11-2014

On Joint Modeling of Topical Communities and Personal Interest in Microblogs

Tuan-Anh Hoang
Singapore Management University, tahoang.2011@smu.edu.sg

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

On Joint Modeling of Topical Communities and Personal Interest in Microblogs

Tuan-Anh Hoang and Ee-Peng Lim
Living Analytics Research Centre
Singapore Management University
tahoang.2011, eplim@smu.edu.sg

Abstract. In this paper, we propose the *Topical Communities and Personal Interest* (**TCPI**) model for simultaneously modeling topics, topical communities, and users' topical interests in microblogging data. **TCPI** considers different topical communities while differentiating users' personal topical interests from those of topical communities, and learning the dependence of each user on the affiliated communities to generate content. This makes **TCPI** different from existing models that either do not consider the existence of multiple topical communities, or do not differentiate between personal and community's topical interests. Our experiments on two Twitter datasets show that **TCPI** can effectively mine the representative topics for each topical community. We also demonstrate that **TCPI** significantly outperforms other state-of-the-art topic models in the modeling tweet generation task.

Keywords: Social media, Microblogs, Topic modeling, User modeling

1 Introduction

Microblogging sites such as Twitter[1] and Weibo[2] allow users to publish short messages, which are called *tweets*, sharing their current status, opinion, and other information. Embedded in these tweets is a wide range of topics. Empirical and user studies on microblog usage have showed that users may tweet about either their personal topics or background topics [11, 29, 13]. The former covers individual interests of the users. The latter is the interest shared by users in topical communities and they emerge when users in the same community tweet about common interests [6]. Background topics are thus the results of interests of the topical communities.

There are previous works on modeling background topics in social media as well as in general document corpuses, e.g., [30, 23, 10]. However, most of these works model a single background topic or a distribution of background topics. In this work, we instead consider the existence of multiple topical communities, each with a different background topic distribution. Examples of such communities include IT professionals, political groups, entertainment fans, etc.. The IT

[1] www.twitter.com
[2] http://www.weibo.com

community covers topics such as technology, science, etc.. The political community covers topics such as welfare, budget, etc.. A user who is associated with a topical community will therefore adopt topics from the interest of the community. The members of these communities may not be socially connected to one another. Hence, when modeling users on social media, we have to consider both the user's personal interests and his topical communities.

In this work, we aim to model topical communities as well as users' topical interests in microblogging data. We want to consider different topical communities, and also to learn topical interests of each user and her dependence on the topical communities to generate content.

A simple way to identify the topical communities is first performing topic modeling on the set of tweets using one of existing models (e.g., LDA [4]) to find out topical interests of the users, then assign the most common topics of all the users to be the topical communities' topics. Such an approach however does not allow us to distinguish between multiple topical communities, nor allow each topical community to have multiple topics. It also does not allow us to quantify, for each user, the degree in which the user depends on topical communities in generating content. We therefore propose to jointly model user topical interests and topical communities' interests in a same framework where each user has a parameter controling her bias towards generating content based on her own interests or based on the topical communities.

Our main contributions in this work consist of the following.

- We propose a probabilistic graphical model, called *Topical Communities and Personal Interest* model (abbreviated as **TCPI**), for modeling topics and topical communities, as well as modeling users' topical interests and their dependency on the topical communities in generating content.
- We develop a sampling method to infer the model's parameters. We further develop a regularization technique to bias the model to learn more semantically clear topical communities.
- We apply **TCPI** model on two Twitter datasets and show that it significantly outperforms other state-of-the-art models in modeling tweet generation task.
- An empirical analysis of topics and topical communities for the two datasets has been conducted to demonstrate the efficacy of the **TCPI** model.

The rest of the paper is organized as follows. We first discuss the related works on modeling topics in social media in Section 2. We then present our proposed model in detail in Section 3. Next, we describe two experimental datasets and report results of experiments in applying the proposed model on the two dataset in Section 4. Finally, we give our conclusions and discuss future work in Section 5.

2 Related Work

In this section, we review previous works that are closely related to our work. These works fall into two categories: (i) the works on analyzing topics in microblogs, and (ii) works on analyzing communities in social networks.

2.1 Topic Analysis

Michelson *et. al.* first examined topical interests of Twitter users by analyzing the named entities mentioned in their tweets [16]. Hong *et. al.* then conducted an empirical study on different ways of performing topic modeling on tweets using the original LDA model [9] and Author-topic model [21]. They found that topic learnt from documents formed by aggregating tweets posted by the same users may help to significantly improve some user profiling tasks. Similarly, Mehrotra *et. al.* investigated different ways of forming documents from tweets in order to improve the performance of LDA model for microblogging data [15]. They found that grouping the tweets containing the same hashtags may lead to a significant improvement. Using the same approach, Ramage *et. al* proposed to use Supervised LDA model [20] to model topics of tweets where each tweet is labeled based on linguistic elements (e.g., hashtags, emoticons, and question marks, etc.) contained in the tweet; and Qiu *et. al.* proposed to jointly modeling topics of tweets and their associated posting behaviors (i.e., tweet, retweet, or reply) [19]. Lastly, the work by Zhao *et. al.* [30] is particularly close to our work. In this work, the authors proposed **TwitterLDA** topic model, which is considered as state-of-the-art topic model for microblogging data. **TwitterLDA** is a variant of LDA, in which: (i) documents are formed by aggregating tweets posted by the same users; (ii) a single background topic is assumed; (iii) there is only one common topic for all words in each tweet; and (iv), each word in a tweet is generated from either the background topic or the user's topic. The plate notation of **TwitterLDA** model is shown in Figure 1 (a), and it's generative process is as follows.

