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 the 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 on social role aware information diffusion, 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 (corresponding to two-step flow theory (Lazarsfeld, Berelson, and Gaudet 1944)). 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, which leads information overload to information everywhere. Results on structural hole spanners show a different story. The proba-
bility for \( v \) to repost a post keeps increasing with the number of her reposting followers who are structural hole spanners. As structural hole spanners are those who bridge different otherwise disconnected communities, they tend to bring information that a certain community is rarely exposed to, thus may be able to interest \( v \) more easily (Burt 2001). This result also suggests that most users tires to bridge information flow between different groups. To summarize, the probability that a user will repost a message depends strongly on the roles of her followers 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 remains 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.

### Social Role-Aware Diffusion Model

#### Formulation

Let \( G = (V, E, X) \) be a social network, where \( V \) is a set of users, \( E \subseteq V \times V \) is a set of links between users, \( e_{uv} \in E \), denotes a directed (follow) link from user \( v \) to \( u \) \((v, u \in V)\), and \( X \) is a \(|V| \times K \) social attribute matrix, with each row \( x_v = \{x_1, \ldots, x_K : x_i \in \mathbb{R}\} \) 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 (Page et al. 1999), network constraint score (Burt 2009; Lou and Tang 2013), 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 \( t \)).

To model the intuition that a user may take different social roles in different diffusion processes, we associate each user with a social role distribution:

**Definition 1. Social Role Distribution.** The social role distribution of user \( v \in V \) is denoted by \( \theta_v \), which is a \( R \)-dimensional vector and satisfies \( \sum \theta_{vy} = 1 \). \( \theta_{vy} \) is the probability that \( v \) plays role \( r \) when diffusing a message.

#### Model Description

We propose a social Role-Aware INformation diffusion model (RAIN) for learning users’ social roles and modeling information diffusion simultaneously. Figure 2 illustrates our model. RAIN determines social role distribution of each user according to both her structural attributes and her behavior in diffusion process. Inspired by the work in (Lou and Tang 2013), we consider three social roles in this paper, namely opinion leaders, structural hole spanners, and ordinary users. Existing work detects social roles of users only based on their social attributes. For example, Burt (Burt 2009) treats users with small network constraint scores as
diffusion time delay
a latent variable denoting the social role of user
number of latent roles
total number of social attributes of users
social role distribution of user
the time when

\( \rho \)

cess generation. Inspired by our exploratory analysis, which
We first introduce the diffusion pro-

two parts: users’ social attributes generation and information
diffusion behaviors. Overall, our generative model contains
not only by her social attributes but also by her information
diffusion behaviors. Overall, our generative model contains
two parts: users’ social attributes generation and information
diffusion process generation.

Generative process. We first introduce the diffusion pro-
cession. 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 param-
eters \( \rho_r \) and \( \lambda_r \) 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 diffu-
sion function (e.g., a threshold function or a cascade func-
tion) parametrized by \( \rho_r \) and \( \lambda_r \) to determine whether a user
will become active. In this paper, to make things concrete,
we focus on the Independent Cascade model.

More specifically, we first generate the influential strength
and diffusion delay with respect to each social role \( r \): \( \rho_r \sim \text{Beta(} \beta \text{)}, \lambda_r \sim \text{Beta(} \gamma \text{)} \). 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 \( \Delta t \) according to the
geometric distribution \( P(\Delta t | \lambda_r) \). At time \( t' = t + \Delta t + 1 \), we
toss a coin: \( z_{tuv}^{t'} \sim \text{Bernoulli}(\rho_{ru}) \), to determine whether \( u \)
will succeed in activating \( v \). At anytime, user \( v \) will become active
if at least one of her fellowees activate her success-
fully. 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 acti-
ivate her inactive followers. The process terminates when no
more activation is possible.

