Probabilistic Topic and Role Model for Information Diffusion in Social Network

No Author Given

No Institute Given

Abstract. Information diffusion, which tackles the issue of how a piece of information spreads and reaches individuals in or between networks, has attracted considerable research efforts due to its widespread applications, such as viral marketing and rumor control. However, the process of information diffusion is complex and its underlying mechanism is still unclear. One important reason is that social influence takes many forms and each form may be determined by different factors. One of the biggest challenges is how to capture all the crucial factors of a social network such as users' interests (which can be represented as topics), users' attributes (which can be summarized as roles) and users' reposting behaviors in a unified manner to model the information diffusion process . Yet in the existing approaches concerning the social network analysis, these crucial factors are considered separately and processed independently. In this paper, we comprehensively investigate the high correlation and mutual influence between users' interests and roles in diffusion process, which are defined by the contents of the reposted messages and her structural properties, respectively. Then, a novel Topic and Role Model (TRM) is proposed, which integrates user topical interest extraction, role recognition and information diffusion modeling into a unified framework. We develop a Gibbs sampling based algorithm to estimate model parameters, and then fit the model to a large and real social network using historical information diffusion data. Experimental results show the validity and competitiveness of our model, compared to several state-of-the-art alternative models.

Keywords: user topic; user role; information diffusion ; social network

1 Introduction

Information diffusion focus on how a piece of information (knowledge) spreads and reaches individuals in or between networks [26, 23]. With the rapid development of online networks, such as Twitter, Facebook, Sina Weibo etc., it becomes easier for people to share or obtain the information they are interested in by posting or retweeting messages online [1, 3, 25]. Information on social media propagates more quickly than in real society [11, 14]. The process of the information diffusion is crucial for spreading technological innovations [19], word of mouth effects in marketing [9], and opinion formulations [24]. In reality, the information diffusion process is complex, as is the influence of one user on another. Previous researches, such as [28, 8, 18] and empirical studies on online social networks, including Twitter [12, 14], Facebook [17], Sina Weibo [19], revealed many interesting phenomena and basic underlying laws. For example,in a social network,one user's final behavior decision is the balance between her own interests(which can be represented as topics) and her trusted friends with similar interests [4, 24, 28]. Users may have different attributes, for example, some users may be popular,and have many followers, but others may be different. How should we model the information diffusion process so that the model can capture the intrinsic relations between all these elements, such as individual attributes, users' topical interests and actions?

Information diffusion analysis has been attracting much interest from researchers. Some extensive efforts have been made in this field [24, 28]. From the perspective of structure of the network availability, information diffusion model can be classified into network structure based methods and non-structure based methods. Non-structure based approaches are limited by the fact that they ignore the topology of the network and only forecast the evolution of the rate at which information globally diffuses[3]. In network structure based methods, there are two representative models, namely Independent Cascade(IC) model [10]and Linear Threshold model(LT) model [13]. They assume that the network structure determines the flow of information and focus on the structure of the process. Since these two models require a diffusion probability between every two users, thus they have high computational complexity. To overcome this problem, [28, 24] introduce the topic model such as LDA [6] to make users with same topic distribution share the same behavior pattern. Zhang [27] employed social influence locality for modeling users' reposting behaviors by taking the phenomenon into consideration that users' behaviors are mainly influenced by close friends in their ego networks. Xu [24] assumed that the user posting behavior is mainly influenced by three factors: breaking news, posts from social friends and user's intrinsic interests. Tang [18] proposed a probabilistic factor graph model that takes the effects of user-to-user topic-wise influence strength for expert findings. Furthermore, topic-aware diffusion models assumed that either the topics associated with the diffusion process are specified in advance or independent with the user structural attributes [8].

Although much progress has been made, the results of existing work are not satisfactory, due to their limitations:

1. Most information diffusion models utilized only portions of the available social network information. For example, Zhang [27] only took the network structure information into consideration, ignoring the differences between the users themselves, such as user's preferences or interest, while [24] assumed that breaking news, posts from social friends and user's interests influence user posting behaviors, ignoring the conformity influence in information diffusion.

*2.*The highly volatile user behaviors usually render it difficult to accurately uncover diffusion patterns for the approaches between individuals. For example, some users seem very active, and prefer to repost the messages from her idoles, while others may not.

3. The underlying mechanism of information diffusion is still unclear. One important reason is that social influence takes many forms and each form may be determined by different factors.

