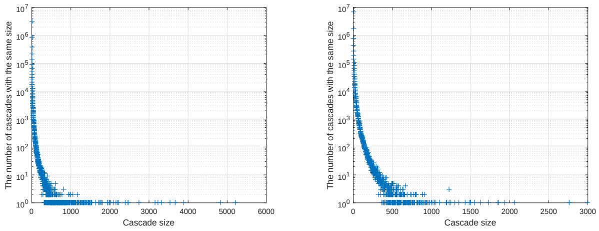
The GT model proposed in this work for information diffusion prediction can be generally applied to different types of social networks. In this section, we present various experiments to evaluate the rationality and effectiveness of our model.

6.1 Experimental Setup

Datasets. Given the social network and action logs as inputs, we evaluate the proposed GT model against two different genres of real-world datasets: Sina Weibo and Flickr.

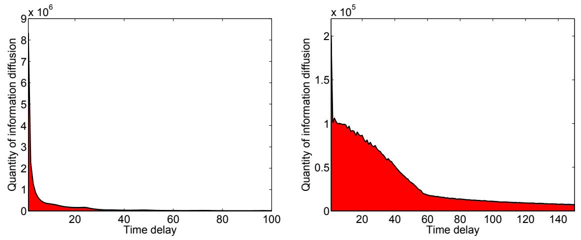
- **Sina Weibo.** This dataset is crawled from Sina Weibo by an open API interface. First, we use the snowball sampling technique to obtain a set of quality users, who mutually form a reasonably large connected component. Specifically, we select 5 seed users related to the Internet field and collect the friends of these seed users and further to the friends' friends. We ultimately obtained 251,639 users and 4,359,915 edges in the social network. Second, we collect approximately 30 million microblogs published by the 251,639 users from 11/07/2011 to 11/28/2011.
- **Flickr.** This dataset was collected and used in the work of [38]. The authors crawled the Flickr social network once per day for the period of 104 consecutive days between November 2 and December 3 in 2006 and February 3 and May 18 in 2007. They collected 2.5 million users and 33 million links in total. They also collected user behaviors, consisting of approximately 34 million favorite-markings over 11 million distinct photos.

Figure 7 presents the distributions of diffusion cascade sizes in two datasets. The long-tail shapes of two distributions mean that the size of most diffusion cascades is small, and only a small number of information get a large diffusion scale. We split each dataset into a training dataset and a testing dataset



(a) Sina Weibo dataset (b) Flickr dataset

Fig. 7: The distributions of diffusion cascade sizes on Sina Weibo and Flickr datasets.



(a) Sina Weibo dataset (b) Flickr dataset

Fig. 8: Information diffusion quantity distributions over time delay for the Sina Weibo and Flickr datasets, respectively.

according to the available information. The training dataset is used to learn the proposed model, while the testing dataset is used for evaluation.

In the Sina Weibo dataset, the diffusion information refers to microblogs. If a user posts or forwards a microblog, this user is activated by the microblog. Considering the properties of the Sina Weibo dataset, we set one hour as a time step. In the Flickr dataset, the diffusion information refers to photos. If a user posts or marks a photo as his/her favorite, then this user is activated by the photo. We observe that the speed of information diffusion in Flickr is much slower than that in Sina Weibo, and therefore set three days as the Flicker's time step.

Parameter K is the size of non-negative vector used to present social influence. From algorithm 4, the larger the parameter K, the higher the prediction accuracy, but the more running time; the smaller the parameter K, the lower the prediction efficiency, but the less running time. Thus it can be seen that, K is an important parameter for the balance between prediction accuracy and efficiency. How to determine the size K of non-negative vector used to present social influence is a key problem. In our experiments, we set the value of parameter K based on the statistic analysis on real diffusion data.