For the social attribute generation process, we assume that
each attribute of a user \( v \) is sampled according to a Gaussian
distribution. Users with the same social roles have similar
social attributes and share the same Gaussian distribution.
Thus, we first generate each user \( v \)’s social role distribution:
\( \theta_v \sim \text{Dir(} \alpha \text{)} \). Then, for each role \( r \), we generate \( K \) Gaussian
parameters: \( (\mu_{rk}, \delta_{rk}) \sim \text{NG}(\tau) \), for \( k = 1, \ldots, K \). 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: \( x_{vk} \sim \text{N}(\mu_{rk}, \delta_{rk}^{-1}) \).
Table 1 summarizes major notations used in
RAIN.

### Likelihood function

For each message \( i \), we define \( A_{it} \) as the set of users who become active at time \( t \), \( D_{it} = A_{it} \cup \cdots \cup A_{it} \) as the set of users who are active by time \( t \),
and the 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 user \( v \), \( z_{ituv}^r = (z_{ituv}^{1}, \ldots, z_{ituv}^r) \) is an indicator vector. \( z_{ituv}^r = 1 \) if user \( u \) succeeds in activating user \( v \) at
\( t \) to diffuse message \( i \), and \( z_{ituv}^r = 0 \) if user \( u \) fails to
activate \( v \) within \( \{t_{ru} + 1, t\} \), where \( t_{ru} \) indicates the time \( u \) was activated to diffuse message \( i \).

We consider the probability that user \( u \) will succeed in
activating one of her fellowers \( v \) at time \( t \) \( (z_{ituv}^r = 1) \), by
considering \( u \)’s social role information:

\[
\phi_{ituv}^r = \sum_r \rho_r \lambda_r (1 - \lambda_r)^{t - t_{ru} - 1} \theta_{ur}
\]

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

\[
P(v \in A_{it}) = \prod_{u \in B(v) \cap D_{it-1}} (\phi_{ituv}^r + \varepsilon_{ituv}) - \prod_{v \in B(v) \cap D_{it-1}} \varepsilon_{ituv}
\]

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

\[
P(v \notin D_{iT}) = \prod_{u \in B(v) \cap D_{iT}} \sum_r (1 - \rho_r) \theta_{ur}
\]

For the social attribute generation part, we have:

\[
P(x_{uk}) = \sum_r \sqrt{\frac{\delta_k}{2\pi}} \exp(-\frac{(x_{uk} - \mu_{rk})^2}{2}) \theta_{ur}
\]

Based on Eqs. (3) to (5), we obtain the following likely-
hood function:

\[
L = \prod_{i=1}^{I} \prod_{t=1}^{T} P(v \in A_{it}) \times \prod_{i=1}^{I} P(v \notin D_{iT}) \times \prod_{u \in V} K \prod_{k=1}^{K} P(x_{uk}) \times \prod_{u \in V} \prod_{r=1}^{R} P(\theta_{ur}) \times \prod_{r=1}^{R} \prod_{k=1}^{K} P(\mu_{rk}, \delta_{rk})
\]
Model learning

We employ Gibbs sampling (Resnik and Hardisty 2010) to estimate the unknown parameters in the proposed model. Specifically, we begin with the posterior for sampling the latent variable \( r \) for each social attribute of a user \( u \):

\[
P(r_{uk} | r_{-uk}, x) = \frac{n_{r_{uk}} + \alpha}{\sum_r(n_{r_{uk}} + \alpha)} \frac{\Gamma(\tau_2 + n_{r_{uk}} - \frac{n_{r_{uk}}}{2})}{\Gamma(\tau_2 + n_{r_{uk}} - \frac{n_{r_{uk}}}{2})} \\
\times \frac{\sqrt{(\tau_1 + n_{r_{uk}})\eta(n_{r_{uk}}, \bar{x}_{r_{uk}}, \bar{s}_{r_{uk}})}}{\sqrt{(\tau_1 + n_{r_{uk}})\eta(n_{r_{uk}}, \bar{x}_{r_{uk}}, \bar{s}_{r_{uk}})}}
\]

(7)

where the counter \( n_{r_{uk}} \) (resp. \( n_{r_{uk}} \)) denotes the number of times \( r \) being sampled with (resp. the \( k \)-th social attribute of) user \( u \); \( \bar{x}_{r_{uk}} \) and \( \bar{s}_{r_{uk}} \) are respectively the mean and variance of the \( k \)-th social attribute associated with role \( r \).

