Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems | School of Computing and Information Systems

4-2017

# Modeling topics and behavior of microbloggers: An integrated approach

Tuan Anh HOANG
*Singapore Management University*, tahoang.2011@phdis.smu.edu.sg

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

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the Databases and Information Systems Commons, Numerical Analysis and Scientific Computing Commons, and the Social Media Commons

## Citation

HOANG, Tuan Anh and LIM, Ee-Peng. Modeling topics and behavior of microbloggers: An integrated approach. (2017). *ACM Transactions on Intelligent Systems and Technology*. 8, (3), 44: 1-37.
**Available at:** https://ink.library.smu.edu.sg/sis_research/3727

This Journal Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Published in ACM Transactions on Intelligent Systems and Technology, 2017 April, Volume 8, Issue 3,
https://doi.org/10.1145/2990507

# Modeling Topics and Behavior of Microbloggers: An Integrated Approach

TUAN-ANH HOANG, L3S Research Center, Germany
EE-PENG LIM, Living Analytics Research Centre, Singapore Management University

Microblogging encompasses both user-generated content and behavior. When modeling microblogging data, one has to consider personal and background topics, as well as how these topics generate the observed content and behavior. In this article, we propose the *Generalized Behavior-Topic* (GBT) model for simultaneously modeling background topics and users' topical interest in microblogging data. GBT considers multiple topical communities (or realms) with different background topical interests while learning the personal topics of each user and the user's dependence on realms to generate both *content* and *behavior*. This differentiates GBT from other previous works that consider either *one realm* only or *content data* only. By associating user behavior with the latent background and personal topics, GBT helps to model user behavior by the two types of topics. GBT also distinguishes itself from other earlier works by modeling multiple types of behavior together. Our experiments on two Twitter datasets show that GBT can effectively mine the representative topics for each realm. We also demonstrate that GBT significantly outperforms other state-of-the-art models in modeling content topics and user profiling.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Data Mining

General Terms: Measurement, Algorithms, Performance

Additional Key Words and Phrases: Social media, microblogging, user behavior, behavior mining, topic modeling, probabilistic graphic model

**ACM Reference Format:**
Tuan-Anh Hoang and Ee-Peng Lim. 2017. Modeling topics and behavior of microbloggers: An integrated approach. ACM Trans. Intell. Syst. Technol. 8, 3, Article 44 (April 2017), 37 pages.
DOI: http://dx.doi.org/10.1145/2990507

## 1. INTRODUCTION

### 1.1. Motivation

Microblogging is the act of users publishing short messages, called *tweets*, on the Internet software platform Twitter to share their current status with their followers. Embedded in these tweets is a wide range of topics. Other than posting tweets, microblogging users also adopt behavior instances of different types as part of their interactions. Examples of *behavior types* include *networking* (i.e., following other users), *user mention* (i.e., mentioning other users in tweets), *hashtag usage* (i.e., inserting hashtags in tweets), and *retweeting* (i.e., forwarding tweets from other users). A user's actual

This work was completed when the first author was with Living Analytics Research Centre, Singapore Management University. 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.
Authors' addresses: T.-A. Hoang, L3S Research Center, Appelstraße 9A, 30167 Hannover; E.-P. Lim, Living Analytics Research Centre, Singapore Management University, 80 Stamford Road, Singapore 178902.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

behavior can therefore be represented as instances of different behavior types. The ease of content posting and behavior adopting from both desktop and mobile devices has made microblogging ideal for information sharing and seeking, social networking, and communication.

In microblogging, one of the paramount problems is to determine the topical interests of users and user communities. This problem is of utmost important for many applications, including user profiling [Pennacchiotti and Popescu 2011; Boutet et al. 2012], personalized recommendation [De Francisci Morales et al. 2012; Hannon et al. 2010; Qiu et al. 2013], and event detection [Sakaki et al. 2010; Diao et al. 2012; Diao and Jiang 2013]. There has been a number of works addressing the problem. They suffer from the following two major shortcomings: (i) they do not consider topical communities when modeling users' personal interest, and (ii) they learn users' interest from either their content (i.e., tweets) only or and their behavior only, but not both. We elaborate on these shortcomings here.

**Personal Interest and Topical Communities.** Empirical and user studies on microblogging usage have shown that the purpose of tweeting can be broadly attributed to users' personal topics or background topics [Java et al. 2007; Zhao and Rosson 2009; Kooti et al. 2012]. The former cover interests of the users themselves. The latter are the interests shared by users in the same topical communities [Grabowicz et al. 2013]. Instead of using the term *community* or *social community*, which usually refers to a social group of densely connected users [Prentice et al. 1994], we use the term *realm* to describe a topical user community. Users within a realm may not have many social ties among them, but they share some common background interest. In general, a user can belong to multiple realms. Thus, when modeling microblogging user content and behavior, we have to consider both the users' personal interest and their realms. Previous works do not consider realms, however. Some do not model background topics at all (e.g., Hong and Davison [2010], Ramage et al. [2010], and Yang et al. [2014]). Others assume that there is only a single background topic (e.g., Hong et al. [2011], Zhao et al. [2011], Qiu et al. [2013], and Xie and Xing [2013]). Without considering realms and background topics, the previous models would not be able to describe the users' personal interests accurately.