For each dataset, the maximum diffusion time delay K will adopt different values in the social influence function. Figure 8 shows the distributions of information diffusion quantity over a time delay in the Sina Weibo and Flickr datasets respectively, both of which have a long-tail shape. In the Sina Weibo dataset, 81.5% of diffusion actions are performed with a time delay of less than 24 hours, and hence we set the parameter K to 24. In the Flickr dataset, 85.0% of diffusion actions are performed with a time delay of less than 90 days, and we set the parameter K to 30.

We make use of information released or reposted by user u to calculate the preference payoff in Sina Weibo dataset. Because the Flickr dataset does not contain any text information about users, here we adopt an alternative solution where users and photo are both profiled by user ID vectors. For a photo i, we adopt the IDs of users who have marked the photo as favorite to profile photo i. Each user usually makes several photos, and each photo corresponds to a profile consisting of the user ID. For a user u, we collect all IDs of users who marked the photos marked by user u to profile user u. After obtaining the profiles of users and profiles, the cosine similarity between these profiles, is used as the preference payoff.

Comparison methods. Here, we compare the proposed GT model against the most similar one in the literature [35], where two time-dependent models, CT model and DT model, are presented for capturing social influence (or influence probability) and applied together with the general threshold model to predict time-dependent information diffusion. Since the CT model provides better prediction performance than the DT model in [35], to show the improvement, we compare our proposed method against the method of combining the CT model with the generalised threshold model. In addition, in our model, we adopt two methods with different parameter setups, pagerank and diffusion cascades, to calculate the user's global influence. This experiment is designed to illustrate how the accuracy of model parameter affects the model performance, thus proving the rationality of our proposed model. In addition, we compare the method using both social payoff and preference payoff with the method using social payoff only. Specifically, The methods evaluated and compared in our experiments are as follows:

- **Method 1:** baseline method combining the CT model [35] with the general threshold model for prediction.
- **Method 2:** in the GT model, we use pagerank to estimate global influence and only make use of social payoff for prediction.
- **Method 3:** in the GT model, we use diffusion cascades to estimate global influence and only make use of social payoff for prediction.
- **Method 4:** in the GT model, we use pagerank to estimate global influence and combine social payoff and preference payoff together for prediction.
- **Method 5:** in the GT model, we use diffusion cascades to estimate global influence and combine social payoff and preference payoff together for prediction.

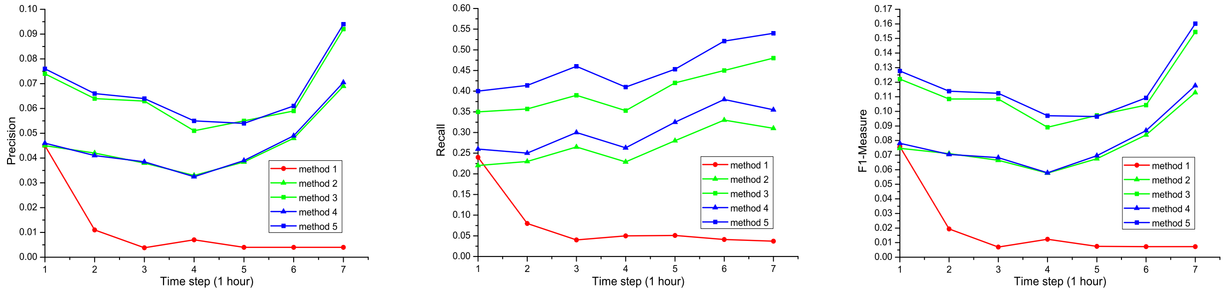


Fig. 9: Prediction performance using different approaches and metrics on the Sina Weibo dataset at the microscopic view.