- Sample the background topic $\phi_B \sim Dirichlet(\beta)$
- For each $k = 1, \cdots, K$, sample the k-th topic $\phi_k \sim Dirichlet(\beta)$
- Sample the dependence on background topic $\mu \sim Beta(\rho)$
- For each user u, sample u's topic distribution $\theta_u \sim Dirichlet(\alpha)$
- Generate tweets for the user u: for each tweet t that u posts:
  1. Sample topic for the tweet $z_t \sim Multinomial(\theta_u)$
  2. Sample the tweet's words: for each word $w_{t,n}$ at slot n:
     - Sample $y_{t,n} \sim Bernoulli(\mu)$
     - If $y_{t,n} = 0$, sample from background topic: $w_{t,n} \sim Multinomial(\phi_B)$; else ($y_{t,n} = 1$), sample from topic z_t: $w_{t,n} \sim Multinomial(\phi_{z_t})$

TwitterLDA model however does not consider multiple background topics, and impractically assume that all users have the same dependency on the unique background topic (as the paramter μ is common for all the users).

It is important to note that our work is similar but not exactly the same with works on finding global topics (e.g., [10, 23]). Global topics are shared by all the users and not specific for any community. On the other hand, topics of each topical community is specific for the community, and are shared mostly by users within the community.

2.2 Community Analysis

Most of the early works on community analysis in social networks are finding social communities based on social links among the users. For example, Newman proposed to discover social communities by finding a network partition that maximizes a measure of "compactness" in community structure called *modularity* [18]; Airoldi *et. al.* proposed a statistical mixed membership model [1]. There are also works on finding topical communities based on user generated content (e.g., [31, 23]), and users' attributes and interest affiliations (e.g., [24, 26, 27]). Ding *et. al.* conducted an empirical study showing that social community structure of a social network may significantly be different from topical communities discovered from the same network [5]. Moreover, most of existing works on analyzing topical community do not differentiate users' personal interests from those of topical communities. They assume that a user's topical interests is determined purely based on her topical communities' interests. This assumption is not practical when applying for microblogging users since they express interest in a vast variety of topics of daily life, and their interests are therefore not always determined by their topical communities.

Lastly, it is also important to note that our work is different from works on finding topical interests of social communities (e.g., [28, 22]). Topical interests of each social community includes most common topics shared by users within the community, and hence may not specific for the community, i.e., two different social communities may have the same topical interests. On the other hand, each topical community is uniquely determined based on its topical interests: different topical communities have significantly different topical interests.

3 Topical Community and Personal Interest Model

3.1 Assumptions

Our model relies on the assumptions that: (i) users generate content topically; and (ii) users generate content according either to their personal interests or some topical communities. The first assumption suggests that, for each user, there is always an underlying topic explaining content of the every tweet she posts. The second assumption suggests that, while different users generally have different personal topical interests, their generated content also share some common topics of the topical communities the users belong to. For example, most of the users tend to tweet about daily activities and entertainment although these topics may not represent their real personal interest. During an election campaign, a user who is not personally interested in politics, may still tweet more about political topics as she follows the prevalent topical community of interests in political topics. Hence, to model users' content accurately, it is important to determine topical communities as well as their own personal interests.

3.2 Generative Process

Based on the assumptions as presented above, we propose the **TCPI** model to model user generated tweet from a vocabulary $\mathcal{V}$. The **TCPI** model has K

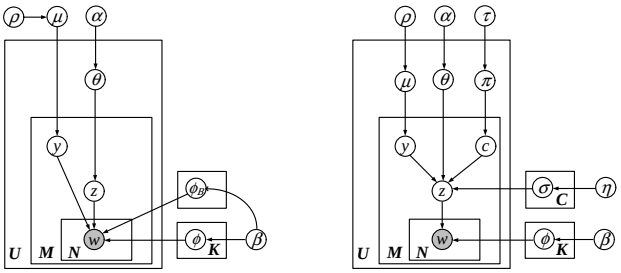

Fig. 1: Plate notation for: (a) **TwitterLDA** and (b) **TCPI** models

latent topics, where each topic k has a multinomial distribution ϕ_k over the vocabulary $\mathcal{V}$. To capture the topical communities, the **TCPI** model assumes that there are C topical communities, where each community c has a multinomial distribution σ_c over the K topics. Each user u also has a personal topic distribution θ_u over the K topics and a community distribution π_u over the C topical communities. Moreover, each user has a dependence distribution μ_u which is a Bernoulli distribution indicating how likely the user tweets based on her own personal interests (μ_u^0) or based on the topical communities ($\mu_u^1 = 1 - \mu_u^0$). Lastly, we assume that θ_u, π_u, σ, and ϕ have Dirichlet priors α, τ, η, and β respectively, while μ_u has Beta prior ρ.

In **TCPI** model, we assume the following generative process for all the posted tweets. To generate a tweet t for user u, we first flip a biased coin $y_{u,t}$ (whose bias to head up is μ_u^0) to decide if the tweet is based on u's personal interests, or based on one of the topical communities u belongs to. If $y_{u,t} = 0$, we then choose the topic z_t for the tweet according to u's topic distribution θ_u. Otherwise, $y_{u,t} = 1$, we first choose a topical community c according to u's community distribution π_u, then we choose z_t according to the chosen community's topic distribution σ_c. As tweets are short with no more than 140 characters, we assume that each tweet has only one topic. Once the topic z_t is chosen, words in t are then chosen according to the topic's word distribution ϕ_{z_t}. In summary, the **TCPI** model has the plate notation as shown in Figure 1 (b) and the generative process as follows.