In reality, the information diffusion process is complex, as is the influence of one user on another. Social influence happens when one's opinions or behaviors are affected by others. It is well known that different types of social ties have essentially different influence on social actions. [4, 24, 28]claimed that one user's final behavior decision is the balance between her own interests(which can be represented as topics) and her trusted friends with similar interests. Obviously, users with similar preferences are more likely to be friends than others in social networks [20, 7]. Topics are the collections of user's interests to post a message and provide the intentions for user's engagement in social networks [16, 24]. Moreover,the structure of the network affects how information may diffuse in a social network, in which the users' structural attributes determine their social roles in different communities [21, 26]. User's social role in turn affects the information diffusion process. For example, 1% users with high Pagerank score [15] acting as opinion leaders post 50% URLs on Twitters information diffusion [22].

Consequently, we face some very interesting questions: Is there any dynamics or mutual influence between user interests and social roles? To what extent do they influence the information diffusion process? If, for instance, a famous artificial intelligent expert and a normal political-science major student both retweet the same two messages: one message about AlphaGO and the other one about the presidential candidate Trump, will the followers of each user retweet the two different messages equally? Specially, will the two users have the same influence strength on their common followers ? This paper offers a new perspective.

We introduce a TRM model to uncover and explore temporal diffusion. We model topics and roles in a unified latent framework, and extract role-aware topic level influence dynamics. In this model, we group the users based on their structural properties and reposed information, and translate the calculations of individual level influence to the role-topic pairwise influence, which can provide a coarse-grained diffusion representation. These effective technologies facilitate our TRM model to accurately capture the backbone of information diffusion, as well as better predict and analyzes the diffusion.

Topics and social roles are both hidden. Pipeline approaches to extract these two factors in sequence fail to capture their interdependence. Though in recent years an array of techniques [26, 7, 16, 28] have been developed for jointly leveraging these two critical factors, they all fall short of properly modeling the correlations between them. Besides the task of simultaneously extracting topics and social roles, we are even required to model the information diffusion process with temporal factors.

To tackle the above issues, this paper develops a latent model, TRM(Topic and Role Model), to extract topics, social roles and role aware topics dynamic in a unified way. We model topics and social roles as latent variables, and set up a generative process for observed network, contents of the reposted messages and time to accurately characterize the role-aware topic level information diffusion process.

Based on the extraction, we design an effective diffusion prediction method. Extensive experiments on large dataset show that our method greatly improves the prediction accuracy. To summarize, we make the following contributions:

- **Novel Perspective.** We systematically study on the building joint models to explore mutual influence for user topics and roles. It brings up a new into the information diffusion process. To the best of our knowledge, such a new angle has not been studied before.
- **Comprehensive Model.** We propose a latent model to uncover the hidden topics and roles as well as capture the information diffusion, which can model the process of information diffusion better than other models. We further devise a Gibbs sampler to estimate the parameters.
- **Inspiring Prediction & Exploration** An effective diffusion prediction approach is developed which leverages the information diffusion patterns with user topics and social roles. We conduct extensive experiments to validate the proposed model over several baselines by employing two large real-world network as experimental datasets. Experimental result shows that the proposed model performs much better than the baseline methods.

2 Related Work

Information diffusion. Although online information diffusion has received increasing interest over the recent years [26, 10, 13, 24], modeling the diffusion in social networks remains an open problem. Central to information diffusion is the estimation of influence strength. In representative information diffusion models, such as Independent Cascade(IC) model [10] and Linear Threshold model(LT) model [13]. Both types of models have the computational problem is that of selecting the set of initial users that are more likely to influence the largest number of users in the social network [4], and also have the over-fitting problem resulting from their large number of unknown parameters to learn. TRM addresses these two problems by allowing users with the same social role and user topical interests to share the same diffusion patterns, thus greatly reduce the number of parameters. Social influence happens when one's opinions or behaviors are affected by others.

Topic-aware influence. While most of the above work utilizes the network structural and timing information to model the information diffusion process, a different line of work has considered analyzing the available textural information and use the latent topics of the messages as the user's needs or interests [23, 16, 4]. Topics are the collections of user's interests to post a message and provide the intentions for user's engagement in social networks [16, 24]. Furthermore, users with similar preferences are more likely to be friends than others in social networks [20, 7].In [23], the authors assumed that the user's retweeting actions are mainly influenced by the three factors: breaking news, posts from social friends and user's intrinsic interest, and proposed a mixture latent topic model to predict the user's repost behaviors. In [16], the authors applied the Hawkes process to model the information diffusion process based on the latent topics of the user's. Most of the topic-aware information diffusion models take into considerations the topic of the user or the twitter, and neglect the user's structural attributes. In contrast to these models, we focus on the diffusion process not only with considering how the topical interests may influence such process, but also considering the different roles of users. Especially, the social role and user topical interest distributions of each user are not only determined by her structural attributes and the contents of the reposted messages respectively, but also by her diffusion behaviors.