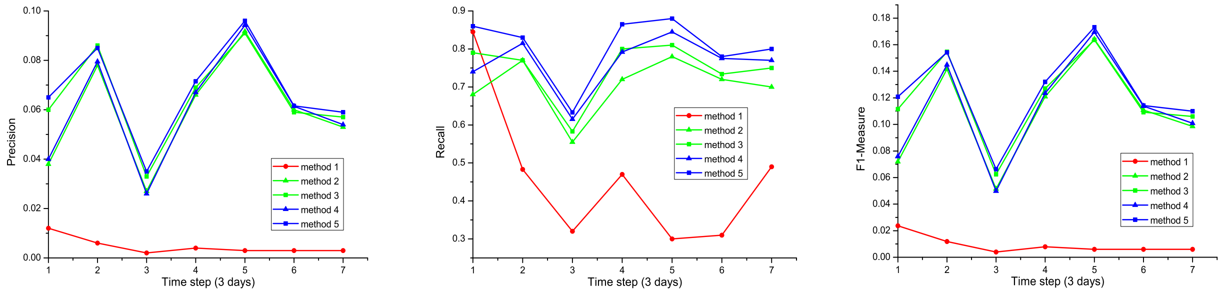


Fig. 10: Prediction performance using different approaches and metrics on the Flickr dataset at the microscopic view.

6.2 Experiments results

Prediction Performance. For each piece of information a in the testing dataset, given its diffusion progress before time t (from time 0 to $t-1$), our goal is to predict which node will be activated at time t.

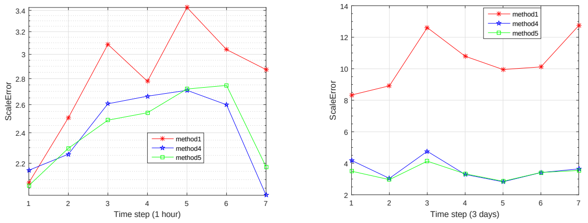
Figures 9 and 10 show the prediction performance of these different approaches with different metrics at 7 time steps (times 2-8) for the Sina Weibo and Flickr datasets, respectively. In these two figures, red curves (method 1) present the results of the baseline method, green curves (method 2 and method 3) present the results of the GT model using social payoff, and blue curves (methods 4 and 5) present the results of the GT model using both social payoff and preference payoff. The curves with triangles (methods 2 and 4) present the results of the GT model adopting pagerank values as global influence, and the curves with squares (methods 3 and 5) present the results of the GT model adopting the property of diffusion cascades as global influence.

From figures 9 and 10, we find that the proposed GT model (methods 2–5) consistently outperforms method 1 [35] that combines the CT model and the general threshold model. The prediction of information diffusion in [35] highly depends on the activation threshold of users, which is hard to set up. The same activation threshold value is assigned for all the users, but different users actually have different activation thresholds. Therefore, the prediction performance of [35] is relatively poor. In contrast, our model, which strategically considers all the interactive users and preference of users, improves the prediction performance dramatically.

Figures 9 and 10 present that the GT model that uses the diffusion cascades method for global influence calculation achieves better prediction performance than the pagerank method. This is mainly because that the pagerank method only analyzes the topology structure of the network, while diffusion cascades are mined from both the network structure and user behaviors. Thus, the diffusion cascades method can provide more accurate influence values than the pagerank method. These results demonstrate that when our model has more accurate parameters, it performs better in the prediction task.

Moreover, the blues curves with triangles or squares are in higher positions than the green curves with the same symbols in Figures. 9 and 10. This indicates that preference payoff is supportive to information diffusion prediction. We also notice that preference payoff helps recall more than precision. This is because that preference payoff is only added to active payoffs but is not added to inactive payoffs. The GT model predicts that more users will adopt information. This causes the values of TP and FP to increase but the value of FN does not change. Methods 4 and 5 adopt different global influence, leading to different social payoffs. Thus, in the process of combining social payoff and preference payoff, we set different β values in order to optimize the prediction results. Figures 9 and 10 show the best results that we have obtained.