- For each $k = 1, \cdots, K$, sample the k-th topic $\phi_k \sim Dirichlet(\beta)$
- For each $c = 1, \cdots, C$, sample the c-th community's topic distribution $\sigma_c \sim Dirichlet(\eta)$
- For each user u
  1. Sample u's topic distribution $\theta_u \sim Dirichlet(\alpha)$
  2. Sample u's community distribution $\pi_u \sim Dirichlet(\tau)$
  3. Sample u's dependence distribution $\mu_u \sim Beta(\rho)$
- Generate tweets for the user u: for each tweet t that u posts:
  1. Sample $y_{u,t} \sim Bernoulli(\mu_u)$
  2. Sample topic for the tweet: if $y_{u,t} = 0$, sample $z \sim Multinomial(\theta_u)$; if $y_{u,t} = 1$, sample a community $c \sim Multinomial(\pi_u)$, then sample $z_t \sim Multinomial(\sigma_c)$

Fig. 2: Probabilities used in **jointly sampling coin and topical community** for tweet t_j^i without regularization

$$p(y_j^i = 0|\mathcal{T}, \mathcal{Y}_{-t_j^i}, \mathcal{C}_{-t_j^i}, \mathcal{Z}, \alpha, \beta, \tau, \eta, \rho) \propto$$

$$\propto \frac{\mathbf{n_y}(0, u_i, \mathcal{Y}_{-t_j^i}) + \rho_0}{\sum_{y=0}^{1}\left(\mathbf{n_y}(y, u_i, \mathcal{Y}_{-t_j^i}) + \rho_y\right)} \cdot \frac{\mathbf{n_{zu}}(z_j^i, u_i, \mathcal{Z}_{-t_j^i}) + \alpha_{z_j^i}}{\sum_{k=1}^{K}\left(\mathbf{n_{zu}}(k, u_i, \mathcal{Z}_{-t_j^i}) + \alpha_k\right)} \tag{1}$$

$$p(y_j^i = 1, c_j^i = c|\mathcal{T}, \mathcal{Y}_{-t_j^i}, \mathcal{C}_{-t_j^i}, \mathcal{Z}, \alpha, \beta, \tau, \eta, \rho) \propto$$

$$\propto \frac{\mathbf{n_y}(1, u_i, \mathcal{Y}_{-t_j^i}) + \rho_1}{\sum_{y=0}^{1}\left(\mathbf{n_y}(y, u_i, \mathcal{Y}_{-t_j^i}) + \rho_y\right)} \cdot \frac{\mathbf{n_{cu}}(c, u_i, \mathcal{C}_{-t_j^i}) + \tau_c}{\sum_{c=1}^{C}\left(\mathbf{n_{cu}}(c, u_i, \mathcal{C}_{-t_j^i}) + \tau_c\right)} \cdot \frac{\mathbf{n_{zc}}(z_j^i, c, \mathcal{Z}_{-t_j^i}, \mathcal{C}_{-t_j^i}) + \eta_{cz_j^i}}{\sum_{k=1}^{K}\left(\mathbf{n_{zc}}(k, c, \mathcal{Z}_{-t_j^i}, \mathcal{R}_{-t_j^i}) + \eta_{ck}\right)} \tag{2}$$

3. Sample the tweet's words: for each word slot n, sample the word $w_{t,n} \sim Multinomial(\phi_{z_t})$

3.3 Model Learning

Consider a set of microblogging users together with their posted tweets, we now present the algorithm for performing inference in the **TCPI** model. We use U to denote the number of users and use W to denote the number of words in the tweet vocabulary $\mathcal{V}$. We denote the set of all posted tweets in the dataset by $\mathcal{T}$. For each user u_i, we denote her j-th tweet by t_j^i. For each posted tweet t_j^i, we denote N_{ij} words in the tweet by $w_1^{ij}, \cdots, w_{N_{ij}}^{ij}$ respectively, and we denote the tweet's topic, coin, and topical community (if exists) by z_j^i, y_j^i, and c_j^i respectively. Lastly, we denote the bag-of-topics, bag-of-coins, and bag-of-topical communities of all the posted tweets in the dataset by $\mathcal{Z}$, $\mathcal{Y}$, and $\mathcal{C}$ respectively.

Due to the intractability of LDA-based models [4], we make use of sampling method in learning and estimating the parameters in the **TCPI** model. More exactly, we use a collapsed Gibbs sampler ([14]) to iteratively and jointly sample the latent coin and latent topical community, and sample latent topic of every posted tweet as follows.

For each posted tweet t_j^i, the j-th tweet posted by user u_i, we use $\mathcal{Y}_{-t_j^i}$, $\mathcal{C}_{-t_j^i}$, $\mathcal{Z}_{-t_j^i}$ to denote the bag-of-coins, bag-of-topical communities and bag-of-topics, respectively, of all other posted tweets in the dataset except the tweet t_i^j. Then the coin y_j^i and the topical community c_j^i of t_j^i are jointly sampled according to equations in Figure 2, and the topic z_j^i of t_j^i is sampled according to equations in Figure 3. Note that when $y_j^i = 0$, we do not have to sample c_j^i, and the current c_j^i (if exists) will be discarded. In these equations, $\mathbf{n_y}(c, u, \mathcal{C})$ records the number of times the coin y is observed in the set of tweets of user u for the bag-of-coins $\mathcal{Y}$. Similarly, $\mathbf{n_{zu}}(z, u, \mathcal{Z})$ records the number of times the topic z is observed in the