The superscript \( -uk \) on the counters indicates exclusion of the current observation (resp. 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 the presentation of Eq. (7) and is defined as:

\[
\eta(\cdot) = [\tau_3 + \frac{1}{2}((n_{r_{uk}} + \tau_1)\bar{x}_{r_{uk}} - \tau_0)^2]^{\frac{1}{2}}
\]

(8)

In Eqs. (7) and (8), \( \tau \) is the parameter of normal-gamma prior. Similarly, we evaluate the posterior for sampling the latent variables \( (t, r, z) \) for each diffusion process:

\[
P(r_{iwv}, \Delta t_{iwv}, z_{iwv} | r_{i-uwv}, \Delta t_{i-uwv}, z_{i-uwv}, y) = \frac{n_{r_{iwv}} + \alpha}{\sum_r(n_{r_{iwv}} + \alpha)} \frac{\Gamma(\gamma_1 + n_{r_{iwv}})}{\Gamma(\gamma_1 + n_{r_{iwv}})} \times \frac{(n_{z_{iwv}} + \gamma_1)\prod_{\Delta t_{i-uwv}}^{t_{iwv}}(s_{z_{iwv}} - n_{z_{iwv}} + \gamma_0 + t) \times \Phi_{\Delta t_{i-uwv}}(\gamma_1 + s_{r_{iwv}} + \gamma_0 + t) \times \Phi_{\Delta t_{i-uwv}}(\gamma_1 + s_{r_{iwv}} + \gamma_0 + t)}{\prod_{\Delta t_{i-uwv}}^{t_{iwv}}(s_{z_{iwv}} - n_{z_{iwv}} + \gamma_0 + t) \times \Phi_{\Delta t_{i-uwv}}(\gamma_1 + s_{r_{iwv}} + \gamma_0 + t) \times \Phi_{\Delta t_{i-uwv}}(\gamma_1 + s_{r_{iwv}} + \gamma_0 + t)}
\]

(9)

where \( n_r \) (resp. \( n_{z_r} \)) denotes the number of times \( r \) sampled (resp. with \( z \)); \( s_r \) denotes the sum of \( \Delta t \) that has been sampled with \( r \). We use \( \Phi \) to indicate \( P(y_i | x_i, \Delta t_{i-uwv}) \) for brevity. Intuitively, \( \Phi \) is used to handle contradictions arise during the sampling process. Please refer to more details about \( \Phi \) and other implementation notes here.

We now estimate model parameters by the sampling results. The updating rules for \( \theta, \lambda, \) and \( \rho \) can be deduced as:

\[
\theta_{uw} = P(\hat{r} = r | r, \Delta t, z, y) = \frac{n_{ur} + \alpha}{\sum_r(n_{ur} + \alpha)}
\]

(10)

\[
\lambda_{vr} = P(\hat{\Delta \tilde{t}} = 1 | \hat{r} = r, r, \Delta t, z, y) = \frac{n_{vr} + \gamma_1}{\gamma_1 + s_r + \gamma_0}
\]

\[
\rho_{uv} = P(\hat{z} = 1 | \hat{r} = r, r, \Delta t, z, y) = \frac{n_{zuv} + \beta_1}{n_{zuv} + \beta_1 + n_{uv} + \beta_0}
\]

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

\[
\mu_{rk} \approx E(\mu_{rk}) = \frac{\tau_0 + n_{r_{uk}}}{\tau_1 + n_{r_{uk}}} \bar{r}_{rk},
\]

\[
\delta_{rk} \approx E(\delta_{rk}) = \frac{\tau_2 + n_{r_{uk}}}{\tau_1 + n_{r_{uk}}} \bar{s}_{rk}
\]

(11)

Experimental Results

All data and codes used here are publicly available.