As shown in Figures 9 and 10, the measurements of [35] decrease over time, demonstrating the deterioration of the prediction ability over time. In [35], a critical task for predicting the node v's behavior is to find a node u among the v's neighbors such that when u is activated, the joint influence of all the v's active neighbors on v is for the first time greater than the



(a) Sina Weibo dataset

(b) Flickr dataset

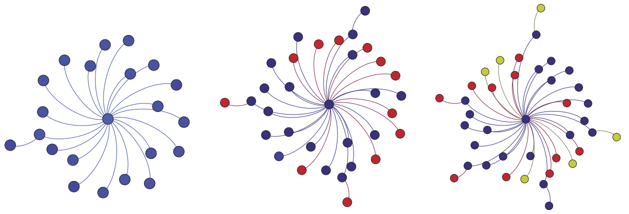
Fig. 11: Prediction performance using different approaches at the macroscopic view on Sina Weibo and Flickr datasets.

activation threshold of v. Then, the prediction method in [35] indicates that v will be activated before time $t_u+\tau_{u,v}$. However, the node u may not be the user who actually triggers the v's activation. In the early stage, the range of information diffusion is small, and there are fewer active nodes around v. Thus, the user u has a larger probability of being the user that triggered v's activation. As time increases and the diffusion range becomes larger, more nodes around v are getting activated and therefore the probability of user u being the real user that triggered the v's activation decreases. The prediction for the node v's activation time, which highly relies on node u, will cause a large deviation. In contrast, our model, as shown in the upper two curves in Figures 9 and 10, can achieve better and time-independent performance. This is because, in our model, the prediction of a node's behavior does not rely on the activation time of any particular neighbor. Instead, this prediction is made by combining both the active and inactive nodes to accurately measure the payoffs for different choices and thereby increases the reliability of prediction.

Figures 9 and 10 present the prediction performance of different methods at the microscopic view. Moreover, we also compare these methods at a macroscopic view. Specifically, we do statistical analysis of all individual prediction behaviors to get the overall diffusion scale, then calculate the $ScaleError$ based on Eq. 7 which is a macroscopic metric.

$$ScaleError = \frac{\sqrt{\sum_{i\in test}(PreScale(i) - RealScale(i))^2}}{\sqrt{\sum_{i\in test}(RealScale(i))^2}} \tag{7}$$

In Eq. 7, $PreScale(i)$ is the diffusion scale of information i that different methods predict, and $RealScale(i)$ is the scale of the real diffusion process of information i. Figure 11 presents the prediction performance of the baseline method and our two best approaches in terms of the $ScaleError$ metric on the Sina Weibo and Flickr datasets. The smaller $ScaleError$ value, the better prediction performance. The prediction results at the macroscopic view are similar to those at the microscopic view. Our method achieves better prediction performance than the baseline method.



(a) (b) (c)

Fig. 12: An real case of the diffusion prediction using the GT model in Sina Weibo. (a) the diffusion process at time t, (b) the real diffusion process at time $t+1$; (c) the prediction diffusion process at time $t+1$. The blue color means the node was activated before time t; the red color means the node is truly activated at time $t+1$; the yellow color means that the node is not truly activated at time $t+1$ but it is predicted to be activated.

A real case study. We analyze our GT model using a real case study which is presented by Figure 12. In Figure 12, subfigure (a) is the diffusion cascade of one tweet in Sina Weibo at time t, subfigure (b) is the real diffusion cascade of the tweet at time $t+1$, and subfigure (c) is the diffusion cascade of the tweet at time $t+1$ predicted by our model. Comparing the subfigure (b) with subfigure (c), we can see that our model can successfully predict most real activated users, while, to a slight extent, also predicts some users to be active who are not activated in fact. How to reduce these false predicted activated users is our further work.

Running time. In our experiments, we also test the running time of the proposed method. The running time mainly contains two parts: running time of parameters (*e.g.* global influence and social influence) learning and running time of diffusion prediction. When we predict the behavior of a user, we only use the information of the user's neighbors. Therefore, the running time of the prediction is not related much to the size of social network, and the prediction process is usually very fast. Parameters learning takes most running time of the experiments. Therefore, here, we test the running time of parameters learning on two datasets which are presented in table 2.