Fig. 3: Probabilities used in **sampling topic** for tweet t_j^i without regularization

$$p(z_j^i = z | y_j^i = 0, \mathcal{T}, \mathcal{Y}_{-t_j^i}, \mathcal{C}, \mathcal{Z}_{-t_j^i}, \alpha, \beta, \tau, \eta, \rho) \propto$$

$$\propto \frac{\mathbf{n_{zu}}(z, u_i, \mathcal{Z}_{-t_j^i}) + \alpha_z}{\sum_{k=1}^{K} \left(\mathbf{n_{zu}}(k, u_i, \mathcal{Z}_{-t_j^i}) + \alpha_k\right)} \cdot \prod_{n=1}^{N_{ij}} \frac{\mathbf{n_w}(w_n^{ij}, z, \mathcal{Z}_{-t_j^i}) + \beta_{z w_n^{ij}}}{\sum_{v=1}^{W} \left(\mathbf{n_w}(v, z, \mathcal{Z}_{-t_j^i}) + \beta_{zv}\right)} \quad (3)$$

$$p(z_j^i = z | y_j^i = 1, \mathcal{T}, \mathcal{Y}_{-t_j^i}, \mathcal{C}, \mathcal{Z}_{-t_j^i}, \alpha, \beta, \tau, \eta, \rho) \propto$$

$$\propto \frac{\mathbf{n_{zc}}(z, c_j^i, \mathcal{Z}_{-t_j^i}, \mathcal{C}_{-t_j^i}) + \eta_{c_j^i z}}{\sum_{k=1}^{K} \left(\mathbf{n_{zc}}(k, c_j^i, \mathcal{Z}_{-t_j^i}, \mathcal{C}_{-t_j^i}) + \alpha_{c_j^i k}\right)} \cdot \prod_{n=1}^{N_{ij}} \frac{\mathbf{n_w}(w_n^{ij}, z, \mathcal{T}_{-t_j^i}, \mathcal{Z}_{-t_j^i}) + \beta_{z w_n^{ij}}}{\sum_{v=1}^{W} \left(\mathbf{n_w}(v, z, \mathcal{T}_{-t_j^i}, \mathcal{Z}_{-t_j^i}) + \beta_{zv}\right)} \quad (4)$$

set of tweets of user u for the bag of topics $\mathcal{Z}$; $\mathbf{n_{zc}}(z, c, \mathcal{Z}, \mathcal{C})$ records the number of times the topic z is observed in the set of tweets that are tweeted based on the topical community c by any user for the bag-of-topics $\mathcal{Z}$ and the bag-of-topical communities $\mathcal{C}$; $\mathbf{n_{cu}}(c, u, \mathcal{C})$ records the number of times the topical community c is observed in the set of tweets of user u; and $\mathbf{n_w}(w, z, \mathcal{T}, \mathcal{Z})$ records the number of times the word w is observed in the topic z for the set of tweets $\mathcal{T}$ and the bag-of-topics $\mathcal{Z}$.

In the right hand side of Equation 1: (i) the first term is proportional to the probability that the coin 0 is generated given the priors and (current) values of all other latent variables (i.e., the coins, topical communities (if exist), and topics of all other tweets); and (ii) the second term is proportional to the probability that the (current) topic z_j^i is generated given the priors, (current) values of all other latent variables, and the chosen coin. Similarly, in the right hand side of Equation 2: (i) the first term is proportional to the probability that the coin 1 is generated given the priors and (current) values of all other latent variables; (ii) the second term is proportional to the probability that the topical community c is generated given the priors, (current) values of all other latent variables, and the chosen coin; and (iii) the third term is proportional to the probability that the (current) topic z_j^i is generated given the priors, (current) values of all other latent variables, and the chosen coin as well as the chosen community.

The terms in the right hand side of Equations 3 and 4 respectively have the similar meaning with those of Equations 1 and 2.

3.4 Sparsity Regularization

As we want to differentiate users' tweets based on personal interests from topical communities and to differentiate one topical community from the others, we would prefer a clear distinction among these latent factors. In other words, we want topical communities' topic distributions and users' topic distributions to be skewed on different topics, and topical communities' topic distribution to be also skewed on different topics. More exactly, in estimating parameters in the **TCPI** model, we need to obtain sparsity in the following distribution.

- Topic specific coin distribution $p(y|z)$ where y is a coin and z is a topic: the sparsity in this distribution is to ensure that each topic z is mostly covered by either users' personal interests or topical communities.
- Topic specific topical community distribution $p(c|z)$ where c is a topical community and z is a topic: the sparsity in this distribution is to ensure that each topic z is mostly covered by one or only a few topical communities.

To obtain the sparsity mentioned above, we use the *pseudo-observed variable* based regularization technique proposed by Balasubramanyan *et. al.* [2] as follows.

Topic specific coin distribution regularization. Since the topic specific coin distributions are determined by both coin and community joint sampling and topic sampling steps, we regularize both these two steps to bias the distributions to expected sparsity.

In coin and topical community joint sampling steps. In each coin and topical community sampling step for the tweet t^i_j, we multiply the right hand side of equations in Figure 2 with a corresponding regularization term $\mathcal{R}_{\text{topCoin-C\&C}}(y|z^i_j)$ which is computed based on empirical entropy of $p(y|z^i_j)$ as in Equation 5.