Consider the example in Figure 1. There are two realms: *Food* and *Politics*. Both *user-A* and *user-B* belong to the two realms; therefore, they sometimes tweet about the realms' topics. For example, *user-A* and *user-B* mention food in *tweet*-3 and *tweet*-7, respectively; they also mention politics in *tweet*-4 and *tweet*-8, respectively. They also adopt the realms' representative behavior instances. Being part of the *Food* realm, they use hashtag *#foods*, and follow and retweet from *HealthyLiving*[1]. Similarly, they use hashtags *#p2*,*#tcot*, *#elections*, and *#MittRomney*, and follow and retweet from *BarackObama*[2] and *MittRomney*[3] due to their association with the *Politics* realm. The existing models, in the absence of realms, would incorrectly treat the two realms' topics as users' personal interests.

A naïve approach to learning both users' personal interests and their realms is to learn them in two steps. The choices of which step to go first leads us to two solutions. The first solution is (i) to perform topic modeling on users' content and behavior, followed by (ii) assigning the most common topics among all the users to be the realms' topics. The second solution is (a) to detect users' realms, then (b) to determine users' personal interests based on the detected realms and users' realm membership. These two solutions are inherently suboptimal due to separation of steps, and also require ad

[1]https://twitter.com/healthyliving.
[2]https://twitter.com/barackobama.
[3]https://twitter.com/mittromney.

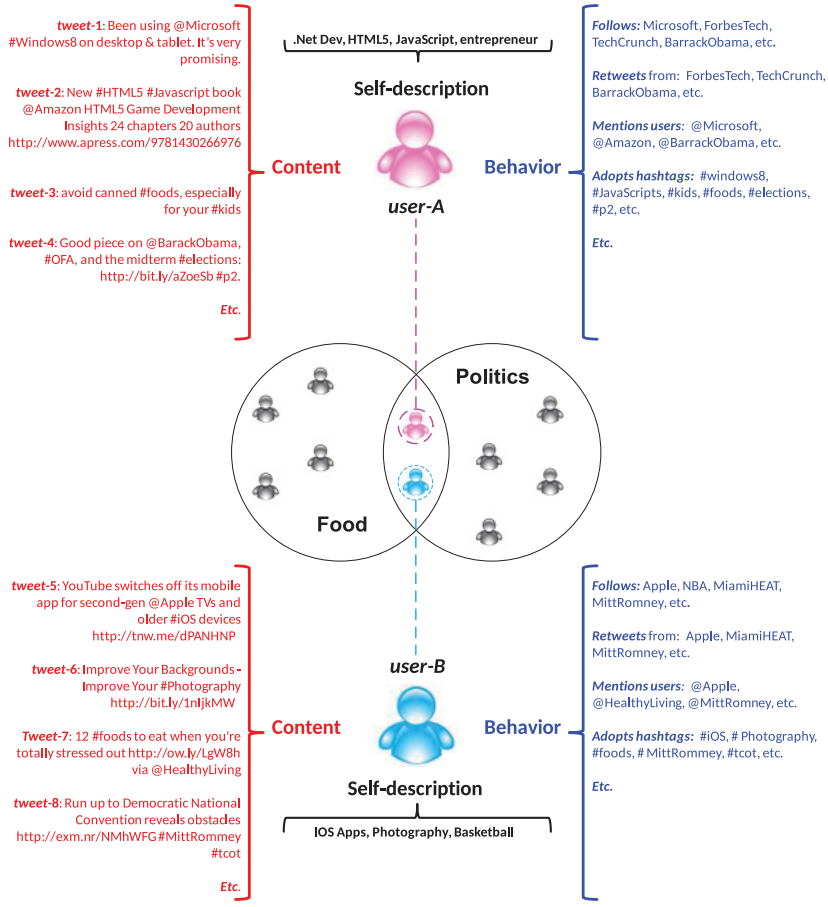



Fig. 1. This example illustrates how microblogging users' personal interest and their realms determine their content and behavior.

hoc treatments for converting the output of the first step to become the input of second step.

**User Content and User Behavior.** Topical interests determine both content and behavior of users. For example, in Figure 1, *user-A* is interested in Microsoft's .NET framework, HTML5, and entrepreneurship (as stated in *user-A*'s self-description); thus, the user mentions and retweets from *Microsoft* and, and adopts hashtags such as #*windows8* and #*JavaScripts*. Also, due to topics of *user-A*'s realms, the user follows, mentions, and retweets from *BarackObama*, and adopts hashtags such as #*kids*, #*food*, #*p2*, and #*elections*. Similarly, *user-B* is interested in IOS applications, thus mentions and retweets from *Apple*; and adopts the hashtag #*ios*. Also, due to *user-B*'s association with the *Politics* realm, the user follows, mentions, and retweets from *MittRomney*, and adopts hashtags such as #*food*, #*tcot*, and #*MittRomney*.

