# Social-Relational Topic Model for Social Networks

Weiyu Guo[1,2], Shu Wu[1], Liang Wang[1], Tieniu Tan[1]
[1]Center for Research on Intelligent Perception and Computing,
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China.
[2]College of Engineering and Information Technology,
University of Chinese Academy of Sciences, Beijing 100049, China.
weiyu.guo@ia.ac.cn, {shu.wu, wangliang, tnt}@nlpr.ia.ac.cn

## ABSTRACT

Social networking services, such as Twitter and Sina Weibo, have tremendous popularity in recent years. Mass of short texts and social links are aggregated into these service platforms. To realize personalized services on social network, topic inference from both short texts and social links plays more and more important role. Most conventional topic modeling methods focus on analyzing formal texts, e.g., papers, news and blogs, and usually assume that the links are only generated by topical factors. As a result, on social network, the learned topics of these methods are usually affected by topic-irrelevant links. Recently, a few approaches use artificial priors to recognize the links generated by the popularity factor in topic modeling. However, employing global priors, these methods can not well capture the distinct properties of each link and still suffer from the effect of topic-irrelevant links. To address the above limitations, we propose a novel Social-Relational Topic Model (SRTM), which can alleviate the effect of topic-irrelevant links by analyzing relational users' topics of each link. SRTM jointly models texts and social links for learning the topic distribution and topical influence of each user. The experimental results show that, our model outperforms the state-of-the-arts in topic modeling and social link prediction.

## Keywords

Topic Modeling; Social Networks; Social Link Generation

## 1. INTRODUCTION

Exploring users' topics on social networks from rich texts and social links is important for marketing activities in real applications. Although a lot of works, e.g., Link-LDA [1], RTM [2] and RankTopic [3], have been proposed for this task, most of them assume that the social links are purely caused by topical factors and are used as the supplement of texts in topic inference. Clearly, this assumption is not suitable for social networks, since many topic-irrelevant factors which can lead to a link generation. For example, on Twitter, President Obama associates with political topics and has a lot of followers. But, as far as we know, many of Obama's followers are not really interested in politics, and these following relationships may be caused by the so-called bandwagon effect [4]. As a result, if we could not recognize such topic-irrelevant links which are not really caused by users' interesting topics, the learned topics tend to be affected by these noise links and are not reliable to reveal personalized interests.

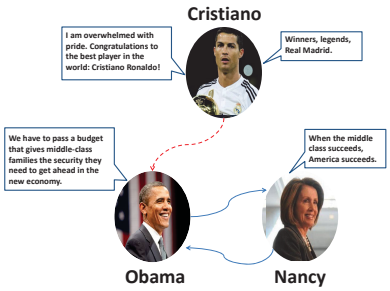


**Figure 1: A toy example of Twitter data, which shows several tweets and following relationships among three users. In this example, Obama and Nancy (Minority Leader of the United States House of Representatives) are following each other, and Cristiano (a football star) is following Obama.**

Recently, some approaches are proposed for dealing with topic-irrelevant links in topic modeling. For example, FLDA [5] embeds a Bernoulli prior to judge whether following links on Micro-bloging platforms are generated by popularity factors. However, since the factors of link generation on social networks are various, the methods which only rely on prior knowledge and can not well reveal the characteristics of each link. Intuitively, many topic-irrelevant links can be indicated by the texts of their relational users. As shown in Figure 1, Obama and Nancy have published similar political views in their tweets, and they follow each other may because of having similar political topics. While Cristiano mainly publishes tweets about football, he follows Obama may be caused by adoring the political celebrity. Therefore, the topical similarity of users, may be in turn to indicate whether the links between users are caused by topical factors.

In this paper, for alleviating the effect of topic-irrelevant links and obtaining reliable topics, we propose a novel SRTM, which can assess whether a specific link on social networks is caused by topical factors, and jointly model the texts and social links into a unified generative process. Moreover, except


*CIKM'15,* October 19˝ 23, 2015, Melbourne, Australia.

DOI: http://dx.doi.org/10.1145/2806416.2806611.

for individual topic distribution, our model can learn the individual influence on different topics and the global popularity of each user. To systematically assess our model, we conduct comparisons with several state-of-the-arts in topic modeling using the perplexity metric and social link prediction with ranking metrics.