3 TRM diffusion model

Online social networks, such as Twitter and Sina weibo, have become the central nexus for discussion of the topics of the day. On social networks, users from all over the world tweet or retweet a variety of topics of interests. Naturally, each user has distinct topical interests or personalized preference. To characterize the heterogeneity among all users, we model each user possesses a distinct probability that he or she is interested in retweetting messages. For example, consider a mini set of two topics: sports and movie. One may retweet the sports topic with a higher probability than the movie topic, while another user may be more interested in retweeting movie than retweeting sports. Given a set of topics, a user generates each word in their tweets from one of the topics based on the distribution specific to this topic. Moreover,the structure of the network affects how information may diffuse in a social network, in which the users' structural attributes determine their social roles in different communities [21]. It is well known that different types of social ties have essentially different influence on social actions. User's social role in turn affects the information diffusion process. For example, 1% users with high Pagerank score [15] acting as opinion leaders post 50% URLs on Twitters information diffusion [22]. Furthermore, social roles and diffusion are not independent of each other in nature. Intuitively, each user may play multiple roles with respect to different structural attributes, thus exhibiting different influential strengths in different diffusion process. For instance, one with high PageRank scores may act as opinion leader when she post on her area of expertise. 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.

3.1 Formulation

Let G = *(V,E,X)* where V is the set of all the users and $E \subseteq V \times V$ is a set of relationships between users. Each factor $e_{ij} =< v_i, v_j >\in E$ represents user v_i follows v_j, in other words, v_i is the follower of v_j and v_j is the followee of v_i in turn. Each user v_i has H-dimensional attribute vector x_i, where H is the number of all attributes. Each factor $x_{vh} \subset X$ denotes the $h-$th attribute of user v. We can define the user's attributes such as Pagerank Score [[15]], in degree

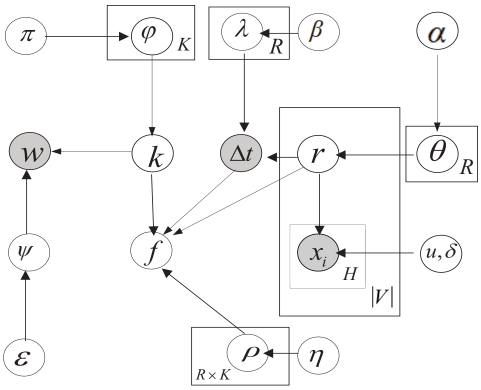


Fig. 1: TRM model

and network constraint score [[7]], based on the structure of the social network. For each user $v \in V$, we use $N(v) = \{u|u \in V, e_{uv} \in E\}$ to denote the set of followees of v. For a message, Whether a user activate her followers may also depend on the role she plays and the her intention she chooses.

To model the intuition that a user may have different interest topics and take different roles in information diffusion process, we associate each with an interest topic and social role distribution respectively:

Definiton 1 ***Topic distribution***. In social networks, a user is usually interested in multiple topics. Formally, each user is associated with a vector $\phi_v \in V^K$, where K is the number of topics($\sum_k \phi_{vk} = 1$).

Definiton 2 ***Role***. Each user may play multiple different roles, denoted as $r = [1, 2, ..., R]$. Each role has a set of parameters for the distribution the attributes conform to. Here we use Gaussian distribution. If a user plays role r, its h-th attribute conforms to $(u_{rh}, {\delta_{rh}}^{-1})$.

Definition 3 ***Role distribution***. Each user has a multinomial distribution over roles, which is denoted as θ. θ_v denotes the probability for user v to play role r, and is subject to $\sum_r \theta_{vr} = 1$.

Definition 4 ***Topic-Role Pair***. Whether a user v would repost a message posted by her followee u depends on the $u's$ role she plays and the topic she chooses. We use ρ to denote the distribution of topic-role pairs over reposting actions. In information diffusion process, the actions of reposting messages only contains two cases, so we can use a Bernoulli distribution to model the distribution of topic-role pairs over actions. In other words, $\rho_r k$ denotes the influence strength that a user plays role r and chooses the topic k to successfully activate one her follower for a message.

3.2 Model description

Based on the above definitions, we explain the proposed TRM model. Our goal is to devise a probabilistic generative model for extracting the user topical interests, learning user social roles and modeling information diffusion simultaneously. Fig.(1) illustrates the model, where f is a diffusion function. TRM determines user topical interests of each user according to both her reposted messages and her behaviors in diffusion process. The basic idea is that the user's actions of reposting messages not only depend on the social role she plays, but also depend on her topical interests. We use the content of user's reposted messages to determine her topic distribution and use user's attributes to determine her role distribution, which are all then used as priors to guide the sampling for user's actions. Overall, the TRM model we proposed consists of three parts: the user's messages generation ,the user's attributes generation and modeling the information diffusion process.