To the best of our knowledge, there is no previous work on modeling microblogging users that employs both user content and user behavior. Most of the existing works either model topics of user content only [Hong and Davison 2010; Zhao et al. 2011] or user behavior only [Ma et al. 2013; Luo et al. 2013]. These works neglect the relationship between user content and user behavior, thus learn the users' interests in a less-than-optimal manner. A user's topical interest may show up in the user's content or behavior,

but not both. For example, in Figure 1, *user-A* is interested in entrepreneurship, motivating the user to follow and retweet from *ForbesTech*[4] and *TechCrunch*[5] even though hardly tweeting on entrepreneurship. Similarly, *user-B* is interested in basketball and follows and retweets from *NBA*[6] and *MiamiHEAT*[7] even though the user may not have tweeted about basketball.

Few other works, consider both user content and user behavior together. Sachan et al. [2012] and Qiu et al. [2013] model the types of user behavior associated with the content. For example, a message may be associated with *tweeting* or *retweeting* types. These works therefore can model only a subset of user behavior types, and do not model the user behavior instances (e.g., who is retweeted, which hashtag is used, and so on). Aggregating users' behavior instances by their types is an oversimplification that leads to less accurate models.

We also consider some obvious approaches that combine users' content (i.e., tweets) and behavior in learning their topical interest:

—*Deriving two separate topic models for users' tweets and behavior. A user's topics are then the concatenation of the user's topics from the two models.* For example, using an existing topic model (e.g., LDA [Blei et al. 2003], Author-Topic [Rosen-Zvi et al. 2004], or TwitterLDA [Zhao et al. 2011]), we obtain topic distributions $\theta_{content}(u)$ and $\theta_{behavior}(u)$ for user $u$ from user $u$'s tweets and behavior, respectively. The topical interest of $u$ is then $[\theta_{content}(u), \theta_{behavior}(u)]$. However, this approach requires a large number of topics and the same topic cannot be associated with both content and behavior.

—*Performing topic modeling on the tweets, followed by assigning each user behavior instance with the topic(s) of its associated tweet(s)*. For example, for each adoption of hashtag $h$, we assign to $h$ the topic(s) of the tweet containing $h$. This approach does not work well for two reasons. First, the topics of some tweets cannot be accurately identified due to their very short and noisy content. Second, the topic of the tweet content does not always fully explain the behavior. For example, recent works have shown that microbloggers use hashtags for many other purposes beyond topic labeling, including personalized bookmarking, named entity markup [Zappavigna 2011], and community membership [Yang et al. 2012].

—*Performing supervised topic modeling on the tweets with their associated behavior instances as labels*. For example, we may apply the Labeled-LDA model [Ramage et al. 2010] on the tweets using their hashtag(s) as topic labels. Again, this approach is not ideal since tweets using the same hashtag do not always share the same topics, as mentioned earlier.

### 1.2. Research Objectives

In this work, we aim to address these shortcomings by introducing realms as well as users' topical interests in modeling of both content and behavior of microbloggers. We seek to learn realms representing collective topical interests, in addition to users' personal topical interests. We also want to model the user's dependence on realms when generating content and adopting behavior.

To meet these objectives, we use an integrated approach. That is, we propose to jointly model user topical interests and realms' topic distributions in the same framework. In this framework, each user is assigned a variable to learn the user's bias towards the

[4]https://twitter.com/ForbesTech.
[5]https://twitter.com/TechCrunch.
[6]https://twitter.com/NBA.
[7]https://twitter.com/MiamiHEAT.

user's personal interests or associated realms. We also propose to jointly model user content and user behavior sharing a common set of latent topics. This integrated approach has several advantages. First, we can learn both users' personal interests and interest of their realms in one step. Second, the approach allows the interests to be modeled accurately by using both user content and user behavior. Third, by modeling both content and behavior with the same set of topics, we are able to infer user behavior using the content and vice versa, as well as to semantically interpret user behavior. For example, we may infer that a user who tweets frequently about political topics is also more likely to mention and retweet from politicians.

### 1.3. Summary of Contributions

In this work, we develop a general framework that allows different types of behavior to be modeled as different bag-of-behavior instances. We then develop a probabilistic graphical model that simultaneously infers latent topics, users' topical interests, and latent realms. Our main contributions in this work are as follows.

—We propose a probabilistic graphical model, called the *Generalized Behavior-Topic* model (GBT), for modeling topical interests of users and their realms, as well as for modeling both user content and user behavior using a common set of topics. In GBT, the dependency of the users on realms in generating content and adopting behavior are variables to be learned. This is a unique contribution of this work since each user's dependence on realms is not observable in the data.
—We develop a simple sampling method to infer the model's variables. We further develop an efficient regularization technique to bias the model to learn more semantically clearer realms. Our learning method is easy to implement and scales with the number of latent topics and realms, as well as the number of observed content words and behavior instances.
—We apply the GBT model to two Twitter datasets and show that it significantly outperforms state-of-the-art topic models for Twitter content.
—An empirical analysis of topics and realms for the two datasets has been conducted to demonstrate the efficacy of the GBT model.
—Last, we further demonstrate the application of the GBT model in some user profiling tasks, showing that it also outperforms other topic models in these tasks.