## 2. RELATED WORK

Recently, a mass of hybrid data that contain textual information and social information are aggregated into social networking Websites. Topic models, such as Link-LDA [1] and RTM [2], which can discover a given number of topics from data sources, are often used to process such hybrid data. Link-LDA views the citations of a document as a kind of words, and uses LDA [6] to deal with such special words. With a pure topic relational assumption, Chang et al. propose RTM model for document networks, which draws topics for citations according to relational topic distributions of citations. However, the social links on social networks, e.g., Micro-blogs, are more complex than the citations. Many factors may lead to the generation of a social link, such as the bandwagon effect or the marketing advertisement. Thus, FLDA [5] improves Link-LDA by introducing a Multinomial-Bernoulli prior to assess whether a following relationship is caused by topical factors. Different from FLDA using an artificial prior to analyze social links, we leverage the similarity of relational topic distributions of each link to measure whether the link is generated by topic-irrelevant factors.

Ranking based models [3, 7] are another direction for dealing with hybrid data. They usually use topic modeling for analyzing the textual information, and conduct ranking algorithms, e.g., pagerank [8], on the structure information. For example, RankTopic [3] uses basic topic model to explore the topic distribution of each node from its texts, and then conducts Topic Sensitive PageRank [9] on the citation network to obtain more reliable topics. In [7], Yan et al. learn the topic distribution of tweets using LDA, and then compute the topical influence of each tweet on a heterogeneous graph by pagerank. However, since ranking based models often assume that the links are purely generated by topics, their learned topics tend to be effected by the topic-irrelevant effect. Moreover, due to the large scale social networks, these models often suffer from a slow convergence.

## 3. SOCIAL-RELATIONAL TOPIC MODEL

In this section, we present the Social-Relational Topic Model (SRTM), which jointly models the texts and social links. Through this model, we estimate the topic-irrelevant links, infer the topic distribution for each user, and analyze the users' influence on different topics.

### 3.1 Definition of SRTM

Our model consists of three major components with each capturing one perspective of our targets. For one following relationship $l$, the graphical representation of our model is shown in Figure 2. The model embeds two LDA models for processing the texts of relational users on either side of the model. In the middle part, the model estimates whether the link associates with topical factors and analyzes users' influence on different topics. More specifically, in SRTM, each user $u$ is viewed as a mixture of $K$ latent topics, i.e., $\theta_u \in \mathbb{R}^K$, corresponding to the texts and social links. SRTM

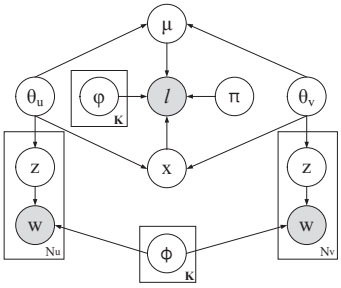


**Figure 2: Graphical representation of SRTM. The gray symbols denote the observed variables, i.e., $w$ denotes a word in texts and $l$ is the following link from the user $u$ to the user $v$. $\theta_u$ and $\theta_v$ denote the topic distribution of $u$ and $v$, respectively. $z$ is a topic in texts, and $x$ denotes a topic on following links. $\phi$ is the topic-word distribution, and $\psi$ is the topic-user distribution. $\pi$ is the Multinomial distribution over users, which indicates the popularity of users. $\mu$ is a switch variable which follows a Bernolli prior and relies on the topic distribution of two related users. $N_u$ and $N_v$ are the number of words in the texts of user $u$ and user $v$, respectively.**

generates a following relationship and estimates whether it is a topic-irrelevant link, by taking relational topic distributions, i.e., $\theta_u$ and $\theta_v$ into account. In a nutshell, the generative process of SRTM is summarized in Algorithm 1, where $\sigma$ is a sigmoid function, i.e., $\sigma(x) = 1/(1 + exp(-x))$, and $\rho_{u,v} = \theta_u \theta_v + \tau$.