Table 1: Notations

SYMBOL	DESCRIPTION
R, K, H, W	number of social roles,topics,attributes and unique words in the dataset
T	the largest timestamp in a given diffusion model
N_d	the number of words in the dth messages
e_{iuv}^{t}	a latent variable denoting whether user v reposts the message i posted by user u at time t
k_d	topic associated with post d
ϕ_v	Multinomial distribution over topics specific to user v
ψ_k	Multinomial distribution over words specific to topic k
θ_v	Multinomial distribution over roles specific to user v
λ_r	geometric distribution over Δt associate with role r
ρ_{rk}	Bernoulli distribution over decision to repost a message associate with topic k and role r
u_{rh}	mean of h-th attribute specific to role r
δ_{rh}	standard deviation of h-th attribute specific to role r

User's messages generation Here, we associate a single hidden variable with each message to indicate its topic due to the limitations of the length of characters in a single message. The generative process of the message d posted by user v can be described as follows:

1. Draw a ϕ_v from from Dirichlet (π)
2. Draw a topic k from multinomial distribution ϕ_v
3. Draw a ψ_k from Dirichlet (ε)
4. For each word $n = 1, ..., N_d$, draw a word w_{dn} from mulitnomial distribution ψ_k.

Social attributes generation Each user may play several roles in different information diffusion processes and is subject to a certain distribution over attributes, denoted by θ_v. Each user is a random mixture over roles and can be denoted by $v = (x_1 ... x_h ... x_H)$. For each individual attribute of user v, we first generate $v's$ role distribution θ_v, then generate a latent role r from user v's role distribution θ_v, next draw the value of attribute h from the normal distribution $(u_{rh}, {\delta_{rh}}^{-1})$. The generative process of the value of attribute h for user v in a social network can be described as follows:

1.Draw a θ_v from Dirichlet (α)

2.Draw a role r from multinomial distribution θ_v.

3.Draw the value of attribute h over role r: $x_{rh} \sim (u_{rh}, {\delta_{rh}}^{-1})$

Model information diffusion process Inspired by our exploratory analysis, which reveals that topical interests and social roles of a user affect the information diffusion process. Specially, we incorporate the user topics and roles in the information diffusion process. We introduce topic-role parameters ρ_{rk} denotes the probability that one user plays role r successfully activate another user specific to topic k and a per-role parameters λ_r denotes the probability that cause a 1-timestamp delay in information diffusion. At anytime, user v will become active at least one of her followees activate her successfully. We use Independent Cascade Model as diffusion function f in TRM model.

More specifically, we first generate the influence strength and diffusion delay corresponding to ρ_{rk} and λ_r, respectively. Consider a message i posted by user u at time t, u will have only one chance to activate her follower v. Firstly, user u draws role r she will play from role distribution θ_u. Next, draws the topic k the message will be interested in for $u's$ follower u from multinomial distribution ϕ_u, and then generate a time delay Δt from geometric distribution λ_r. At time $t' = t + \Delta t + 1$, we draw the e^t_{iuv} from Bernoulli(ρ_{rk}) to determine whether user u will succeed in activating user v.

For each action of repost :
1. Draw a role r from multinomial distribution θ_u.
2. Draw the topic k from multinomial distribution ϕ_u.
3. Draw ρ_{rk} from beta(β) prior.
4. Draw the λ_r from beta(η).
6. Draw Δt from geometric distribution λ_r.
7. Toss a coin, draw the e^t_{iuv} from Bernoulli(ρ_{rk}).

3.3 Model Learning

Learning the model is to find a configuration for the parameters $\{\theta, \phi, \rho, \lambda\}$ to maximizes the log-likelihood objective function. The posterior probability of k_d, which denotes that the latent topic k for the post d of user u to activate her follower is calculated by:

$$p(k_d = k|k_{-ud}) = \frac{n_{-uk} + \pi}{\sum\limits_k (n_{-uk} + \pi)} \times \prod_{n=1}^{N_d} (\frac{n^k_{w_{dn}} + \beta}{\sum\limits_w (n^k_w + \beta)}) \tag{1}$$

where the counter n_{uk} denotes the number of times topic k being sampled with user u; w_{dn} means the n-th word in post d and n^k_w denotes the number of times word w being assigned to topic k. The superscript$-uk$ on the counters indicates exclusion of the current observation(resp.the message d posted by user u) from the counts.

