Modeling Social Role-Aware Information Diffusion

Yang Yang[†], Jie Tang[†], Cane Wing-ki Leung [‡], Yizhou Sun[⋆], Qicong Chen[†], Juanzi Li[†], and Qiang Yang[‡]

[†]Department of Computer Science and Technology, Tsinghua University, China

[‡]Huawei Noah's Ark Lab, Hong Kong

[⋆]College of Computer and Information Science, Northeastern University, USA

{yangyang, jietang, ljz}@keg.cs.tsinghua.edu.cn,

{cane.leung, qiang.yang}@huawei.com, yzsun@ccs.neu.edu

Abstract

Information diffusion, which studies how information is propagated in social networks, has attracted considerable research effort recently. However, most existing approaches do not distinguish between different social roles that nodes may play in the diffusion process. In this paper, we study the interplay between users' social roles and their influence on information diffusion. In particular, we propose a generative model that integrates social role extraction and diffusion modeling into a unified framework. We then present a Gibbs-sampling based learning method to estimate the unknown parameters of the proposed model based on historical diffusion data. The proposed model can be applied in several scenarios. For instance, at the micro-level, the proposed model can be used to predict whether a user will repost a given message; while at the macro-level, it is able to predict both the scale and the duration of a diffusion process. We evaluate the proposed model on a real social media data set. Compared with several alternative methods, our model shows better performance in both micro- and macro-level prediction tasks.

1 Introduction

Information diffusion, also known as *diffusion of innovations*, is the study of how information propagates in or between networks (Rogers 2010). Central to information diffusion is the *influence* of individual nodes (or users in online social networks). In representative information diffusion models, such as the Linear Threshold (LT) model (Granovetter 1978) and the Independent Cascade (IC) model (Goldenberg, Libai, and Muller 2001), every directed link from a user v to another user u in a given network is associated with a non-negative weight, to reflect how much influence user v has on user u in information diffusion.

In reality, the information diffusion process is complex, as is the influence of one user on another. How information may diffuse in a network is affected by structure of the network, in which users' structural properties reflect their *social roles* in different communities (Wasserman and Faust 1994). Users' social roles in turn affect the influence they may have on other users, and hence the information diffusion process. Based on Twitter where a tweet corresponds to a piece of information and retweeting corresponds to information diffusion, a study reveals that 25% of information diffusion is controlled by 1% of users serving the role of *structural hole spanners*, who are bridges between otherwise disconnected communities in a network (Lou and Tang 2013). Another study shows that 50% of URLs on Twitter are posted by less than 1% of users who act as *opinion leaders*, who are people taking central positions in a community (Wu et al. 2011). Compared with posts originated from ordinary users, those from opinion leaders not only attracted much more retweets (larger diffusion scales), but also have longer lifespans (longer diffusion lengths). All these findings suggest that it is crucial to consider users' social roles in information diffusion modeling.

Social roles and diffusion are not independent of each other in nature. To further motivate our study, we present an exploratory analysis on a large social network with 200 million users and 174 million microblog messages. Each post (message) in this network is considered a piece of information, while reposting (or retweeting in Twitter) corresponds to the diffusion of information. We analyze how users taking three roles, namely opinion leaders, structural hole spanners and ordinary users, influence other users' probability of reposting a message.

Figure 1 provides the results. When an opinion leader reposts a message, the probability that her follower v will subsequently repost the message is 12 times higher than the case where the message is reposted by an ordinary user in the first place. More interestingly, if the number of reposting opinion leaders, all followed by v, reaches 3, the probability that v will subsequently repost decreases significantly, but keeps increasing after that. Regarding this finding, we conjecture that 2-3 opinion leaders are sufficient to spread a piece of information throughout a community, making their followers unwilling to repost a message that most of her friends would have known already. However, when a message attracts the attention of more than 3 opinion leaders in a community, it may have become so influential and popular that reposting the message becomes a social norm that other users might want to adopt. Results on structural hole spanners show a different story. The probability for v to repost i keeps increasing with the number of her reposting followees who are structural hole spanners. As structural hole spanners are those who bridge different otherwise disconnected commu-

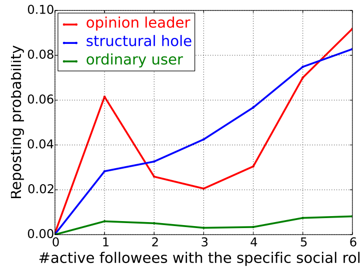


Figure 1: Diffusion influence analysis. We study how users with different roles affect other users' probability of reposting a certain message. In the figure, y-axis denotes the probability that a user v will repost a certain message. X-axis denotes the number of v's followees who reposted the message before v did.

nities, they tend to bring information that a certain community is rarely exposed to, thus may be able to interest v more easily. To summarize, the probability that a user will repost a message depends strongly on the roles of her followees who reposted the message. It is therefore crucial to capture users' social roles when modeling the information diffusion process.