Fig. 4: **Topic specific coin distribution regularization terms** used in sampling **coin** and/or **topical community** for tweet t^i_j

$$\mathcal{R}_{\text{topCoin-C\&C}}(y|z^i_j) = exp\Bigg(- \frac{\Big(H_{y^i_j=y}\big(p(y'|z^i_j)\big) - \mu_{\text{topCoin}}\Big)^2}{2\sigma^2_{\text{topCoin}}}\Bigg) \tag{5}$$

In topic sampling steps. In each topic sampling step for the tweet t^i_j, we multiply the right hand side of equations in Figure 3 with a corresponding regularization term $\mathcal{R}_{\text{topCoin-Topic}}(z|t^i_j)$ which is computed based on empirical entropy of $p(y|z)$ as in Equation 6.

Fig. 5: **Topic specific coin distribution regularization terms** used in sampling **topic** for tweet t^i_j

$$\mathcal{R}_{\text{topCoin-Topic}}(z|t^i_j) = exp\Bigg(- \sum_{z'=1}^{K} \Bigg[\frac{\Big(H_{z^i_j=z}\big(p(y|z')\big) - \mu_{\text{topCoin}}\Big)^2}{2\sigma^2_{\text{topCoin}}}\Bigg]\Bigg) \tag{6}$$

In Equations 5, $H_{y^i_j=y}\big(p(y'|z^i_j)\big)$ is the empirical entropy of $p(y'|z^i_j)$ when $y^i_j = y$. Similarly, in Equations 6, for each topic z', $H_{z^i_j=z}\big(p(y|z')\big)$ is the empirical entropy of $p(y|z')$ when $z^i_j = z$. The two parameters μ_{topCoin} and σ_{topCoin} is

Fig. 6: **Topic specific topical community distribution regularization terms** used in sampling **coin** and/or **topical community** for tweet t_j^i

$$\mathcal{R}_{\text{topComm-C\&C}}(y, c|z_j^i) = exp\Big(- \frac{\Big(H_{y_j^i=y, c_j^i=c}\big(p(c'|z_j^i)\big) - \mu_{\text{topComm}}\Big)^2}{2\sigma^2_{\text{topComm}}}\Big) \quad (7)$$

Fig. 7: **Topic specific topical community distribution regularization terms** used in sampling **topic** for tweet t_j^i

$$\mathcal{R}_{\text{topComm-Topic}}(z|t_j^i) = exp\Big(- \sum_{z'=1}^{K} \Big[\frac{\Big(H_{z_j^i=z}\big(p(c|z')\big) - \mu_{\text{topComm}}\Big)^2}{2\sigma^2_{\text{topComm}}}\Big]\Big) \quad (8)$$

respectively the expected mean and expected variance of the entropy of $p(y|z)$. These expected mean and expected variances are pre-defined parameters. Obviously, with a low expected mean μ_{topCoin}, these regularization terms (1) increase weight for values of y, c, and z that give lower empirical entropy of $p(y|z)$, and hence increasing the sparsity of these distributions; but (2) decrease weight for values of y, c, and z that give higher empirical entropy of $p(y|z)$, and hence decreasing the sparsity of these distributions.

Topic specific topical community distribution regularization. Similarly, since the topic specific topical community distributions are determined by both coin and topical community joint sampling and topic sampling steps, we regularize both these two steps to bias the distributions to expected sparsity.

In coin and topical community joint sampling steps. In each coin and topical community sampling step for the tweet t_j^i, we also multiply the right hand side of equations in Figure 2 with a corresponding regularization term $\mathcal{R}_{\text{topComm-C\&C}}(y, c|z_j^i)$ which is computed based on empirical entropy of $p(c'|z_j^i)$ as in Equation 7.

In topic sampling steps. In each topic sampling step for the tweet t_j^i, we also multiply the right hand side of equations in Figure 3 with a corresponding regularization term $\mathcal{R}_{\text{topComm-Topic}}(z|t_j^i)$ which is computed based on empirical entropy of $p(c|z)$ as in Equation 8.

In Equations 7, $H_{y_j^i=y, c_j^i=c}\big(p(c'|z_j^i)\big)$ is the empirical entropy of $p(c'|z_j^i)$ when $y_j^i = y$ and $c_j^i = c$. Similarly, in Equations 8, for each topic z', $H_{z_j^i=z}\big(p(c|z')\big)$ is the empirical entropy of $p(c|z')$ when $z_j^i = z$. The two parameters μ_{topComm} and σ_{topComm} is respectively the expected mean and expected variance of the entropy of $p(c|z)$. These expected mean and expected variances are pre-defined parameters. Obviously, with a low expected mean μ_{topComm}, these regularization terms (1) increase weight for values of y, c, and z that give lower empirical entropy of $p(c|z)$, and hence increasing the sparsity of these distributions; but (2) decrease weight for values of y, c, and z that give higher empirical entropy of $p(c|z)$, and hence decreasing the sparsity of these distributions.

Table 1: Statistics of the experimental datasets

Dataset	**SE**	**Two-Week**
#user	14,595	24,046
#tweets	3,030,734	3,181,583

In our experiments, we used sampling method with the above regularization setting $\mu_{\text{topCoin}} = \mu_{\text{topComm}} = 0$, $\sigma_{\text{topCoin}} = 0.3$, $\sigma_{\text{topComm}} = 0.5$. We also used symmetric Dirichlet hyperparameters with $\alpha = 50/K$, $\beta = 0.01$, $\rho = 2$, $\tau = 1/C$, and $\eta = 50/K$. Given the input dataset, we train the model with 600 iterations of Gibbs sampling. We took 25 samples with a gap of 20 iterations in the last 500 iterations to estimate all the hidden variables.