According to [[5]] ,we adopt :

$$\begin{aligned} u_{rh} &\approx E(u_{rh}) = \frac{\tau_0\tau_1 + n_{rh}\overline{x}_{rh}}{\tau_1 + n_{rh}} \\ \delta_{rh} &\approx E(\delta_{rh}) {=} \frac{2\tau_2{+}n_{rh}}{2\tau_3{+}n_{rh}s_{rh} + \frac{\tau_1 n_{rh}(\overline{x}_{rh} - \tau_0)^2}{\tau_1 + n_{rh}}} \end{aligned} . \tag{2}$$

Similarily, after Gibbs sampling, parameters $\{\theta, \phi, \rho, \lambda\}$ can be estimated by:

$$\begin{aligned} \theta_{ur} = \frac{n_{ur} + \alpha}{\sum\limits_r (n_{ur} + \alpha)} \quad \phi_{uk} = \frac{n_{uk} + \pi}{\sum\limits_k (n_{uk} + \pi)} \\ \rho_{rk} = \frac{n_{erk(e=1)} + \beta_1}{n_{0rk} + \beta_0 + n_{1rk} + \beta_1} \quad \lambda_r = \frac{n_r + \eta_1}{s_r + \eta_0 + \eta_1} \end{aligned} . \tag{3}$$

where $\overline{r}, \overline{k}, \overline{\Delta t}, \overline{e}$ respectively represent a new observation of $r, k, \Delta t$ and e.

4 Experiment

4.1 Experimental setup

We evaluate the effectiveness of the proposed model on two real-world datasets belonging two different social networks:

•**Weibo** is a dataset from Sina Weibo Website, the largest microblogging service in China. The Weibo data we used in our experiment is from [19] with 66,348 users, 13,487,120 repost actions. We select the original posts that were reposted by larger than 6 users, and use remaining 129,560 original posts for experiments. For a given twitter from a user, we would like to predict who will repost the twitter.
•**Citation Network Database(CND)**[28] is extracted from DBLP, ACM, MicroSoft Academic Graph, and other sources. We select the original paper that were citated by larger than 6 users, and use remaining 67,414 original papers for experiments. For a given paper, we would like to predict which author will cite this paper next.

Since the retweet or cite action prediction is much similar a ranking problem, we prefer the precision at top ranked results as the evaluations of our proposed model. Specially, given a message or a paper i produced by user v, we calculate the retweet or cite probability of each $v's$ followers, and we use P@10(Precision of top-10 predictions), P@50, P@100, and MAP(Mean Average Precision) to evaluate the ranking prediction results for each message and aggregate the results for all messages together.

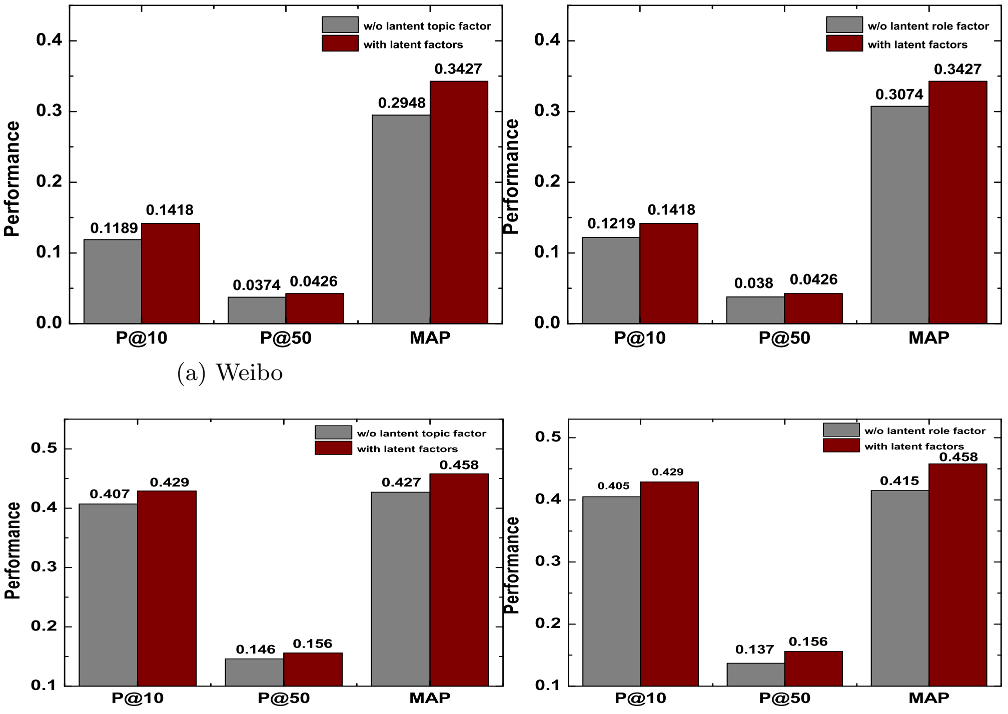
We compare TRM with several representative methods for user's prediction Count, LDA [6], MUPB [24],Social Role-Aware Information Model(Rain) [26].
•**Count:** In Count, the probability of a user reposting a message is in direct proportion to the number of followees who have reposted message i. This method assumes that a user's reposting decision only depends on her followees.
•**LDA:** In LDA [6], the probability of a user reposting a message based on the topic distributions of message i and v's all reposted posts in the training data set. This method only takes contents of the messages into account, and ignores the whole and local social network structural attributes.
•**MUPB:** MUPB [24] assumed that the user posting behavior is mainly influenced by three factors: breaking news, posts from social friends and user's intrinsic interests. Although this method can to some extent help extract the topics of users, however, it failed to capture the real motivation of users to publish content, as user behavior can easily affected by the structure of the network, other than user interest.
•**Rain:** Rain [26] predicts whether user v will repost message i based on the user $v's$ role distribution and the information diffusion attributes corresponding to each role. This method ignored the user's intrinsic interests satisfactions by reposting the message i.
•**TRM:** This is the proposed method. We empirically set the model parameters as $R = 10, K = 10, \alpha$=0.1, π=0.1, ε=0.1 $\beta = (1, 1), \eta = (1, 1)$