Intuitively, a user may play multiple roles with respect to different communities or social circles, thus exhibiting different influential strengths in different diffusion processes. For instance, one may act as an opinion leader when speaking on her area of expertise, and a structural hole spanner when forwarding a piece of news from her colleagues to her family members. How to effectively uncover the social roles users play in information diffusion processes remain an open problem. In this paper, we approach this problem through a *role-aware information diffusion model*. There are two intuitions behind our model. Firstly, a user may play multiple social roles in a network as noted. We therefore propose to learn a probability distribution over social roles for each user, allowing a user to play different roles in different diffusion processes. Secondly, as social roles and diffusion process are interrelated, we can exploit the observed diffusion in a network to help infer the unobserved roles of users and the influence of each role. As such, our model takes as input a social network and its information diffusion traces. It then jointly learns the social role distributions of users and the influence of each role by utilizing both users' structural properties and their behaviors as observed in the diffusion traces. We summarize our technical contributions as follows:

- We propose the problem of role-aware information diffusion modeling in online social networks.
- We formulate a generative model and devise a Gibbs sampler that integrates social roles learning and diffusion modeling into a unified probabilistic framework.
- Employing a large real-world network as experimental data, we conduct extensive experiments to validate the proposed model over several baselines.

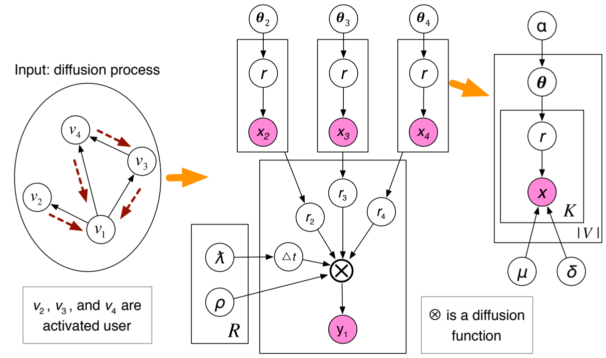


Figure 2: Illustration of our social role-aware diffusion model. Notice that r_2 is the social role that v_2 plays when she tries to activate v_1; an r with no subscript indicates the role sampled for generating a user's social attributes.

2 Social Role-Aware Diffusion Model

2.1 Formulation

Let $G = (V, E, X)$ be a social network. V is a set of users, $E \subseteq V \times V$ is a set of links between users, $e_{vu} \in E$, where $v, u \in V$, denotes a directed (follow) link from v to u, and X is a $|V| \times K$ social attribute matrix, with each row $\mathbf{x}_v = \{x_1, \ldots, x_K | x_i \in \mathbb{R}\}$ representing K *social attributes* of the user v. The K social attributes to use can be defined based on application-specific needs. Examples include PageRank score, network constraint score, node degree, etc. For each node $v \in V$, we use $B(v) = \{u | u \in V, e_{vu} \in E\}$ to denote the set of followees of v.

Different pieces of messages will be propagated over G. When a user v posts or reposts a specific message i at time t, we say that the user v is activated with respect to i at t (and will stay active after t).

Intuitively, a user may take different roles in different information diffusion processes. For instance, she may act as an opinion leader when spreading messages about a specific topic of her interest, and a structural hole spanner when propagating a piece of news from her colleagues to her family members. We model this intuition by associating each user with a *social role distribution*:

Definition 1. **Social Role Distribution.** The social role distribution of a user $v \in V$ is denoted by $\theta_{\mathbf{v}}$, which is a R-dimensional vector and satisfies $\sum_r \theta_{vr} = 1$. θ_{vr} is the probability that v plays the role r when diffusing a certain message.

2.2 Model Description

Figure 2 illustrates our model. Overall, the model determine the social role distribution of each user according to both her structural attributes and her behavior in diffusion process. Thus our generative model contains two parts: users' social attributes generation and information diffusion process generation.

Inspired by the work in (Lou and Tang 2013), we consider three social roles in this paper, namely *opinion leaders*, *structural hole spanners* (who propagates information from one community to another in social networks), and *ordinary users*. Existing work detect social roles of users only based on their social attributes. For example, Burt (Burt 2009) treats users with small network constraint scores as structural hole spanners, and opinion leaders are often measured by PageRank (Page et al. 1999). However, these methods limit from restricting users to act as the same role when propagating different messages. In our model, the social role distribution of each user is determined not only by her social attributes but also by her behaviors in the information diffusion process.