4 Experimental Evaluation

4.1 Datasets

Using snowball sampling, we collected the following two datasets for evaluating the **TCPI** model.

SE Dataset. This dataset is collected from a set of Twitter users who are interested in technology, and particularly in software development. To construct this dataset, we first utilized 100 most influential software developers in Twitter provided in [12] as the seed users. These are highly-followed users who actively tweet about software engineering topics, e.g., *Jeff Atwood*[3], *Jason Fried*[4], and *John Resig*[5]. We further expanded the user set by adding all users following at least five seed users. Lastly, we took all tweets posted by these users from August 1st to October 31st, 2011 to form the first dataset, called **SE** dataset.

Two-Week Dataset. The second dataset is a large corpus of tweets collected just before the 2012 US presidential election. To construct this corpus, we first manually selected a set of 56 *seed users*. These are highly-followed and politics savy Twitter users, including major US politicians, e.g., Barack Obama, Mitt Romney, and Newt Gingrich; well known political bloggers, e.g., America Blog, Red State, and Daily Kos; and political desks of US news media, e.g., CNN Politics, and Huffington Post Politics. The set of users was then expanded by adding all users following at least three seed users. Lastly, we used all the tweets posted by these users during the two week duration from August 25th to September 7th, 2012 to form the second dataset, known as the **Two-Week** dataset.

We employed the following preprocessing steps to clean both datasets. We first removed stopwords from the tweets and filtered out tweets with less than 3 non stop-words. Next, we excluded users with less than 50 (remaining) tweets. This minimum thesholds are necessary so that, for each user, we have enough number of tweet observations for learning both influence of the user's personal interests and that of the topical communities in tweet generation.

Table 1 shows the statistics of the two datasets after the preprocessing steps. As shown in the table, the two datasets after filtering are still large, with about

[3] http://en.wikipedia.org/wiki/Jeff_Atwood
[4] http://www.hanselman.com/blog/AboutMe.aspx
[5] http://en.wikipedia.org/wiki/John_Resig

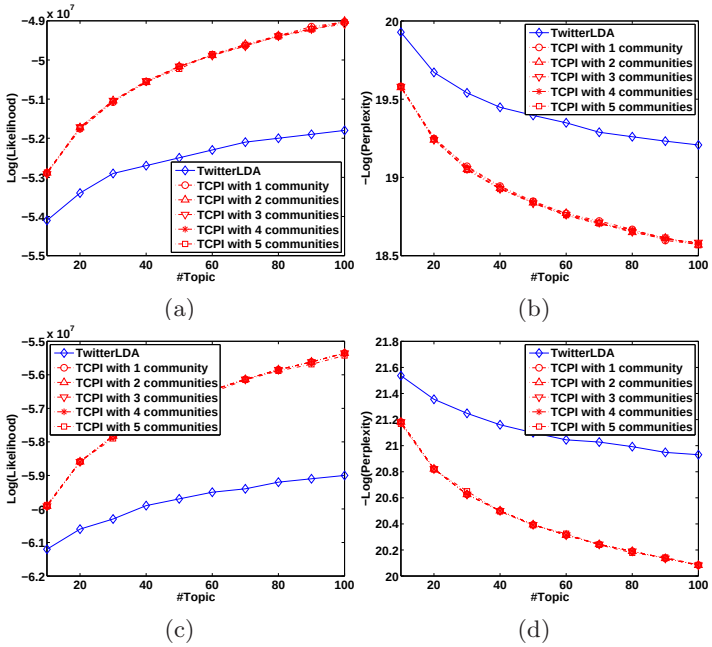

Fig. 8: Loglikelihood and Perplexity of **TwitterLDA** and **TCPI** in: ((a) and (b)) **SE**, and ((c) and (d)) **Two-Week** datasets

200 tweets per user in **SE** dataset; and 120 tweets per user in **Two-Week** dataset. This allows us to learn the latent factors accurately.

4.2 Evaluation Metrics

To examine the ability of **TCPI** model in modeling tweet generation, we compare **TCPI** with **TwitterLDA** model. We adopt *likelihood* and *perplexity* for evaluating the two models. For each user, we randomly selected 90% of tweets of the user to form a training set, and use the remaining 10% of the tweets as the test set. We then learn the **TCPI** and **TwitterLDA** models using the training set, and using the learnt models to generate the test set. Lastly, for each model, we compute the likelihood of the training set and perplexity of the test set. The model with a higher likelihood, or lower perplexity is considered better for the task.

4.3 Performance Comparison

Figures 8 (a) and (b) show the performance of **TwitterLDA** and **TCPI** models in topic modeling on **SE** dataset. Figures 8 (c) and (d) show the performance of the models on **Two-Week** dataset. As expected, larger number of topics K gives larger likelihood and smaller perplexity, and the amount of improvement diminishes as K increases. The figures show that: (1) **TCPI** significantly outperforms **TwitterLDA** in topic modeling task; and (2) **TCPI** is robust against the number of topical communities as its performance does not significantly change as we increase the number of the communities from 1 to 5.

4.4 Background Topics and Topical Communities Analysis

We now examine the background topics and topical communities found by the **TwitterLDA** and **TCPI** models respectively. Considering both time and space complexities, and since it is not practical to expect a large number of topics falling in topical communities, we set the number of the topical communities in **TCPI** model to 3, and set the number of topics in both models to 80.