From table 2, we can see that learning the global influence parameters only needs several hundreds seconds or less on the networks with tens of millions of edges, and learning social influence parameters takes similar time on the diffusion logs containing tens of millions of user behaviors, which validates the efficiency of our method. We also notice that learning the social influence parameter on Sina Weibo is faster than that on Flickr, which is caused by the characteristic of the dataset. The Sina Weibo dataset directly provide the reposting behaviors between users, however, the Flickr dataset does not own similar information. Therefore, we need to use Algorithm 4 to extract the behaviors of information diffusion, which needs more time. As seen from the running time, our method

TABLE 2: Running time of parameters learning.(Sec)

	pagerank	cascade size	social influence
Sina Weibo	80.788	305.154	164.790
Flickr	633.484	647.887	1345.675

is fast and scalable. Some methods [31], [32] need too much running time, so we do not compare our method with them. The method in the literature [35] is most related to our work and also efficient, therefore, we consider it as the baseline in the experiments.

7 Conclusions

We have presented a novel information diffusion model (*i.e.* GT model) in this paper. It treats the nodes of a social network as autonomous, intelligent and rational agents, and jointly considers all of the interacting users and their preferences in the social network to make strategical decisions. By introducing the time-related user payoffs based on actual diffusion data, the proposed GT model has the capability of predicting the temporal dynamic of information diffusion process. User payoffs contain social payoff and preference payoff. Both the global influence of users and social influence between users are exploited for the calculation of user payoffs, where the social influence is presented in a novel manner by a non-negative vector of a fixed length that can fully capture complex dynamics of the user interaction. The similarity between the information and user preference is considered as preference payoff in this paper. Finally, we present the proposed algorithm for information diffusion prediction based on the proposed GT model. Experimental results on different genres of datasets with different evaluation metrics have justified the proposed prediction method.

Several challenges remain. In this work, we adopt a simple vector space model to profile users and information. In the future, we may adopt Linear discriminant analysis (LDA) to construct user profile and information profile to obtain better performance. Besides that, an effective method can be further designed to calculate preference payoff for the newly arriving users who do not have any post, and explore how to update the preference payoff of users. Finally, our current work considers the social networks where information propagates are static. However, both nodes and links in social networks are changing over time. Modeling information diffusion in dynamic social networks is a more challenging problem which will be further studied in our future work.

Acknowledgment

D. Li was supported in part by the Hong Kong Scholar project of China (No. ALGA4131016116), the China Postdoctoral fund (No. 2016M600250) and the major science and technology project of Shandong Province (No. 2015ZDXX0201B02). S. Zhang was supported in part by the Natural Science Foundation of China (No. 61672188). X. Sun was supported in part by the Natural Science Foundation of China (No. 61602128) and the Natural Science Foundation of Shandong Province (No. ZR2016FQ13). H. Zhou was supported in part by UK EPSRC under Grants EP/N508664/1 and EP/N011074/1, and Royal Society-Newton Advanced Fellowship under Grant NA160342. X. Li was supported in part by the National Natural Science Foundation of China (Grant No. 61761130079).

References

[1] J. Niu and L. Wang, "Structural properties and generative model of non-giant connected components in social networks," *Science China Information Sciences*, vol. 59, no. 12, p. 123101, 2016. 1

[2] R. Ji, Y. Gao, W. Liu, X. Xie, Q. Tian, and X. Li, "When location meets social multimedia: A survey on vision-based recognition and mining for geo-social multimedia analytics," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 1, p. 1, 2015. 1

[3] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. 7th SIGKDD*, 2001, pp. 57–66. 1

[4] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proc. 8th SIGKDD*, 2002, pp. 61–70. 1

[5] J. Leskovec, A. Krause, and et al, "Cost-effective outbreak detection in networks," in *Proc. 13th SIGKDD*, 2007, pp. 420–429. 1