Table 2: Performance comparison on two datasets evaluated by Precision@n and MAP.

	Weibo				CND			
Method	P@10	P@50	P@100	MAP	P@10	P@50	P@100	MAP
Count	0.007	0.006	0.006	0.013	0.089	0.029	0.017	0.127
LDA	0.112	0.039	0.020	0.085	0.153	0.049	**0.030**	0.198
MUPB	0.405	0.137	0.079	0.415	0.122	0.038	0.022	0.307
Rain	0.407	0.146	0.083	0.427	0.121	0.038	0.021	0.299
TRM	**0.429**	**0.156**	**0.088**	**0.458**	**0.143**	**0.043**	0.024	**0.345**

4.2 Experimental Results

Better Performance. The performance comparison on the two datasets evaluated by P@10, P@50, P@100 and MAP is illustrated in Table 2. We can discover that TRM model clearly outperforms Count, LDA, MUPB, and Rain nearly on all metrics in Weibo (+0.076- 0.445 improvement



(a) Weibo

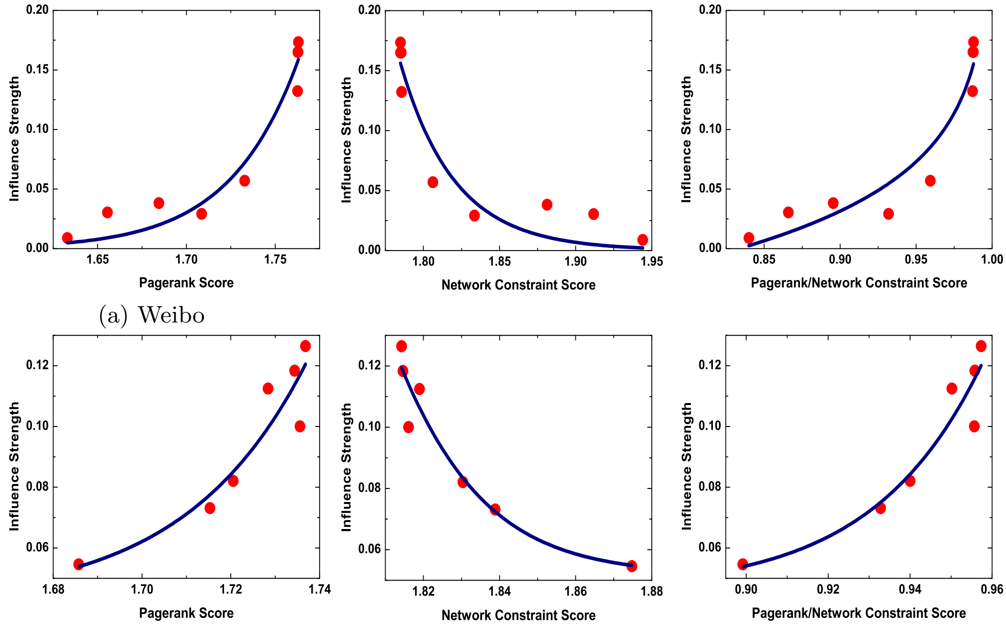
(b) CND

Fig. 2: Effect of Mutual Influence: the contribution of topic and roles on repost prediction.

in terms of average MAP) and CND(+0.049-0.218 improvement in terms of average MAP). Due to the lack of supervised information, Count performs worst on both datasets. Whereas, the Count and LDA model outperform better in CND than Weibo. Since a author usually only focuses one or two fields of study, thus the users in CND database usually cite the papers with the similar topics. The Count and LDA model all ignore the user's social structural attributes. Prediction of the user's reposts action based on LDA only depends on the user's history reposts, and ignores the situation where a user needs or topic distributions may change over time. TRM also outperforms MUPB and Rain on all metrics. Although MUPB and Rain also considers user topics and social roles, respectively, they still ignores the correlations and mutual influence between topics and roles.