The rest of the article is organized as follows. We discuss the related works on topic modeling of user content and behavior in Section 2. We present our proposed model in Section 3. We describe two experimental datasets and report the results of applying the proposed model on the two datasets in Section 4. We then report the results of evaluating the proposed model and other topic models in some user-profiling tasks in Section 5. Finally, we present our conclusions and discuss future work in Section 6.

## 2. RELATED WORK

In this section, we review previous works closely related to ours. These works fall into three categories: (i) works on analyzing topics in microblogging data, (ii) works on analyzing user behavior in microblogging data, and (iii) works on analyzing communities in social networks.

### 2.1. Topic Analysis

Michelson and Macskassy [2010] empirically analyzed microblogging users' topical interests by examining named entities mentioned in tweets. Hong and Davison [2010] then conducted an empirical study on different ways of performing topic modeling on tweets using the original LDA model [Blei et al. 2003] and Author-Topic model [Rosen-Zvi et al. 2004]. They found that the topics learned from documents formed by

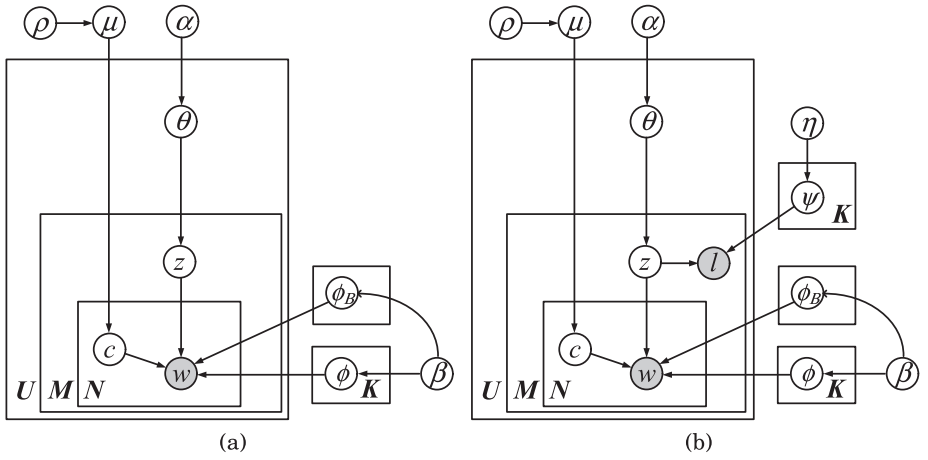



Fig. 2. (a) TwitterLDA and (b) QBLDA models.

aggregating tweets posted by the same users could help in user profiling. Similarly, Mehrotra et al. [2013] investigated ways of forming documents from tweets in order to improve the performance of the LDA model. They found that grouping tweets by hashtag could lead to an improvement in quality of the learned topics. Ramage et al. [2010] further proposed using the Labeled LDA model [Ramage et al. 2009] to model topics of tweets, in which each tweet is labeled by its linguistic elements (e.g., hashtags or emoticons). Lim and Buntine [2014] and Tan et al. [2014b] proposed incorporating the sentiment of tweets into the LDA model. Wang et al. [2014] proposed regularizing the LDA model by user network.

TwitterLDA [Zhao et al. 2011], the state-of-the-art of microblogging content topic models, is a variant of LDA, in which (i) documents are formed by aggregating tweets of the same users, (ii) a single background topic is assumed, (iii) each tweet has only one topic shared by all words of the tweet, and (iv) each word in a tweet is generated from either the background topic or the tweet's topic. The plate notation of the TwitterLDA model is shown in Figure 2(a), and its generative process is as follows.

—Sample the background topic's word distribution $\phi_B \sim Dirichlet(\beta)$
—For each topic $k$ ($k = 1, \ldots, K$), sample the topic's word distribution $\phi_k \sim Dirichlet(\beta)$
—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 the topic: $z \sim Multinomial(\theta_u)$
(2) Sample the words: For the $i$th word of the tweet
—Sample the source: $c \sim Bernoulli(\mu)$
—If $c = 0$, sample the word from background topic: $w \sim Multinomial(\phi_B)$; else ($c = 1$), sample the word from the tweet's topic: $w \sim Multinomial(\phi_z)$

The TwitterLDA model, however, does not consider multiple background topics, and assumes that all users have the same dependency on the unique background topic as the parameter $\mu$ is shared by all the users.

Vosecky et al. [2014] proposed jointly modeling multiple types of named entities embedded in tweets. Yan et al. [2013] and Cheng et al. [2014] proposed modeling the generation of co-occurrence of pairs or words instead of modeling the occurrence of each single word. Lin et al. [2014] proposed exploiting the sparsity of both topic

distributions and topic–word distributions in modeling topics of tweets. Last, Yang et al. [2014] proposed a classification approach to assign tweets to predefined topics.

All these works consider only user content. Our work, on the other hand, considers both user content and user behavior.