[6] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proc. 16th SIGKDD*, 2010, pp. 1029–1038. 1

[7] S. Cheng, H. Shen, J. Huang, W. Chen, and X. Cheng, "Imrank: influence maximization via finding self-consistent ranking," in *Proc. 37th SIGIR*, 2014, pp. 475–484. 1

[8] S. Chen, J. Fan, G. Li, J. Feng, K.-l. Tan, and J. Tang, "Online topic-aware influence maximization," *Proceedings of the VLDB Endowment*, vol. 8, no. 6, pp. 666–677, 2015. 1

[9] D. Li, Z.-M. Xu, N. Chakraborty, A. Gupta, K. Sycara, and S. Li, "Polarity related influence maximization in signed social networks," *PloS one*, vol. 9, no. 7, p. e102199, 2014. 1

[10] Z. Wang, E. Chen, Q. Liu, Y. Yang, Y. Ge, and B. Chang, "Maximizing the coverage of information propagation in social networks." in *IJCAI*, 2015, pp. 2104–2110. 1

[11] B. Xu and L. Liu, "Information diffusion through online social networks," in *Proc. 2010 International Conference on Education Management and Management Science*, 2010, pp. 53–56. 1

[12] J. Zhao, J. Wu, and K. Xu, "Weak ties: Subtle role of information diffusion in online social networks," *Physical Review E*, vol. 82, no. 1, p. 016105, 2010. 1

[13] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Patterns of cascading behavior in large blog graphs," in *Proc. 2007 SDM*, 2007, pp. 551–556. 1, 2

[14] D. Liben-Nowell and J. Kleinberg, "Tracing information flow on a global scale using internet chain-letter data," *Proceedings of the National Academy of Sciences*, vol. 105, no. 12, pp. 4633–4638, 2008. 1, 2

[15] B. Golub and M. O. Jackson, "Using selection bias to explain the observed structure of internet diffusions," *Proceedings of the National Academy of Sciences*, vol. 107, no. 24, pp. 10 833–10 836, 2010. 1, 2

[16] D. Wang, Z. Wen, H. Tong, C.-Y. Lin, C. Song, and A.-L. Barabási, "Information spreading in context," in *Proc. 20th WWW*, 2011, pp. 735–744. 1, 2

[17] D. Li, Z. Xu, Y. Luo, and et al, "Modeling information diffusion over social networks for temporal dynamic prediction," in *Proc. 22nd CIKM*, 2013, pp. 1477–1480. 2

[18] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical review letters*, vol. 86, no. 14, p. 3200, 2001. 2

[19] R. M. May and A. L. Lloyd, "Infection dynamics on scale-free networks," *Physical Review E*, vol. 64, no. 6, p. 066112, 2001. 2

[20] H. W. Watson and F. Galton, "On the probability of the extinction of families," *The Journal of the Anthropological Institute of Great Britain and Ireland*, vol. 4, pp. 138–144, 1875. 2

[21] M. G. Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," *Proc. 28th ICML*, 2011. 2

[22] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su, "Understanding retweeting behaviors in social networks," in *Proc. 19th CIKM*, 2010, pp. 1633–1636. 2

[23] T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern, "Predicting information spreading in twitter," in *Workshop on computational social science and the wisdom of crowds, nips*, vol. 104, no. 45, 2010, pp. 17 599–601. 2

[24] H. Fei, R. Jiang, Y. Yang, B. Luo, and J. Huan, "Content based social behavior prediction: a multi-task learning approach," in *Proc. 20th CIKM*, 2011, pp. 995–1000. 3

[25] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influence in heterogeneous networks," in *Proc. 19th CIKM*, 2010, pp. 199–208. 3

[26] N. Du, L. Song, M. G. Rodriguez, and H. Zha, "Scalable influence estimation in continuous-time diffusion networks," in *Proc. 27th NIPS*, 2013, pp. 3147–3155. 3