**Algorithm 1** Generative process of SRTM

```
Draw π ~ Dir(ϵ);
for all each topic k = {1, 2, ..., K} do
    Draw φ ~ Dir(β);
    Draw ψ ~ Dir(δ);
end for
for all each user i = {1, 2, ..., M} do
    Draw topic proportions θ_i|α ~ Dir(α);
    for all each word w_{i,n} in the texts of u do
        Draw an assignment z_{i,n}|θ_i ~ Mult(θ_i);
        Draw a word w_{i,n}|z_{i,n}, φ ~ Mult(φ_{z_{i,n}});
    end for
end for
for all each user u = {1, 2, ..., M} do
    for all each following relationship l_{u,v} of u do
        Draw a switch μ_{u,v}|θ_u, θ_v, τ ~ Ber(σ(ρ_{u,v}));
        if μ_{u,v} == 1 then
            Draw an assignment x_{u,v}|θ_u, θ_v ~ Mult(θ_u ⊙ θ_v);
            Draw a followed user e_{u,v}|x_{u,v}, ψ ~ Mult(ψ_{x_{u,v}});
        else
            Draw a followed user e_{u,v} ~ Mult(π_v);
        end if
    end for
end for
```

To generate texts of users, each user is treated as a mixture of latent topics from which words are drawn. Similar to LDA, for the $u$-th user, the model first draws the topic distribution $\theta_u$ from a Dirichlet prior with a hyper-parameter $\alpha$. Then, to generate the $n$-th word in the texts of the user, a topic assignment $z_{u,n}$ is drawn from $\theta_u$. Finally, the word $w_{u,n}$ is picked from the topic-word distribution $\phi_{z_{u,n}}$.

Clearly, the topic distribution is affected by the social information, and the generative mechanism of following relationships are different from the words in textual contents.

In Algorithm 1, we sculpture a two-stage stochastic process for the generation of following relationships. Specifically, we first assess whether a following relationship is related to users' topics, using a switch variable $\mu$. For a following relationship $l_{u,v}$, there is a switch $\mu_{u,v}$ which is drawn from a Bernoulli prior. If $\mu_{u,v} = 1$, the following relationship is assumed to be associated with the user's topics, then a topic assignment $x_{u,v}$ is drawn according to the relational topic distributions, i.e., $\theta_u$ and $\theta_v$. Indicated by $x_{u,v}$, the following relationship $l_{u,v}$ is sampled from the Multinomial distribution $\varphi_{x_{u,v}}$, which corresponds to the topic-specific influence of users. If $\mu_{u,v} = 0$, the following relationship is viewed to be generated by topic-irrelevant factors, and is sampled from another Multinomial distribution $\pi$, which mainly indicates the global popularity of users. Moreover, according to the observation in Figure 1, the similarity of textual contents can help us to estimate whether the social link associates with topics. Therefore, for the following relationship $l_{u,v}$, we bring $\theta_u$ and $\theta_v$ to be the preconditions of $\mu_{u,v}$. More specifically, for drawing the value of $\mu_{u,v}$, we calculate the inner product of $\theta_u$ and $\theta_v$ and use it as the parameter of Bernoulli prior.

## 3.2 Model Inference by Gibbs Sampling

Given $M$ users and the hyper-parameters $\alpha$, $\beta$, $\delta$, $\epsilon$ and $\tau$, the joint probability distribution for the observed variables of the model can be written as

$$P(W, G \mid \Theta) \propto \prod_{i=1}^{M} P(\theta_i \mid \alpha) \prod_{n=1}^{N_i} P(z_{i,n}|\theta_i) \prod_{k=1}^{K} P(w_{i,n}|z_{i,n}, \beta) \times \prod_{u=1}^{M} \prod_{v=1}^{F_u} (P(\mu_{u,v} = 1 \mid \rho_{u,v}) P(x_{u,v}|\theta_u, \theta_v) \prod_{k=1}^{K} P(e_{u,v}|x_{u,v}, \delta) + P(\mu_{u,v} = 0 \mid \rho_{u,v})\, P(e_{u,v} \mid \epsilon)) \tag{1}$$

where $\Theta = \{\theta, \phi, \psi, \mu, \pi, \alpha, \beta, \delta, \epsilon, \tau\}$ is the set of parameters. $F_u$ is a set which contains all users followed by $u$-th user, and $G = \{l_{u,v} | u \in M, v \in F_u\}$ is the social graph. $W$ represents the observed word set.