The work in Qiu et al. [2013] is most related to our work. With same assumptions as with TwitterLDA, the authors proposed modeling topics of tweets using both the tweets' content and the types of their *associated behavior instances* (i.e., either a tweet is a *(original) tweet* or *retweet*, and so on). Their proposed model, denoted here by the QBLDA model, has the plate notation shown in Figure 2(b), and the generative process similar to that of the TwitterLDA model, except for one more step to generate the behavior type associated with the tweet. In QBLDA, after the topic $z$ of tweet $t$ is sampled, its associated behavior type $l$ is generated from the topic-specific behavior type distribution $\psi_z$ where $\psi_z$ is a multinomial distribution over $L$ types of user behavior. Similar to the topics' word distribution, $\psi_k$ has Dirichlet prior $\eta$ for each topic $k$. That is,

—$\psi_k \sim Dirichlet(\eta)$ for each topic $k$ and
—$l \sim Multinomial(\psi_z)$

QBLDA models user behavior types only (e.g., *retweet* behavior type), but not the behavior instances (e.g., who is retweeted). This model therefore considers only a subset of user behavior types that are exclusively associated with tweets (i.e., there is only one type of behavior associated with each tweet). Also, in the QBLDA model, users' content may be replicated in learning the users' interests, leading to a less accurate model. For example, if a tweet is retweeted 3 times, then 4 copies of the tweet are required: 1 associated with *tweet* behavior type, and 3 associated with *retweet* behavior type. Our work is more general than that of QBLDA by decoupling user content with its associated user behavior instances. Our goal is to include the unified modeling of user behavior instances of multiple types to more accurately model user interest.

Our work is also related to the works on event detection and trend analysis in social media (e.g., Hu et al. [2012], Diao et al. [2012], Gao et al. [2012], Yin et al. [2013], and Diao and Jiang [2013]). Events are different from realms as the former consist of topics that are bursty and popular within a short duration of time. In contrast, realms consist of topics that are not necessary bursty, but remain popular for a much longer time. The works on bursty event detection and trend analysis also do not consider user behavior. Last, our work is also similar but not exactly the same as works on modeling global topics (e.g., Hong et al. [2011] and Xie and Xing [2013]). Although global topics are also popular and last for a long time, they are shared by all users as opposed to users of some realm. Modeling global topics is therefore a special case of our work when the number of realms is degenerated to one.

### 2.2. User Behavior Analysis

There has been a number of works analyzing user behavior in microblogging data. Kwak et al. [2010, 2011], Wu et al. [2011], and Feller et al. [2011] studied the patterns of following behavior. Hannon et al. [2010], Yin et al. [2011], and Barbieri et al. [2014] proposed models for recommending following behavior. Suh et al. [2010], Conover et al. [2011], Wu et al. [2011], and Tan et al. [2014a] studied retweeting behavior. Welch et al. [2011] conducted empirical research showing that a user's retweeting behavior is a stronger indicator of the user's topical interest than the user's following behavior. Yang and Counts [2010], Dabeer et al. [2011], Cui et al. [2011], Chen et al. [2012], Pan et al. [2013], and Yan et al. [2012] proposed models for retweeting behavior. However, most of these works (i) consider only a single type of user behavior or (ii) do not

consider content when modeling user behavior. Our model extends the state-of-the-art by modeling different types of user behavior simultaneously when modeling content.

There are also existing works that jointly model user-generated content and user behavior employing the same latent space. The works by Erosheva et al. [2004], Nallapati et al. [2008], Yano et al. [2009], and Ma et al. [2015] are among them. These works neither model realms nor consider the existence of different types of user behavior, however.

### 2.3. Community and Realm Analysis

In social networks, communities may be formed by users developing dense social ties with other users or sharing common interests with others. This results in different types of communities: social, topical, or hybrid [Prentice et al. 1994; Grabowicz et al. 2013]. Most of the early works on community mining focus on finding social communities that have dense social links among the community users. Newman [2006] proposed discovering social communities by finding a network partition that maximizes a measure of "compactness" in community structure called *modularity*. Airoldi et al. [2008] et al. propose a statistical mixed membership model.

Research works on finding topical communities (i.e., realms) include those based on user-generated content (e.g., Zhou et al. [2006] and Xie and Xing [2013]), and user attributes (e.g., Yang and Leskovec [2012] and Yang et al. [2013, 2014]). Ding [2011] conducted an empirical study showing that social community structure of a social network can be significantly different from realms discovered from the same network. However, most of these works do not differentiate users' personal interest from that of topical communities. They assume that a user's topical interest is determined purely based on the user's topical communities' interests. This assumption is not practical in the microblogging context since microbloggers cover a vast range of interest topics, which are not always determined by their topical communities. Our model therefore seeks to differentiate a user's personal interest from that of the user's realms.

Last, it is important to note that our work is different from works on finding topical interest of social communities (e.g., Liu et al. [2009], McCallum et al. [2005], Li et al. [2010], Lim and Datta [2012], Sachan et al. [2012], Yin et al. [2012], and Sachan et al. [2014]). Topical interest of each social community in these works refers to the most common topics shared by users within a social community, thus may not be unique to the community. Two different social communities may share the same topical interest. Our proposed model requires each realm to be uniquely determined based on its topical interest. Different realms are required to have distinctive interest.