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 tweets (posting logs) of over 200 million users. If there exists a following link from a user \( u \) to another user \( v \), 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 (tweets) 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 1st, 2011 as the training set, and those of November 2nd, 2011 as the test set to evaluate the 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 only present the results on the 4 most popular topics: campus, horoscope, movie, and history.

Tasks. We evaluate the proposed model, RAIN, based on the following two tasks. (1) At the micro-level, how accurate is RAIN in predicting whether a user will repost a given message? (2) At the macro-level, how well does RAIN predict the scale and duration of a diffusion process?

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 RAIN and several baseline methods (described below). Note that

\(^1\)http://arxiv.org/pdf/1104.1562v1
\(^2\)http://t.qq.com/
Table 2: Performance of repost prediction on several topics.

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

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 followers 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 message \( i \) based on three features: the number of \( v \)'s followers who have reposted \( i \), the number of \( v \)'s followers 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).

RAIN. This is the proposed social role-aware diffusion model. For each message \( i \), both this method and IC model use simulation to estimate 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 2 shows the performance of RAIN 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%). RAIN outperforms baselines by 32.6% in terms of MAP (mean average precision on all instances). 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. RAIN 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.

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 RAIN, which summarize the structural properties of users taking a certain role, to uncover the meaning of the latent roles learned by RAIN. For instance, a latent role with high PageRank score is considered to be representing the opinion leader role. Next, we group users into opinion leaders, structural hole spanners, and ordinary users. Finally, we use RAIN 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, RAIN 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 community to another more easily (e.g., “movie”, “history”, and “travel”).

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 RAIN can predict for each
Barbieri et al. (2013) studied social influence from a topic not a simple monotonic function of the number of her active neighbors, we show that the probability of a user becoming active is monotonically with the number of her active neighbors. However, according to the experiments conducted in this paper, the tendency of an inactive user to become active increases the proportion of the original posts with a particular number of reposts. Overall, RAIN performs better, while the baseline method tends to overestimate diffusion scale.

Scale and duration 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, RAIN performs better, while the baseline method tends to overestimate diffusion scale. 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.

Related Work

Recent years have seen extensive modeling efforts on the information diffusion (Lerner 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 (LT) models (Granovetter 1978) and Independent Cascade (IC) models (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 becoming active is not a simple monotonic function of the number of her active neighbors, but is relevant to the social roles of her followees.

Social influence and conformity is another related topic. Barbieri et al. (2013) studied social influence from a topic modeling perspective. Myers et al. (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. Tang et al. (2009) studied the problem of learning influence probabilities between users in social networks. Tang et al. (2013) further investigate how conformity influence users’ behaviors and Zhang et al. (2014) extended the problem with awareness of social roles. Rodriguez et al. (2013) applied the survival theory to generalize some existing diffusion models into a multiplicative model. In contrast to our work, these studies focus only on the diffusion process without considering how different types of users may influence such process.

Conclusion

In this paper, we study a novel problem of social role-aware information diffusion, with an emphasis on understanding the interplay between users’ social roles and their influence on information diffusion. We propose a social role-aware information diffusion (RAIN) 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.

Acknowledgements. Yang Yang and Jie Tang are supported by National 863 project (No. 2014AAA015103), National 973 projects (No. 2014CB440506, No. 2012CB316006, No. 2011CB302302), NSFC (No. 61222212), and a research fund from Huawei Inc. Qiang Yang and Cane Leung have been supported in part by National 973 project 2011CB340304 and Hong Kong RGC Projects 621013, 620812, and 621211. Yizhou Sun is supported by Yahoo! ACE Award and NEU TIER 1 Grant.
References