To deal with the coupling of variables in our model, we use collapsed Gibbs sampling to learn variables distributions in Eq.1. Since, in SRTM, the distribution of following relationships is a joint distribution of two-level mixtures and simultaneously associates with two topic mixtures, we need to take both $\theta_u$ and $\theta_v$ into account when computing the posterior distributions of $x$, which is the topic distribution on following relationships. More specifically, the posterior distributions for Gibbs sampling in SRTM are given

$$p(z_{u,w} = k \mid z_{-(u,w)}, \alpha, \beta) \propto (d_{uk}^{-(u,w)} + s_{uk} + \alpha) \times \frac{(W_{kw}^{-(u,w)} + \beta)}{\sum_{w'}^{N} W_{kw'}^{-(u,w)} + \beta} \tag{2}$$

$$p(\mu_{u,v} = 1 \mid x_{-(u,v)}, \alpha) \propto sigmoid\Big(\sum_{k=1}^{K} \frac{d_{uk} + s_{uk}^{-(u,v)} + \alpha}{N_u + L_u + K\alpha} \times \frac{d_{vk} + s_{vk} + \alpha}{N_v + L_v + K\alpha} + \tau\Big) \tag{3}$$

$$p(x_{u,v}, \mu_{u,v} = 0 \mid x_{-(u,v)}, \alpha, \epsilon) \propto (1 - p(\mu_{u,v} = 1 \mid \mu_{-(u,v)}, \alpha)) \times \frac{\mu_{*,v,0} - 1 + \epsilon}{\mu_{*,*,0} - 1 + M\epsilon} \tag{4}$$

$$p(x_{u,v} = k, \mu_{u,v} = 1 \mid x_{-(u,v)}, \alpha, \delta) \propto p(\mu_{u,v} = 1 \mid x_{-(u,v)}, \alpha) \times (d_{uk} + s_{uk}^{-(u,v)} + \alpha) \times (d_{vk} + s_{vk} + \alpha) \times \frac{V_{ku}^{-(u,v)} + \delta}{\sum_{v'}^{M} V_{kv'}^{-(u,v)} + \delta} \tag{5}$$

where $z_{uw}$ denotes the topic of the $w$-$th$ word for the $u$-$th$ user, and $x_{u,v}$ is the topic of the $v$-$th$ link for the $u$-$th$ user. Let $z_{-(uw)}$ denote the topics for all words except $z_{uw}$, and $x_{-(u,v)}$ follow an analogous definition. We use $\mu_{u,v}$ as a factor indicator (topical or non topical) of the $v$-$th$ link for the $u$-$th$ user. Moreover, for recording the intermediate process, we bring several counters into the above equations, i.e., $W_{kw}$, $V_{kv}$, $d_{uk}$ and $s_{uk}$. $W_{kw}$ is the number of times that the $w$-$th$ word is assigned to the $k$-$th$ topic and $V_{kv}$ is the number of times that the $v$-$th$ user is assigned to the $k$-$th$ topic. $d_{uk}$ records the number of times that the $u$-$th$ user is assigned to the $k$-$th$ topic from texts. $s_{uk}$ is the number of times that the $u$-$th$ user is assigned to the $k$-$th$ topic from following links.

After the sampling algorithm runs for an appropriate number of iterations (until the chain has converged to a stationary distribution), the estimates for the parameters, i.e., $\theta$, $\phi$, $\psi$ and $\pi$ can be obtained via the following equations:

$$\theta_{uk} \propto \frac{d_{uk} + s_{uk} + \alpha}{N_u + L_u + K\alpha} \quad \pi_v \propto \frac{\mu_{*,v,0} + \epsilon}{\mu_{*,*,0} + M\epsilon} \tag{6}$$

$$\phi_{kw} \propto \frac{W_{kw} + \beta}{\sum_{w'}^{N} W_{kw'} + \beta} \quad \psi_{kv} \propto \frac{V_{kv} + \delta}{\sum_{v'}^{M} V_{kv'} + \delta} \tag{7}$$

where $L_u$ denotes the number of following links of the $u$-$th$ user. $\mu_{*,v,0}$ denotes the number of times that the $v$-$th$ user is followed by other users because of non-topical factors, and $\mu_{*,*,0}$ is the total number of times that the following behaviors are caused by non-topical factors in the dataset.