## 3. GENERALIZED BEHAVIOR-TOPIC MODEL

In this section, we present our proposed *Generalized Behavior-Topic* (GBT) model in detail. We begin by introducing notations and concepts. Next, we describe the principles in designing the model and its generative process. We also highlight some properties of the GBT model and its differences from the state-of-the-art models. Last, we present an algorithm for learning the model's variables that utilizes a regularization technique to bias the learning process for clearer topics and realms.

### 3.1. Notations and Preliminaries

We summarize the notations in Table I. We use $\mathcal{U}$ to denote the set of all users, and use $U$ to denote the number of users, that is, $U = |\mathcal{U}|$. For each user $u \in \mathcal{U}$, we denote the set and the number of tweets posted by $u$ by $\mathcal{T}_u$ and $T_u$, respectively, that is, $T_u = |\mathcal{T}_u|$. Then, $\mathcal{T}$ and $T$ denote the set and the number of all tweets posted by all users, respectively, that is, $\mathcal{T} = \cup_u \mathcal{T}_u$ and $T = \sum_u T_u$. The $j$th tweet posted by $u$ is denoted by $t_u^j$ and the

Table I. Notations

| | |
|---|---|
| $\mathcal{U}$/ $U$ | Set/ number of users, that is, $U = \lvert\mathcal{U}\rvert$ |
| $\mathcal{T}_u$/ $T_u$ | Set/ number of tweets posted by user $u$, that is, $T_u = \lvert\mathcal{T}_u\rvert$ |
| $\mathcal{T}$/ $T$ | Set/ number of tweets posted by all users, that is, $\mathcal{T} = \cup_u \mathcal{T}_u$ and $T = \sum_u T_u$ |
| $t_u^j$ | Tweet number $j$ ($j = 1, \ldots, T_u$) of user $u$ |
| $N_u^j$ | Number of words in tweet $t_u^j$ |
| $w_u^{ji}$ | Word number $i$ ($i = 1, \ldots, N_u^j$) in tweet $t_u^j$ |
| $\mathcal{W}$ | Bag-of-words from all tweets |
| $\mathcal{V}_t$/ $V_t$ | Word vocabulary/ number of words in vocabulary, that is, $V_t = \lvert\mathcal{V}_t\rvert$ |
| $L$ | Number of behavior types |
| $\mathcal{B}_u^l$ | Bag-of-behavior instances of type-$l$ ($l = 1, \ldots, L$) that user $u$ adopts |
| $B_u^l$ | Number of behavior instances of type-$l$ that user $u$ adopts, that is, $B_u^l = \lvert\mathcal{B}_u^l\rvert$ |
| $\mathcal{B}$ | Bag-of-behavior instances of all types and adopted by all users, that is, $\mathcal{B} = \{\mathcal{B}_u^l : \forall u \in \mathcal{U}$ and $\forall l = 1, \ldots, L\}$ |
| $b_u^{lj}$ | $j$th behavior instance of type-$l$ ($j = 1, \ldots, B_u^l$) that user $u$ adopts |
| $\mathcal{V}_b^l$/ $V_b^l$ | Type-$l$ behavior vocabulary/ number of behavior instances in the vocabulary, that is, $V_b^l = \lvert\mathcal{V}_b^l\rvert$ |
| $K$/ $R$ | Number of topics/ realms |
| $\phi_k$/ $\lambda_k^l$ | Word/ type-$l$ behavior instance distribution of $k$th topic |
| $\sigma_r$ | Topic distribution of realm $r$ |
| $\theta_u$/ $\pi_u$ | Topic/ realm distribution of user $u$ |
| $\mu_u$ | Dependence distribution of user $u$ |
| $\alpha/\beta/\eta/\rho/\tau/\gamma_l$ | Dirichlet (beta) conjugate priors of $\theta_u/\phi_k/\sigma_r/\mu_u/\pi_u/\lambda_k^l$ |
| $c_u^j$/ $r_u^j$/ $z_u^j$ | Source /realm/ topic of tweet $t_u^j$ |
| $c_u^{lj}$/ $r_u^{lj}$/ $z_u^{lj}$ | Source/ realm/ topic of behavior $b_u^{lj}$ |
| $\mathcal{C}$/ $\mathcal{R}$/ $\mathcal{Z}$ | Bag-of-sources/ realms/ topics of all the tweets and behavior instances |
| $\mathcal{C}_{-t_u^j}$/ $\mathcal{R}_{-t_u^j}$/ $\mathcal{Z}_{-t_u^j}$ | Bag-of-sources/ realms/ topics of all behavior instances and tweets except $t_u^j$ |
| $\mathcal{C}_{-b_u^{lj}}$/ $\mathcal{R}_{-b_u^{lj}}$/ $\mathcal{Z}_{-b_u^{lj}}$ | Bag-of-sources/ realms/ topics of all behavior instances and tweets except $b_u^{lj}$ |
| $\mathcal{O}_{-t_u^j}$ | Short form of the tuple $(\mathcal{T}, \mathcal{B}, \mathcal{C}_{-t_u^j}, \mathcal{R}_{-t_u^j}, \mathcal{Z}_{-t_u^j}, \alpha, \beta, \eta, \rho, \gamma_1, \ldots, \gamma_L)$ |
| $\mathcal{O}_{-b_u^{lj}}$ | Short form of the tuple $(\mathcal{T}, \mathcal{B}, \mathcal{C}_{-b_u^{lj}}, \mathcal{R}_{-b_u^{lj}}, \mathcal{Z}_{-b_u^{lj}}, \alpha, \beta, \eta, \rho, \gamma_1, \ldots, \gamma_L)$ |
| $\mathbf{n_c}(c, u, \mathcal{C})$ | #times source $c$ is observed in set of tweets and behavior instances of user $u$ for bag-of-sources $\mathcal{C}$ |
| $\mathbf{n_{zu}}(z, u, \mathcal{Z})$ | #tweets + #behavior instances of user $u$ that have source 0 and have topic $z$ for bag-of-topics $\mathcal{Z}$ |
| $\mathbf{n_{zr}}(z, r, \mathcal{Z}, \mathcal{R})$ | #tweets + #behavior instances that have source 1 and have topic $z$ and realm $r$ for bag-of-topics $\mathcal{Z}$, and bag-of-realms $\mathcal{R}$ |
| $\mathbf{n_w}(w, z, \mathcal{T}, \mathcal{Z})$ | #times word $w$ is observed in topic $z$ for set of tweets $\mathcal{T}$ and bag-of-topics $\mathcal{Z}$ |
| $\mathbf{n_b^l}(b, z, \mathcal{B}, \mathcal{Z})$ | #times type-$l$ behavior $b$ is observed in topic $z$ for bag-of-behavior instances $\mathcal{B}$ and bag-of-topics $\mathcal{Z}$ |