Generative process. We first introduce the diffusion process generation. Generally, inspired by our exploratory analysis, which reveals that the social role of a user affects her influential strength and diffusion delay, we introduce per-role parameters ρ_r and λ_r as the probability that users playing role r will activate another user successfully and will cause a 1-timestamp diffusion delay respectively. We then use a diffusion function (e.g., a threshold function or a cascade function) parametrized with ρ_r and λ_r to determine whether the user will become active. In this paper, we implement the diffusion function based on the Independent Cascade model. We leave the implementation of the Linear Threshold model as future work.

For the details, we first generate the influential strength and diffusion delay with respect to each social role r: $\rho_r \sim \text{Beta}(\beta)$, $\lambda_r \sim \text{Beta}(\gamma)$. Consider message i which is first posted by user u at time t, u will have a chance to activate each inactive follower v: first, we sample the role r, which user u is playing when she tries to activate v: $r \sim \text{Mult}(\theta_u)$. Next, we generate a diffusion delay $\triangle t$ according to the geometric distribution $P(\triangle t|\lambda_r)$. At time $t' = t+\triangle t+1$, we toss a coin: $z_{iuv}^{t'} \sim \text{Bernoulli}(\rho_r)$, to determine whether u will succeed in activating v. At anytime, user v will become active if at least one of her followees activate her successfully. Notice that multiple activation attempts are sequenced in an arbitrary order. After v becomes active, she will then execute the diffusion process we just described to try to activate her inactive followers. The process terminates when no more activation is possible.

For the social attribute generation process, we first generate each user v's social role distribution: $\theta_v \sim \text{Dir}(\alpha)$. Then, for each role r, we generate K Gaussian parameters: $(\mu_{rk}, \delta_{rk}) \sim \text{NG}(\tau)$, for $k = 1, ..., K$. Next, for the k-th attribute of user v, we generate a latent variable: $r \sim \text{Mult}(\theta_v)$. Finally, we generate that attribute: $x_{vk} \sim \text{N}(\mu_{rk}, \delta_{rk}^{-1})$.

Likelihood function. We start with the notations required to define the likelihood function. For each message i, A_{it} as the set of users who become active at time t, and $D_{it} = A_{i0} \cup \cdots \cup A_{it}$ as the set of users who are active by time t. We further define a binary variable y_{itu} to denote whether user u is activated ($y_{itu} = 1$) or not ($y_{itu} = 0$) with respect to message i at time t. For each user v, $\mathbf{z}_{i*v}^t = (z_{iuv}^t)_{u \in B(v) \cap D_{it-1}}$ is an indicator vector. $z_{iuv}^t = 1$ if user u succeeds in activating user v at time t to diffuse message i, and $z_{iuv}^t = 0$ if user u fails to activate v within time $[t_{iu}+1, t]$, where t_{iu} indicates the time u was activated to diffuse message i.

We consider the probability that the user u will succeed in activating one of her followers v at time t ($z_{iuv}^t = 1$), by considering u's social role information:

$$\varphi_{iuv}^t = \sum_r \rho_r \lambda_r (1-\lambda_r)^{t-t_{iu}-1} \theta_{ur}. \tag{1}$$

We further define a binary variable y_{itu} to denote whether user u is activated ($y_{itu} = 1$) or not ($y_{itu} = 0$) with respect to message i at time $t.D_{it}$ denoting the set of users who are active by time t. If user v is not activated by user $u \in B(v) \cap D_{it-1}$ within the time period $[t_{iu}+1, t]$, then $z_{iuv}^t = 0$ with probability:

$$\varepsilon_{iuv}^t = \sum_r \theta_{ur}[\rho_r(1-\lambda_r)^{t-t_{iu}} + 1 - \rho_r]. \tag{2}$$

Based on Eqs. (1) and (2), the probability that user v is active at time t can be expressed as:

$$P(v \in A_{it}) = \prod_{u \in B(v) \cap D_{it-1}} (\varphi_{iuv}^t + \varepsilon_{iuv}^t) - \prod_{u \in B(v) \cap D_{it-1}} \varepsilon_{iuv}^t. \tag{3}$$

Further, the probability that user v is never activated by the last timestamp T can be written as:

$$P(v \notin D_{iT}) = \prod_{u \in B(v) \cap D_{iT}} \sum_r (1-\rho_r)\theta_{ur}. \tag{4}$$