## 3.3 Social Link Generation

After the procedure of parameter inference, using the estimated parameters, we can construct a function to describe the generative process of social links in SRTM. In this function, we take the individual topic distribution, individual topic-specific influence and user popularity into account. More specifically, given a user $u$ and a candidate $v$, we can calculate the value which indicates the likelihood of $u$ following $v$ as below,

$$P(u, v) \propto (1 - \sigma(\rho_{u,v}))\pi_v + \sigma(\rho_{u,v}) \sum_{k=1}^{K} \theta_{uk}\, \theta_{vk}\, \psi_{k,v}. \tag{8}$$

Notice that including proposed SRTM model, the latent space models, e.g., LDA, Link-LDA and FLDA, can easily be embedded in this function. The value computed from this function describes the generation likelihood of the social link from user $u$ to user $v$.

# 4. EXPERIMENT

To investigate the performance of SRTM, we perform experiments on two real word datasets, i.e., Sina Weibo and Twitter, which contain a mass of following links and user published short texts. Sina Weibo is a popular Micro-bloging service of China, and the dataset we used contains 1,704,142 users and their related information, i.e., published statuses and social links. The dataset of Twitter is collected from twitter.com, which contains 5,847,699 users, 24,257,080 statuses and 126,964,950 social links. We randomly split training (80%) and testing (20%) data for experiments.

Table 1: Performance comparison on Sina Weibo (K=32, 64) with MAP@3, 10, 20 and AUC.

| Method | Sina Weibo (K=32) | | | | Sina Weibo (K=64) | | | |
|---|---|---|---|---|---|---|---|---|
| | MAP@3 | MAP@10 | MAP@20 | AUC | MAP@3 | MAP@10 | MAP@20 | AUC |
| MF | 0.435 | 0.453 | 0.436 | 0.849 | 0.459 | 0.468 | 0.448 | 0.853 |
| LDA | 0.064 | 0.085 | 0.091 | 0.613 | 0.079 | 0.102 | 0.106 | 0.622 |
| Link-LDA | 0.523 | 0.498 | 0.460 | 0.867 | 0.495 | 0.477 | 0.441 | 0.866 |
| FLDA | 0.533 | 0.506 | 0.469 | 0.870 | 0.496 | 0.475 | 0.440 | 0.868 |
| **SRTM** | **0.577** | **0.547** | **0.510** | **0.873** | **0.579** | **0.548** | **0.511** | **0.872** |

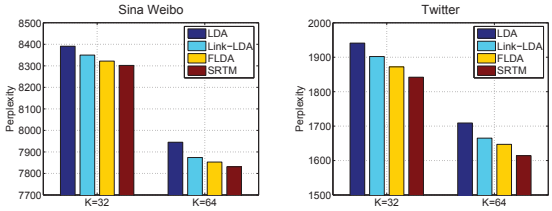


**Figure 3: The perplexity of LDA styled models on Sina Weibo and Twitter.**

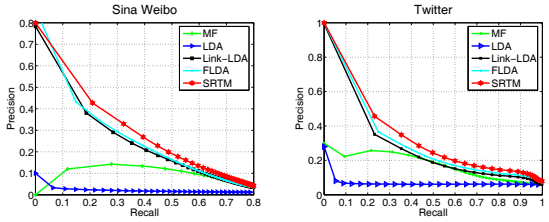


**Figure 4: Precision-Recall curves on Sina Weibo (K=32) and Twitter (K=32).**

We conduct comparisons with several state-of-the-art methods, including MF [10], LDA [6], Link-LDA [1] and FLDA [5]. Among these compared methods, MF and LDA only take links or texts into account, while Link-LDA and FLDA take both links and texts into account. In order to evaluate the LDA styled models, we use the perplexity which is widely used for topic modeling. To further study our model, we embed it into the TRA framework for social link prediction, and use Area Under the ROC Curve (AUC) and Mean Average Precision (MAP) as metrics.