number of words in $t_u^j$ is denoted by $N_u^j$. We denote the $i$th word of tweet $t_u^j$ by $w_u^{ji}$. The bag of words from all the tweets is denoted by $\mathcal{W}$. Last, we denote the vocabulary of all words by $\mathcal{V}_t$ and the number of words in the vocabulary by $V_t$, that is, $V_t = \lvert\mathcal{V}_t\rvert$.

We denote the number of behavior types by $L$, and denote the types by type-1 to type-$L$, respectively. A user may adopt the same behavior instance multiple times. We therefore use a bag-of-behavior instances to represent a user's behavior instances of each type.

We denote the bag-of-behavior instances of type-$l$ that $u$ adopts by $\mathcal{B}_u^l$ and the number of behavior instances in the bag by $B_u^l$, that is, $B_u^l = \lvert\mathcal{B}_u^l\rvert$. Similar to words in tweets, we denote the $j$th behavior instance of type-$l$ that $u$ adopts by $b_u^{lj}$. The bag-of-behavior instances of all types and all users is denoted by $\mathcal{B}$, that is, $\mathcal{B} = \{\mathcal{B}_u^l : \forall u \in \mathcal{U}$ and $\forall l = 1, \ldots, L\}$. Finally, we denote the vocabulary of type-$l$ behavior (i.e., the set of all behavior

instances of type-$l$) by $\mathcal{V}_b^l$, and the number of behavior instances in the vocabulary by $V_b^l$, that is, $V_b^l = |\mathcal{V}_b^l|$.

We now give formal definitions of the main concepts used in this article.

*Definition* 1 (*Topic*). A topic is a semantically coherent theme of words and behavior instances. Formally, a topic $z$ is represented by $(1+L)$-tuple $(\phi_z, \lambda_z^1, \ldots, \lambda_z^L)$, where $\phi_z$ is a multinomial distribution over word vocabulary $\mathcal{V}_t$, and $\lambda_z^l$ is a multinomial distribution over type-$l$ behavior vocabulary $\mathcal{V}_b^l$ for $\forall l = 1, \ldots, L$. $\phi_z$ and $\lambda_z^l$ are called the word distribution and type-$l$ behavior distribution of topic $z$, respectively.

For example, a political campaign topic would have high probabilities for words such as *policies*, *debates*, and *votes*, but low probabilities for other words. The topic also has high probabilities for behavior instances such as mentions of and retweeting from politicians, and low probabilities for other behavior instances. Another topic about foods would have high probabilities for words such as *coffee* and *drinks*, and low properties for other words. It would also have high probabilities for behavior instances such as adoptions of #*restaurants* and #*cuisine* hashtags, as well as mentions of and retweeting from famous food businesses and bloggers, and low probabilities for other behavior instances.

*Definition* 2 (*Users' Topic Distribution*). The topic distribution of a user represents the user's preference levels for different topics. Formally, the topic distribution of user $u$ is a multinomial distribution $\theta_u$ over the set of all topics.

For example, a user interested in sports would have a high probability for sports topics and a low probability for other topics. Similarly, another user interested in technology would have a high probability for technology topics and a low probability for other topics.

*Definition* 3 (*Realms' Topic Distribution*). The topic distribution of a realm represents the common topical interests of the realm's user members. Formally, the topic distribution of realm $r$ is a multinomial distribution $\sigma_r$ over the set of all topics.