Another important component in the proposed model is social attribute generation for users. In our model, we assume that each attribute of a user is sampled according to a Gaussian distribution. Formally,

$$P(x_{uk}) = \sum_r \sqrt{\frac{\delta_{rk}}{2\pi}} \exp\{-\frac{\delta_{rk}(x_{uk}-\mu_{rk})^2}{2}\}\theta_{ur}. \tag{5}$$

Based on Eqs. (3) to (5), we define the likelihood function as the joint probability of observed diffusion process and users' structural attributes as follows

$$\begin{aligned}\log L = &\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{v \in A_{it}} \log P(v \in A_{it}) + \sum_{i=1}^{I}\sum_{v \notin D_{iT}} \log P(v \notin D_{iT}) \\ &+ \sum_{u \in V}\sum_{k=1}^{K} \log P(x_{uk}) + \sum_{u \in V}\sum_{r=1}^{R} \log P(\theta_{ur}|\alpha) \\ &+ \sum_{r=1}^{R}\{\log P(\rho_r|\beta) + \log P(\lambda_r|\gamma)\} + \sum_{r=1}^{R}\sum_{k=1}^{K} \log P(\mu_{rk}, \delta_{rk}|\tau).\end{aligned} \tag{6}$$

2.3 Model learning

We employ Gibbs sampling (Resnik and Hardisty 2010) to estimate the unknown parameters of the proposed model.

We begin with the posterior for sampling the latent variable r for each social attribute of a user u:

$$P(r_{uk}|\mathbf{r}_{\neg uk},\mathbf{x}) = \frac{n_{ur_{uk}}^{\neg uk}+\alpha}{\sum_r (n_{ur}^{\neg uk}+\alpha)} \frac{\Gamma(\tau_2+\frac{n_{r_{uk}k}}{2})}{\Gamma(\tau_2+\frac{n_{r_{uk}k}^{\neg uk}}{2})} \times \frac{\sqrt{(\tau_1+n_{r_{uk}k}^{\neg uk})}\eta(n_{r_{uk}k}^{\neg uk},\bar{x}_{r_{uk}k}^{\neg uk},s_{r_{uk}}^{\neg uk})}{\sqrt{(\tau_1+n_{r_{uk}k})}\eta(n_{r_{uk}k},\bar{x}_{r_{uk}k},s_{r_{uk}})}, \tag{7}$$

where the counters n_{ur} (and n_{rk}) denote the number of times r being sampled with (the k-th social attribute of) user u. $\bar{x}_{rk}$ and s_{rk} are the mean and variance of the k-th social attribute with role r. The notation $\neg uk$ on the counters indicates exclusion of the current observation (the k-th structural attribute of user u) from the counts. One challenge in Eq. (7) is the calculation of Gamma functions, which we approximated in this work using Stirling's formula (Abramowitz and Stegun 1970). The function $\eta(\cdot)$ is used to simplify Eq. (7) and is defined as:

$$\eta(\cdot) = [\tau_3 + \frac{1}{2}(n_{r_{uk}k}s_{r_{uk}k} + \frac{\tau_1 n_{r_{uk}k}(\bar{x}_{r_{uk}k}-\tau_0)^2}{\tau_1+n_{r_{uk}k}})]^{(\tau_2+\frac{n_{r_{uk}k}}{2})}. \tag{8}$$

In Eqs. (7) and (8), τ is the parameter of normal-gamma prior. In practice, according to (Murphy 2007), we set τ_0 to the mean of all attributes, τ_1 to the number of instances, τ_2 to half of τ_1, and τ_3 to half of the sum of squared deviations to all features.

Similarly, we evaluate the posterior for sampling the latent variables $(\mathbf{t},\mathbf{r},\mathbf{z})$ for each diffusion process:

$$\begin{aligned} &P(r_{iuv},\triangle t_{iuv},z_{iuv}|\mathbf{r}_{\neg iuv},\triangle\mathbf{t}_{\neg iuv},\mathbf{z}_{\neg iuv},\mathbf{y}) \\ &= \frac{n_{ur_{iuv}}^{\neg iuv}+\alpha}{\sum_r (n_{ur}^{\neg iuv}+\alpha)} \times \frac{n_{z_{iuv}r_{iuv}}^{\neg iuv}+\beta_1^{z_{iuv}}\beta_0^{1-z_{iuv}}}{n_{1r_{iuv}}^{\neg iuv}+\beta_1+n_{0r_{iuv}}^{\neg iuv}+\beta_0} \\ &\times \frac{(n_{r_{iuv}}^{\neg iuv}+\gamma_1)\prod_{t=0}^{\Delta t-2}(s_{r_{iuv}}^{\neg iuv}-n_{r_{iuv}}^{\neg iuv}+\gamma_0+t)}{\prod_{t=0}^{\Delta t-1}(\gamma_1+s_{r_{iuv}}^{\neg iuv}+\gamma_0+t)} \times \Phi, \end{aligned} \tag{9}$$