Effect of Mutual Influence. We also examined the nature and the effectiveness of the associated latent factors on the mutual , and Fig.(2) demonstrates their feasibility in our modeling. Note that if we do not incorporate the latent role or topic factor, our TRM model becomes the traditional MUPB or Rain approach, respectively. It shows that the latent factors consistently enhance the precision(evaluated by P@10,P@50 and MAP)for the repost prediction. For example, the latent topic and role factor significantly improve the MAP by 3.1%(from 42.7% to 45.8%), and improve the MAP by 4.3%(from 41.5% to 45.8%) for the repost prediction, respectively. These results also illustrate that the user's topics and social roles are very crucial to model the information diffusion.

Social role analysis. The learned parameters ρ represent the influence strength of a user for different topics and roles. The method also learns u_{rh}, which denotes the mean value of social attribute k for role r, and ρ_{rk} denotes the topic-role pair activate probability of topic k for



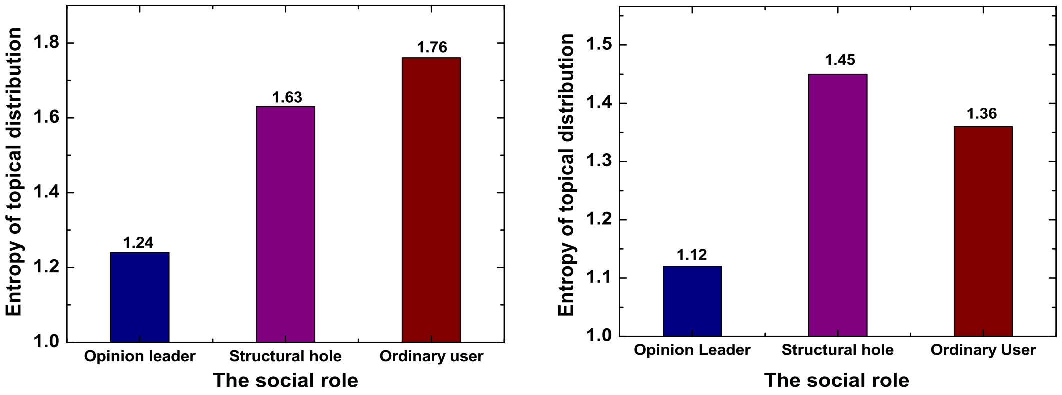
(a) Weibo

(b) CND

Fig. 3: Role analysis on two datasets: the correlation of influence strength with *Pagerank Score, Network Constraint Score and Pagerank/Network Constraint Score*

role r. So we can use the $P(r) = \sum_k \rho_{rk}$ to denotes the influence strength for role r. Fig.(3) shows the correlations between a role's social attribute and its influence strength. We discover that the correlation follows a logarithm function. We try different forms of functions to fit the remaining data points and select logarithm function of R^2. When fitting the data points, we first remove the roles with a small number of related users. As Fig.(3a), Fig.(3b) show that people who have larger Pagerank score or smaller Network constraint score will have greater influence in the information diffusion process than ordinary people. This can be explained that people with larger Pagerank score tend to have more followers, and her posted messages will be more likely to be reposted. And people with smaller network constrain score tend to be a structural hole to connect two or more communities, and her posted messages will more easily be propagated to different social network communities.

Correlation between user topic and social role. We further study the correlation between user topic and social role. To conduct this experiment, we first analyze estimated Gaussian parameters of TRM to uncover the meaning of the latent roles learned by TRM. Inspired by the work of [15, 7, 28, 26]], we group users into opinion leader, structural hole spanners and ordinary users. For instance, a latent role with small network constraint score is considered to be representing the structural hole spanners. Furthermore, the learned parameter ϕ represent the topic distribution for different users. Inspired by [12], we compute the entropy of user's topical distributions to measure how much topical a user's interests or topics are. For a user v, the entropy of her topical interests distribution is computed as follows: $Hp(v) = -\sum_{k}^{K} \phi_{vk} \log(\phi_{vk})$. To further analysis the correlations between users' topics and roles, we continue to calculate the average entropy for each role as follows:$Hp(r) = \frac{1}{|N_r|} \sum_{v \in A_r} Hp(v)$, where A_r denotes the set of



(a) Weibo

(b) CND

Fig. 4: The average entropy of topical distributions for each role on two datasets.