[27] Y. Yang, J. Tang, C. W.-k. Leung, Y. Sun, Q. Chen, J. Li, and Q. Yang, "Rain: Social role-aware information diffusion." in *Proc. 29th AAAI*, 2015, pp. 367–373. 3

[28] B. Chang, H. Zhu, Y. Ge, E. Chen, H. Xiong, and C. Tan, "Predicting the popularity of online serials with autoregressive models," in *Proc. 23rd CIKM*, 2014, pp. 1339–1348. 3

[29] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proc. 23rd WWW*, 2014, pp. 925–936. 3

[30] S.-C. Hung, T.-T. Kuo, and S.-D. Lin, "Novel topic diffusion prediction using latent semantic and user behavior," in *Proc of the ASE BigData & Social Informatics 2015*, 2015, p. 39. 3

[31] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang, "Social action tracking via noise tolerant time-varying factor graphs," in *Proc. 16th SIGKDD*, 2010, pp. 1049–1058. 3, 14

[32] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura, and H. Motoda, "Learning diffusion probability based on node attributes in social networks," in *Proc. 19th Foundations of Intelligent System*, 2011, pp. 153–162. 3, 14

[33] 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*, vol. 12, no. 3, pp. 211–223, 2001. 3

[34] M. Granovetter, "Threshold models of collective behavior," *American journal of sociology*, vol. 83, no. 6, pp. 1420–1443, 1978. 3

[35] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proc. 3rd WSDM*, 2010, pp. 241–250. 3, 6, 11, 12, 14

[36] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1, pp. 107–117, 1998. 6, 7

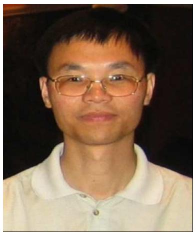
[37] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting tf-idf term weights as making relevance decisions," *ACM Transactions on Information Systems*, vol. 26, no. 3, p. 13, 2008. 7

[38] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proc. 18th WWW*, 2009, pp. 721–730. 10

Dong Li received his M.S. and Ph.D. degrees from Harbin Institute of Technology in 2010 and 2015, respectively. He had spent one year visiting Carnegie Mellon University during his PhD period. He is currently working in Shandong Univerity, Weihai, China and is also a Post-Doctoral Researcher at Harbin Institute of Technology, Harbin, China. His research interests include information diffusion, recommendation system and social data mining.

Shengping Zhang received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2013. He is currently a full Professor with the School of Computer Science and Technology, Harbin Institute of Technology at Weihai. He had been a Post-Doctoral Research Associate with Brown University, Providence, RI, USA, and a Visiting Student Researcher with University of California at Berkeley, Berkeley, CA, USA. He has authored or co-authored over 50 research publications in refereed journals and conferences.

Xin Sun received the B.S. and M.S. degrees from the School of Computer Science and Technology, Harbin Institute of Technology, China, in 2008 and 2010, respectively, and the Ph.D. degree from the Harbin Institute of Technology, in 2015. She is currently a lecturer at Harbin Institute of Technology, Weihai, China.



Huiyu Zhou is a Lecturer in the School of Electronics, Electrical Engineering and Computer Science at Queens University Belfast, United Kingdom. He obtained a Bachelor of Engineering degree in Radio Technology from Huazhong University of Science and Technology of China, and a Master of Science degree in Biomedical Engineering from the University of Dundee of United Kingdom, respectively. He was then awarded a Doctor of Philosophy degree in Computer Vision from the Heriot-Watt University, Edinburgh, United Kingdom. He has taken part in the consortiums of a number of research projects in medical image processing, computer vision, intelligent systems and data mining. Dr. Zhou has published widely in the field.

Sheng Li is currently a full professor at Harbin Institute of Technology, Harbin, China. His research interests include natural language processing, information retrieval and social data mining.

Xuelong Li (M'02-SM'07-F'12) is a full professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China.