where n_r (and n_{zr}) denotes the number of times r sampled (with z); s_r denotes the sum of $\triangle t$ that has been sampled with r. We use Φ to indicate $\frac{P(\mathbf{y}|\mathbf{z},\triangle\mathbf{t})}{P(\mathbf{y}_{\neg iuv}|\mathbf{z}_{\neg iuv},\triangle\mathbf{t}_{\neg iuv})}$ for brevity. Intuitively, Φ is used to handle contradictions arise during the sampling process. It equals to 0 in the following three cases: (1) user v is inactive with respect to message i ($y_{iTv}=0$), but z_{iuv} is sampled as 1; (2) $y_{iTv}=1$, but $z_{iuv}=0$ and no user has activated v ($\sum_{k,\neg u} z_{ikv}=0$); (3) $y_{iTv}=1$, but $\triangle t \neq t_{iv}-t_{iu}$, and no other user h has attempted to activate v at time t_{iv} ($\sum_{h,\neg u}(t_{ih}+\triangle t_{ihv}=t_{iv})=0$). The third case shows that the observed $\mathbf{y}$ imposes certain restrictions on the latent $\triangle\mathbf{t}$. In all other cases, we have $\Phi=1$.

We now estimate model parameters by the sampling results. The updating rules for θ, λ, and ρ can be easily deduced as in LDA (Heinrich 2005):

$$\begin{aligned} \theta_{ur} &= P(\tilde{r}=r|\mathbf{r},\triangle\mathbf{t},\mathbf{z},\mathbf{y}) = \frac{n_{ur}+\alpha}{\sum_r(n_{ur}+\alpha)} \\ \lambda_r &= P(\triangle\tilde{t}=1|\tilde{r}=r,\mathbf{r},\triangle\mathbf{t},\mathbf{z},\mathbf{y}) = \frac{n_r+\gamma_1}{\gamma_1+s_r+\gamma_0} \\ \rho_r &= P(\tilde{z}=1|\tilde{r}=r,\mathbf{r},\triangle\mathbf{t},\mathbf{z},\mathbf{y}) = \frac{n_{(z=1)r}+\beta_1}{n_{1r}+\beta_1+n_{0r}+\beta_0}, \end{aligned} \tag{10}$$

where $\tilde{r}$, $\triangle\tilde{t}$ and $\tilde{z}$ respectively represent a new observation of r, $\triangle t$ and z. Note that the updating rules of both μ_{rk} and δ_{rk} involve an integration that is hard to compute. Hence, we approximate μ_{rk} and δ_{rk} as $E(\mu_{rk})$ and $E(\delta_{rk})$ respectively according to (Bernardo and Smith 2009):

$$\begin{aligned} \mu_{rk} &\approx E(\mu_{rk}) = \frac{\tau_0\tau_1+n_{rk}\bar{x}_{rk}}{\tau_1+n_{rk}}, \\ \delta_{rk} &\approx E(\delta_{rk}) = \frac{2\tau_2+n_{rk}}{2\tau_3+n_{rk}s_{rk}+\frac{\tau_1 n_{rk}(\bar{x}_{rk}-\tau_0)^2}{\tau_1+n_{rk}}}. \end{aligned} \tag{11}$$

3 Experimental Results

In this section, we present experimental results to validate the effectiveness of the proposed social role-aware diffusion model. Our data sets and codes are publicly available[1].

3.1 Experimental Setup

Data set. We conduct experiments on real data from Tencent Weibo[2], a popular Twitter-like microblogging service in China. The complete data set contains the *directed following networks* and *posting logs* of over 200 million users. If there exists a following link from a user v to another user u, we say that v is a follower of u, and that u is a followee of v. Similar to Twitter, there are two types of posts in Tencent Weibo, namely original posts and reposts (or retweets). The reposting log of an original post essentially represents an information diffusion process. We extracted the complete following relationships between users and all posting logs of November 1, 2011 as the training set, and those of November 2, 2011 as the test set for evaluating our proposed model. In total, we have 184,491 users, and 4,588,559 original posts. We removed from both the training and test sets original posts that were reposted by fewer than 5 users, and use the remaining 242,831 original posts for experiments.