users to be assigned to role r, N_r denotes the number of users in A_r. The higher the entropy, the less topical the role are. Fig.(4) demonstrates the correlations between user topics and roles. The higher the entropy, the less topical the role are. It follows that the most topical would be a user that is interested in only a single topic, whereas the least topical would be a user is interested in all topics with equally preferences. Thus the phenomenon in Fig.(4) may be explained as follows: most of opinion leaders post the messages about their areas of expertise and they tend to focus on regional and specialized topics, the structural spanners have broad interests than opinion leaders since they usually focus on more general topics which tend to propagate from one community to another more easily, the ordinary users have more broad interests since they behave more randomly. In Fig.(4b), the structural spanner has higher entropy value than ordinary users. The phenomenon may be explained as: when a person have published lots of papers in different regions, she may become more open-minded and tend to accept new ideas from others.

5 Conclusions

In this paper, we study how to model the information diffusion process by capturing the user topics and social roles. We devise a probabilistic generative model framework, TRM to extract the user topics, recognize the social roles and model information diffusion. Our model allows for mining the correlations between users' topics and roles. Our experiments on a real social network data set show some interesting results that opinion leader tend to be more topical than the other roles, and the structural hole spanners tend to focus on more general topics. In addition, our method also outperform state-of-art baselines on the task of repost prediction in a large real social network.

References

1. F. Abel, Q. Gao, G. J. Houben, and K. Tao. Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In *International Web Science Conference*, pages 1–8, 2011.
2. E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(5):1981–2014, 2008.

3. C. Aslay, N. Barbieri, F. Bonchi, and R. Baeza-Yates. Online topic-aware influence maximization queries. In *International Conference on Extending Database Technology*, 2014.
4. N. Barbieri, F. Bonchi, and G. Manco. Topic-aware social influence propagation models. *Knowledge and Information Systems*, 37(3):555–584, 2012.
5. J. M. Bernardo and A. Smith. Bayesian theory, vol. 405, 2009.
6. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
7. R. S. Burt. *Structural holes: The social structure of competition*. Harvard university press, 2009.
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*, 8(6):666–677, 2015.
9. P. Domingos and M. Richardson. Mining the network value of customers. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66, 2001.
10. J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.
11. M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *Acm Transactions on Knowledge Discovery from Data*, 5(4):1019–1028, 2012.
12. P. A. Grabowicz, N. Ganguly, and K. P. Gummadi. Distinguishing between topical and non-topical information diffusion mechanisms in social media. 2016.
13. M. Granovetter. Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443, 1978.
14. H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010.
15. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.
16. J. C. L. Pinto and T. Chahed. Modeling multi-topic information diffusion in social networks using latent dirichlet allocation and hawkes processes. In *Tenth International Conference on Signal-Image Technology and Internet-Based Systems*, pages 339–346, 2014.
17. V. Sekara, A. Stopczynski, and S. Lehmann. Fundamental structures of dynamic social networks. *Computer Science*, 2015.
18. J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Acm Sigkdd Conference on Knowledge Discovery & Data Mining*, pages 807–816, 2009.
19. C. Tu, Z. Liu, and M. Sun. Prism: Profession identification in social media with personal information and community structure. In *Social Media Processing*, pages 15–27. Springer, 2015.
20. M. A. Wahba and L. G. Bridwell. Maslow reconsidered: A review of research on the need hierarchy theory. In *Academy of Management Proceedings*, volume 1973, pages 514–520. Academy of Management, 1973.
21. S. Wasserman and K. Faust. Social network analysis methods and applications. *Contemporary Sociology*, 91(435):219–220, 1994.
22. S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 705–714. ACM, 2011.
23. F. Xiong, Y. Liu, Z.-j. Zhang, J. Zhu, and Y. Zhang. An information diffusion model based on retweeting mechanism for online social media. *Physics Letters A*, 376(30):2103–2108, 2012.
24. Z. Xu, Y. Zhang, Y. Wu, and Q. Yang. Modeling user posting behavior on social media. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 545–554. ACM, 2012.
25. J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *International Conference on Weblogs and Social Media*, 2010.
26. Y. Yang, J. Tang, C. W.-k. Leung, Y. Sun, Q. Chen, J. Li, and Q. Yang. Rain: Social role-aware information diffusion. In *AAAI*, pages 367–373, 2015.
27. J. Zhang, B. Liu, J. Tang, T. Chen, and J. Li. Social influence locality for modeling retweeting behaviors. In *International Joint Conference on Artificial Intelligence*, pages 2761–2767, 2013.
28. J. Zhang, J. Tang, H. Zhuang, W. K. Leung, and J. Li. Role-aware conformity influence modeling and analysis in social networks. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.