Table 2: Top words of **background topic** found in **SE** dataset by **TwitterLDA** model

life,making,video,blog,change,reading,job,home,thought,line
team,power,game,business,money,friends,talking,starting,month,company

Table 3: Top topics of **topical communities** found in **SE** dataset by **TCPI** model

Community Id	Community Label	Top topics		
		Topic Id	Topic Label	Probability
0	Daily life	61	Daily stuffs	0.535
		79	Traveling	0.086
		25	Food and drinks	0.063
1	Apple's product	50	iOS	0.274
		74	Networking services	0.146
		37	iPhone and iPad	0.091
2	Software development	24	Programming	0.614
		9	Conference and meeting	0.105
		15	Operating systems	0.056

Table 2 shows the top words of the background topic found by **TwitterLDA** model in **SE** dataset, and Table 3 shows the top topics of each topical community found by **TCPI**. Note that, other than background topic, the labels of other topics are manually assigned after examining the topics' top words (shown in Tables 4) and top tweets. For each topic, the topic's top words are the words having the highest likelihoods given the topic, and the topic's top tweets are the tweets having the lowest perplexities given the topic. The label of each topical community is also manually assigned based on examining the community's top topics. The tables show that: (i) the background topic found by **TwitterLDA** model is not sematically clear; and (ii) the topical communitiess and their extreme topics found by **TCPI** model are both semantically clear and reasonable. In **SE** dataset, other than *Daily life* community as reported in [11], it is expected that professional communities *Software Development* and *Apple's product* exist in the dataset as most of its users are working in IT industry. This agrees with the findings by Zhao *et. al.* [29] that people also use Twitter for gathering and sharing useful information relevant to their profession.

Similarly, Table 5 shows the top words of the background topic found by **TwitterLDA** model in **Two-Week** dataset, and Table 6 shows the top topics of each topical community found by **TCPI** model. Again, the topics' labels are

Table 4: Top words of topics found in **SE** dataset by **TCPI** model

9	Conference and meeting	conference,meeting,team,weekend,code,session,home event,book,friends,friday,coffee,room,folks lunch,presentation,job,slides,minutes,beer
15	Operating sytems	windows,linux,mac,laptop,ubuntu,server,machine desktop,running,computer,systems,usb,ssd,lion software,#linux,apple,macbook,installing,win8,power
24	Program-ming	code,javascript,git,ruby,java,github,rails,data,api,server,tests php,node,python,language,blog,simple,programming,testing,files
25	Food and drinks	coffee,eating,chicken,dinner,cream,ice,lunch,beer cheese,bacon,chocolate,breakfast,recipe,delicious pizza,salad,wine,pumpkin,bread,butter
37	iPhone and iPad	iphone,apple,ipad,event,ipod,video,ios,retina,macbook,#apple,screen #iphone5,mac,battery,lightning,camera,connector,imac,nano,price
50	iOS	mac,ios,iphone,windows,chrome,apple,lion,ipad,google,screen,mountain android,text,safari,version,browser,itunes,desktop,keyboard,tweetbot
61	Daily stuffs	home,kids,house,#fb,life,coffee,dog,car,wife,room bed,thought,cat,playing,wearing,making,music,baby,friends,weekend
74	Networking services	email,facebook,google,spam,emails,page,blog,service,link,gmail password,mail,users,linkedin,api,inbox,client,links,message,user
79	Traveling	home,train,san,city,ride,bike,airport,weather,car,bus,rain weekend,francisco,traffic,london,road,minutes,heading,#fb,plane

Table 5: Top words of **background topic** found in **Two-Week** dataset by **TwitterLDA** model

life,making,home,america,called,house,change,thought,video,talking line,american,money,country,job,obama,friends,fact,lost,hell

Table 6: Top topics of **topical communities** found in **Two-Week** dataset by **TCPI** model

Community Id	Community Label	Top topics		
		Topic Id	Topic Label	Probability
0	Daily life	1	Daily stuffs	0.622
		32	Happenings in DNC and RNC 2012	0.062
		25	Food and drinks	0.052
1	Republicans' activities	10	Republican candidates	0.210
		32	Happenings in DNC and RNC 2012	0.196
		0	Presidential candidates' speeches	0.066
2	Campaigning speeches	0	Presidential candidates' speeches	0.203
		18	Speeches at DNC 2012	0.175
		16	Goverment and people	0.108

manually assigned after examining the topics' top words (shown in Tables 7) and top tweets; and the communities' labels are also manually assigned based on examining the communities' top topics. Also, the tables show that: (i) the

Table 7: Top words of topics found in **Two-Week** dataset by **TCPI** model

0	Presidential candidates speeches	obama,romney,gop,media,lies,speech,party,ryan fact,#dnc2012,#tcot,convention,dems,truth republicans,facts,mitt,democrats,campaign,liberal
1	Daily stuffs	life,home,kids,class,mom,house,car,bed,god,friends,room thought,baby,weekend,friend,person,family,hair,game,dog
10	Republican candidates	romney,#gop2012,mitt,#rnc2012,speech,#rnc,ann,ryan christie,america,paul,obama,president,chris rubio,#romneyryan2012,#tcot,american,convention,condi
16	Goverment and people	america,obama,government,god,party,country,american #tcot,freedom,rights,democrats,#dnc2012,americans,gop gop,constitution,liberty,nation,war,power,states
18	Speeches at DNC 2012	obama,#dnc2012,#tcot,biden,#dnc,joe,dnc clinton,#dncin4words,speech,#p2,america,president,bill bill,romney,god,michelle,barack,chair,dems
25	Food and drinks	coffee,chicken,ice,cream,eating,dinner,cheese,beer,lunch,bacon chocolate,breakfast,pizza,wine,#dnc,salad,milk,ate,making,home
32	Hapenning in DNC and RNC 2012	#dnc2012,#gop2012,convention,#rnc2012,speech,#rnc romney,obama,rnc,#dnc,dnc,tampa,ryan gop,stage,mitt,biden,charlotte,paul,music

background topic found by **TwitterLDA** model is not sematically clear; and (ii) the topical communities and their extreme topics found by **TCPI** model are both semantically clear and reasonable. In **Two-Week** dataset, other than *Daily life*, it is expected that political communities *Republicans' activities*, and *Campaining speeches* exist in the dataset as it was collected during a politically active period with many political events related to the American 2012 presidential election, e.g., the national conventions of both democratic (DNC 2012[6]) and republican (RNC 2012[7]) parties.