We further categorize posts in our data set based on their topics, as existing work has discovered that information diffusion behavior of users is dependent on the topic of the information (Yang and Leskovec 2010). Specifically, we first use LDA (Blei, Ng, and Jordan 2003) to extract latent topics from all the posts in our data set, and assign each post to the topic to which it is most relevant. Due to the space limitation, we just demonstrate the results of 4 most popular topics: *campus*, *horoscope*, *movie*, and *history*.

Tasks. We evaluate the proposed model based on the following two tasks. (1) At the micro-level, how accurate is the role-aware diffusion model in predicting whether a user

[1] http://arnetminer.org/role-aware-diffusion/
[2] http://t.qq.com/

will repost a given message? (2) At the macro-level, can the role-aware diffusion model predict the scale and duration of a diffusion process?

3.2 Micro-Level Evaluation

Evaluation setting. Given an original post (message) on a particular topic, we aim to identify users who will most likely repost this message. Specifically, for each original post in the test set, we rank all users according to their probability of reposting the given message as predicted by the proposed model and several baseline methods (described below). Note that on average, each original message in our data set was only reposted by 0.008% of users. We consider the following baselines in our experiments:

Count. Given an original post i, this method ranks users, in descending order, by the number of followees who have reposted i. This method assumes that a user's reposting decision only depends on her followees' decisions.

SVM. This method predicts whether user v will repost i based on three features: the number of v's followers who have reposted i, the number of v's followees who have reposted i, and the number of times v reposted a message posted by the author of i before. Similar features have been utilized in (Zhang et al. 2013). This method then trains a Ranking SVM (Joachims 2002; 2006) to predict v's probability of reposting i. For Ranking SVM, we use TreeRankSVM (Airola, Pahikkala, and Salakoski 2011) to handle our large-scale data.

IC Model. This method employs the traditional Independent Cascade (IC) model (Goldenberg, Libai, and Muller 2001; Kempe, Kleinberg, and Tardos 2003). We estimate the parameters of the IC model from the training set by the learning algorithm proposed in (Kimura et al. 2011).

Role-aware. This is the proposed social role-aware diffusion model. For each message i, both this method and IC model use the simulation method to calculate the probability of a user being activated and rank all users by that. We empirically set the model parameters as: $R = 10$, $\alpha = 0.1$, $\beta = (1, 1)$, and $\gamma = (1, 1)$.

Performance comparison. Table 1 shows the performance of the proposed model and baselines in the micro-level prediction task. Overall, all models perform unsatisfactorily, which is not surprising due to the small percentage of positive instances in the data set (around 0.008%). Our model outperforms the baseline methods by 32.6% in terms of MAP on average. Due to the lack of supervised information, Count performs worst on all topics. SVM generates mixed performance. It performs well on "local" topics (e.g., "horoscope", as people tend to be interested in posts about their own constellations), but falls short on more "global" topics (e.g., "movie"). This can be explained by the fact that SVM optimizes the reposting probability of each user independently by considering only her local diffusion features, while neglecting the overall mechanism behind the whole diffusion process. For IC, its performance is hindered by the over-fitting problem resulting from its large number of unknown parameters to learn. The proposed social role-aware diffusion model addresses such a problem by allowing users with the same social role to share the same diffusion patterns, thus greatly reduces the number of model parameters.

Table 1: Performance of repost prediction on several topics.

Topic	Method	P@10	P@50	P@100	MAP
Campus	Count	0.028	0.010	0.006	0.068
	SVM	0.098	0.045	0.032	0.127
	IC Model	**0.231**	0.142	0.102	0.259
	Role-aware	0.228	**0.145**	**0.106**	**0.263**
Horoscope	Count	0.019	0.010	0.006	0.005
	SVM	0.124	**0.162**	0.088	**0.263**
	IC Model	0.149	0.111	0.098	0.125
	Role-aware	**0.171**	0.121	**0.102**	0.130
Movie	Count	0.015	0.007	0.004	0.009
	SVM	0.094	0.111	0.060	0.199
	IC Model	0.227	0.147	**0.147**	0.236
	Role-aware	**0.229**	**0.173**	0.144	**0.238**
History	Count	0.191	0.056	0.033	0.096
	SVM	0.154	0.051	0.030	0.221
	IC Model	0.206	0.134	**0.135**	0.230
	Role-aware	**0.225**	**0.171**	0.134	**0.262**