Figure 3 demonstrates the effectiveness of compared methods on topic modeling, in term of perplexity. Since the training texts usually can not represent all topical information and some topics can be revealed by social links, LDA only taking texts into account can not well model such hybrid data. Since Link-LDA treats links as a kind of words and FLDA introduces a prior for the bandwagon effect of social links, they improve performance of LDA. Owing much to estimating topic-irrelevant links with relational topic distributions, our model yields lower perplexity than both Link-LDA and FLDA on two datasets.

In addition, we show keywords and associated influencers of several topics, which are explored by SRTM from Sina Weibo dataset. Table 2 indicates that our model can not only explain latent properties of users using the keywords extracted from texts, but also find out the users who are the influencers on individual topic. For example, under the topic of Economy, SRTM can not only provide keywords, e.g., investment, market and bank, to describe this topic, but also can discover the influencers, e.g., Xianping Lang who is a famous economist in China.

To evaluate performance of link generation, Table 1 shows the comparison evaluated by ranking metrics on Sina Weibo with two different factor dimensionalities. Owing much to take topic-irrelevant links, popularity of users and topic-specific influence of users into account, our model consistently outperforms the compared methods in terms of MAP@N and AUC. In particular, on MAP@N, improvements of SRTM over compared methods are real significant. This observation indicates that SRTM is very suitable for top-N recommendation, which is the most fundamental problem in practical applications. Figure 4 indicates the precision-recall curves on two datasets. Since short texts on social network contain much noise and lack formal linguistic structures, the performance of LDA is often worse than MF which only takes social links into account. Furthermore, we notice that Link-LDA and FLDA, jointly modeling the texts and links, have great superiority over LDA and MF only considering texts or links. This phenomenon reveals that on social network, the texts and social links can be used to complement each other in topic inference. In all experiments, due to well estimating topic-irrelevant links, SRTM consistently outperforms Link-LDA and FLDA.

**Table 2: A sample of topics and their influencers discovered by SRTM from Sina Weibo (K=32). Each topic is shown with top-8 keywords and top-3 influencers.**

| "Technique" | "Economy" | "Entertainment" | "Sport" |
|---|---|---|---|
| data | investment | movie | game |
| user | market | director | sport |
| Internet | company | story | fans |
| baidu | bank | release | football |
| problem | dollar | American | basketball |
| analysis | estate | comedy | champion |
| retrieval | policy | life | Milan |
| recommendation | funds | video | player |
| Yaqin Zhang | Xianping Lang | Kangyong Cai | Jianxiang Huang |
| Xiaochuan Wang | Shusong Ba | Zhang Wen | Lin Gao. |
| Hongjiang Zhang | Hongbin Song | Bo Huang. | Jianlian Yi |

## 5. ACKNOWLEDGMENTS

This work is jointly supported by National Basic Research Program of China (2012CB316300), and National Natural Science Foundation of China (61403390, U1435221, 61175003, 61420106015).

## 6. REFERENCES

[1] Erosheva et al. Mixed-membership models of scientific publications. *PNAS*, 101(suppl 1):5220–5227, 2004.
[2] Chang et al. Relational topic models for document networks. In *ICAIS*, pages 81–88, 2009.
[3] Dongsheng et al. Ranktopic: Ranking based topic modeling. In *ICDM*, pages 211–220, 2012.
[4] Sang-Min et al. A recommendation model using the bandwagon effect for e-marketing purposes in iot. *IJDSN*, 501:475163, 2015.
[5] Bin et al. Scalable topic-specific influence analysis on microblogs. In *WSDM*, pages 513–522, 2014.
[6] David M. Blei et al. Latent dirichlet allocation. *JMLR*, 3(1):993–1022, 2003.
[7] Rui Yan et al. Tweet recommendation with graph co-ranking. In *ACL*, pages 516–525, 2010.
[8] Page et al. The pagerank citation ranking: Bringing order to the web. 1999.
[9] Haveliwala et al. Topic-sensitive pagerank. In *WWW*, pages 517–526, 2002.
[10] Wei-Sheng et al. A learning-rate schedule for stochastic gradient methods to matrix factorization. In *PAKDD*, pages 442–455. 2015.