5 Conclusion

In this paper, we propose a novel topic model called **TCPI** for simultaneously modeling topical communities and users' topical interests in microblogging data. Our model differentiates users' personal interests from their topical communities while learning both the two set of latent factors at the same time. We also report experiments on two Twitter datasets showing the effectiveness of the proposed model. **TCPI** is shown to outperform TwitterLDA, another state-of-the-art topic model for modeling tweet generation.

In the future, we would like to consider the scalability of the proposed model. Possible solutions for scaling up the model are approximated and distributed implementations of Gibbs sampling procedures [17], and stale synchronous parallel implementation of variational inference procedures [7]. Moreover, it is potentially helpful to incorporate prior knowledge into the proposed model. Examples of the prior knowledge are topic indicative features [3], and groundtruth community labels for some users [25, 8].

[6] http://en.wikipedia.org/wiki/2012_Democratic_National_Convention
[7] http://en.wikipedia.org/wiki/2012_Republican_National_Convention

6 Acknowledgements

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

References

1. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Mixed membership stochastic blockmodels. J. Mach. Learn. Res. 9 (2008)
2. Balasubramanyan, R., Cohen, W.W.: Regularization of latent variable models to obtain sparsity. In: SDM13 (2013)
3. Balasubramanyan, R., Dalvi, B.B., Cohen, W.W.: From topic models to semi-supervised learning: Biasing mixed-membership models to exploit topic-indicative features in entity clustering. In: ECML/PKDD (2) (2013)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. (2003)
5. Ding, Y.: Community detection: Topological vs. topical. Journal of Informetrics 5(4), 498–514 (2011)
6. Grabowicz, P.A., Aiello, L.M., Eguiluz, V.M., Jaimes, A.: Distinguishing topical and social groups based on common identity and bond theory. In: WSDM (2013)
7. Ho, Q., Xing, E., et. al.: More effective distributed ml via a stale synchronous parallel parameter server. In: NIPS (2013)
8. Hoang, T.A., Cohen, W.W., Lim, E.P.: On modeling community behaviors and sentiments in microblogging. In: SDM14 (2014)
9. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: SOMA '10 (2010)
10. Hong, L., Dom, B., Gurumurthy, S., Tsioutsiouliklis, K.: A time-dependent topic model for multiple text streams. In: KDD (2011)
11. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: Understanding microblogging usage and communities. In: WebKDD/SNA-KDD '07 (2007)
12. Jurgen, A.: Twitter top 100 for software developers. In: http://www.noop.nl/2009/02/twitter-top-100-for-software-developers.html (2009)
13. Kooti, F., Yang, H., Cha, M., Gummadi, P.K., Mason, W.A.: The emergence of conventions in online social networks. In: ICWSM12 (2012)
14. Liu, J.S.: The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. J. Amer. Stat. Assoc (1994)
15. Mehrotra, R., Sanner, S., Buntine, W., Xie, L.: Improving lda topic models for microblogs via tweet pooling and automatic labeling. In: SIGIR (2013)
16. Michelson, M., Macskassy, S.A.: Discovering users' topics of interest on twitter: A first look. In: AND '10 (2010)
17. Newman, D., Asuncion, A., Smyth, P., Welling, M.: Distributed algorithms for topic models. The Journal of Machine Learning Research 10, 1801–1828 (2009)
18. Newman, M.E.J.: Modularity and community structure in networks. PNAS (2006)
19. Qiu, M., Jiang, J., Zhu, F.: It is not just what we say, but how we say them: Lda-based behavior-topic model. In: SDM (2013)
20. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: ECML (2009)
21. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: UAI (2004)

22. Sachan, M., Dubey, A., Srivastava, S., Xing, E.P., Hovy, E.: Spatial compactness meets topical consistency: Jointly modeling links and content for community detection. In: WSDM (2014)
23. Xie, P., Xing, E.P.: Integrating document clustering and topic modeling. In: UAI (2013)
24. Yang, J., Leskovec, J.: Community-affiliation graph model for overlapping network community detection. In: ICDM (2012)
25. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. In: ICDM (2012)
26. Yang, J., McAuley, J., Leskovec, J.: Community detection in networks with node attributes. In: ICDM (2013)
27. Yang, J., McAuley, J., Leskovec, J.: Detecting cohesive and 2-mode communities indirected and undirected networks. In: WSDM (2014)
28. Yin, Z., Cao, L., Gu, Q., Han, J.: Latent community topic analysis: Integration of community discovery with topic modeling. ACM TIST (2012)
29. Zhao, D., Rosson, M.B.: How and why people twitter: The role that micro-blogging plays in informal communication at work. In: GROUP '09 (2009)
30. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: ECIR (2011)
31. Zhou, D., Manavoglu, E., Li, J., Giles, C.L., Zha, H.: Probabilistic models for discovering e-communities. In: WWW'06 (2006)