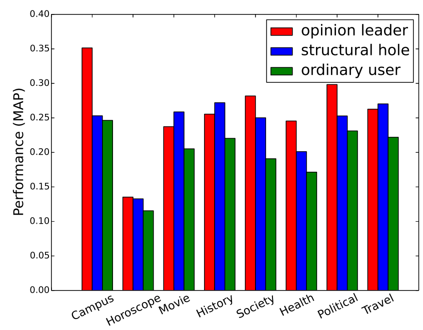
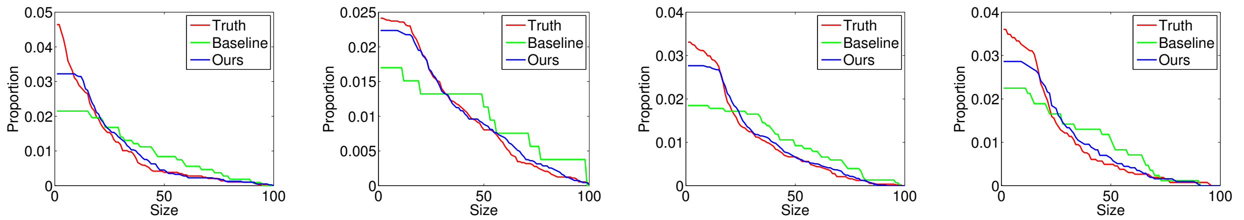


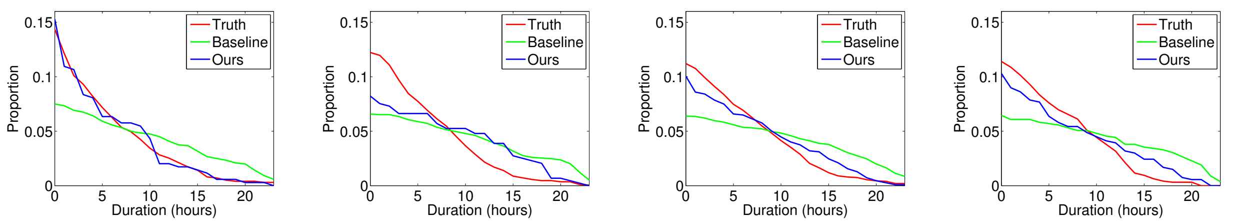
Figure 3: Social role analysis.

Social role analysis. We further study how social roles influence the diffusion process of messages with different topics. To conduct this experiment, we first analyze the estimated Gaussian parameters of the proposed model, which summarize the structural properties of users taking a certain role, to uncover the meaning of the latent roles learned by our model. For instance, a latent role with high PageRank score is considered to be representing the opinion leader. Next, we group users into opinion leaders, structural hole spanners, and ordinary users. Finally, we use the proposed model to perform per-group predictions and analyze the results. We present four more topics in this experiment: *society*, *health*, *political*, and *travel*. As Figure 3 shows, our model can better predict the diffusion behavior of opinion leaders and structural hole spanners, as ordinary users tend to behave more randomly. Furthermore, opinion leaders can be better predicted on more regional and specialized topics (e.g., "campus", "society" and "political"), while structural hole spanners can be better predicted on more general topics, which tend to propagate from one domain to another more



(a) Campus (b) Horoscope (c) Movie (d) History

Figure 4: Diffusion scale distributions of the different topics in the test set.



(a) Campus (b) Horoscope (c) Movie (d) History

Figure 5: Diffusion duration distributions of the different topics in the test set.

easily (e.g., "movie", "history", and "travel").

3.3 Macro-Level Evaluation

Evaluation setting. At the macro-level, we use the fitted model to predict the *scale* and *duration* of a diffusion process. Specifically, we first trace the diffusion process of each topic by selecting all original posts relevant to that topic. Then, we evaluate how accurate the proposed model can predict for each topic its diffusion scale, defined as the number of reposts of the original posts under that topic, and the diffusion duration, defined as the last reposting time of these posts. We use the IC model as the baseline for comparison.

Diffusion scale prediction. Figs. 4(a)-(d) show the diffusion scale prediction results for the 8 different topics. The x-axis in each sub-figure denotes the number of reposts, and the y-axis denotes the proportion of original posts with a particular number of reposts. Overall, our method performs better, while the baseline method tends to overestimate diffusion scale.

Diffusion duration prediction. Figs. 5(a)-(d) show the diffusion duration prediction results of the two models. The x-axis in each sub-figure denotes the time interval between the posting time of an original post and the latest repost time of it, while the y-axis shows the proportion of the original posts with a particular diffusion duration.

4 Related Work

Information diffusion, which can be viewed as the spread of innovations (Rogers 2010), studies how new ideas and messages spread in or between networks. Recent years have seen extensive modeling efforts on the information diffusion process (Lerman and Ghosh 2010; Gomez Rodriguez, Leskovec, and Krause 2010; Leskovec et al. 2007; Sadikov et al. 2011), with the two types of fundamental models being *Linear Threshold models* (LT) (Granovetter 1978) and *Independent Cascade models* (IC) (Goldenberg, Libai, and Muller 2001). Both types of models assume that the tendency of an inactive user to become active increases monotonically with the number of her active neighbors. However, according to the experiments conducted in this paper, we show that the probability of a user become active is not a simple monotonic function of the number of her active neighbors, but is relevant to the user's social role.

Based on the two fundamental models, Barbieri et al. (Barbieri, Bonchi, and Manco 2013) studied social influence from a topic modeling perspective. In particular, they focused on modeling the latent topics of the diffused information, and proposed topic-aware extensions to the LT and IC models. Myers et al. (Myers, Zhu, and Leskovec 2012) considered external influence in information diffusion. In their model, information can be diffused to a node through links in the given network or through influence of external sources. Rodriguez et al. (Rodriguez, Leskovec, and Schölkopf 2013) applied the survival theory to generalize some existing diffusion models into a multiplicative model. In contrast to our work, the aforementioned studies focus only on the diffusion process without considering how different types of users may influence such process.

5 Conclusion

In this paper, we study the interplay between users' social roles and their influence on information diffusion. We propose a novel social role-aware diffusion model, which integrates social role extraction and diffusion modeling into a unified framework. We evaluate the proposed model on a real social media data set at both micro- and macro-levels. Compared with several alternative methods, our model shows better performance.

References

Abramowitz, M., and Stegun, I. 1970. Handbook of mathematical functions.

Airola, A.; Pahikkala, T.; and Salakoski, T. 2011. An improved training algorithm for the linear ranking support vector machine. In *ICANN 2011*.

Barbieri, N.; Bonchi, F.; and Manco, G. 2013. Topic-aware social influence propagation models. *Knowledge and information systems* 37(3):555–584.

Bernardo, J. M., and Smith, A. F. 2009. *Bayesian theory*, volume 405. Wiley. com.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR* 3:993–1022.

Burt, R. S. 2001. Structural holes versus network closure as social capital. *Social capital: Theory and research* 31–56.

Burt, R. S. 2009. *Structural holes: The social structure of competition*. Harvard University Press.

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

Gomez Rodriguez, M.; Leskovec, J.; and Krause, A. 2010. Inferring networks of diffusion and influence. In *KDD'10*, 1019–1028.

Granovetter, M. 1978. Threshold models of collective behavior. *American journal of sociology* 83(6):1420.

Heinrich, G. 2005. Parameter estimation for text analysis. *Web: http://www. arbylon. net/publications/text-est. pdf*.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *KDD'02*, 133–142.

Joachims, T. 2006. Training linear svms in linear time. In *KDD'06*, 217–226.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *KDD'03*, 137–146.

Kimura, M.; Saito, K.; Ohara, K.; and Motoda, H. 2011. Learning information diffusion model in a social network for predicting influence of nodes. *Intelligent Data Analysis* 15(4):633–652.

Lerman, K., and Ghosh, R. 2010. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM'10* 10:90–97.

Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N. S.; and Hurst, M. 2007. Patterns of cascading behavior in large blog graphs. In *SDM'07*, volume 7, 551–556.

Lou, T., and Tang, J. 2013. Mining structural hole spanners through information diffusion in social networks. In *WWW'13*, 825–836.

Murphy, K. P. 2007. Conjugate bayesian analysis of the gaussian distribution. *def* $1(2\sigma 2)$:16.

Myers, S. A.; Zhu, C.; and Leskovec, J. 2012. Information diffusion and external influence in networks. In *KDD'12*, 33–41.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University.

Resnik, P., and Hardisty, E. 2010. Gibbs sampling for the uninitiated. Technical report, DTIC Document.

Rodriguez, M. G.; Leskovec, J.; and Schölkopf, B. 2013. Modeling information propagation with survival theory. *ICML'13* 666–674.

Rogers, E. M. 2010. *Diffusion of innovations*. Simon and Schuster.

Sadikov, E.; Medina, M.; Leskovec, J.; and Garcia-Molina, H. 2011. Correcting for missing data in information cascades. In *WSDM'11*, 55–64.

Wasserman, S., and Faust, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press. chapter 9.

Wu, S.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Who says what to whom on twitter. In *WWW'11*, 705–714.

Yang, J., and Leskovec, J. 2010. Modeling information diffusion in implicit networks. In *ICDM'10*, 599–608.

Zhang, J.; Liu, B.; Tang, J.; Chen, T.; and Li, J. 2013. Social influence locality for modeling retweeting behaviors. In *AAAI'13*, 2761–2767.