For example, the political realm would have high probabilities for topics such as political parties, campaigns, and elections; the IT realm would have high probabilities for topics such as programming, software development, and big IT companies.

*Definition* 4 (*Realm Membership*). A user's realm membership refers to the user's preference in different realms. Formally, realm membership of user $u$ is represented by a multinomial distribution $\pi_u$ over the realms, which is called realm distribution of $u$.

For example, a politician would have high probabilities for the political realms and a celebrity would have high probabilities for entertainment realms

### 3.2. Model Design Principles

The GBT model is designed to simulate the process of generating some observed tweet and behavior data from $K$ topics and $R$ realms. Here, $K$ and $R$ are data-dependent parameters and determined empirically. In the GBT model, tweets and behavior instances are assumed to be conditionally independent given the topics, realms, and users. Every tweet and behavior instance of user $u$ is assigned a topic. This topic is determined based on either $u$'s personal interest or $u$'s realms. To capture the bias of $u$ towards $u$'s personal interest or realms when determining the topics of $u$'s tweets and behavior instances, we assign to $u$ a dependence distribution $\mu_u$. Here, $\mu_u = (\mu_u^0, \mu_u^1)$ is a Bernoulli distribution. $\mu_u^0$ is the bias of $u$ towards $u$'s personal interest, and $\mu_u^1 = 1 - \mu_u^0$ is $u$'s bias

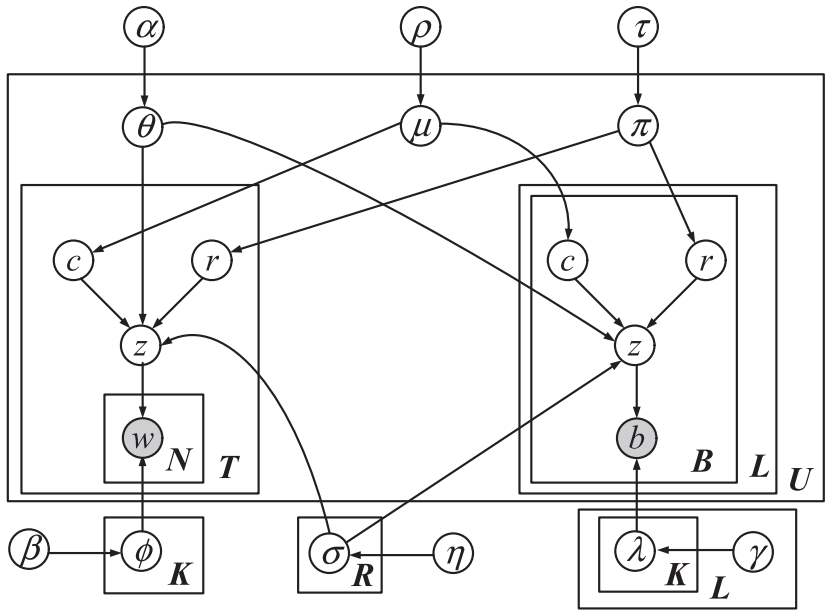



Fig. 3. Plate diagram of GBT model.

towards $u$'s realms. Once a tweet's topic is determined, its words are then determined based on the topic's word. Similarly, once a behavior instance's topic is determined, the instance is then determined based on the topic's behavior distributions.

### 3.3. Generative Process

The generative process of the GBT model has a plate diagram as shown in Figure 3 and is described here.

For every topic $k = 1, \ldots, K$, we assume that the topic's word distribution $\phi_k$ and behavior distributions $\lambda_k^l$ ($l = 1, \ldots, L$) follow known conjugate Dirichlet priors $\beta$ and $\gamma_l$, respectively. Similarly, for every user $u$, we also assume that $u$'s topic distribution $\theta_u$ and realm distribution $\pi_u$ follow known conjugate Dirichlet priors $\alpha$ and $\tau$, respectively, while the user's dependence distribution $\mu_u$ follows a known conjugate beta prior $\rho$. Last, for every ream $r$, we assume that $r$'s topic distribution $\sigma_r$ follows a known conjugate Dirichlet prior $\eta$.

To generate a tweet $t$ posted by user $u$, we first sample a binary variable $c$ from the user's dependence distribution $\mu_u$. The variable $c$ is called the *source* of tweet $t$, and is used to decide if the tweet will be based on $u$'s personal interest, or one of $u$'s realms. If $c = 0$, we then choose the topic $z$ for the tweet according to $u$'s topic distribution $\theta_u$. Otherwise, that is, $c = 1$, we first choose a realm $r$ according to $u$'s realm distribution $\pi_u$, then we choose $z$ according to the chosen realm's topic distribution $\sigma_r$. As tweets are short, with no more than 140 characters, we assume that each tweet has only one topic, and the tweet's words are conditionally independent given its topic. Once the topic $z$ is chosen, words in $t$ are then chosen according to the topic's word distribution $\phi_z$, and each word is chosen independently from the others.

Similarly, we assume the same process for all adopted behavior instances, except that, for a behavior instance $b$ of type-$l$, once its topic $z$ is chosen, the instance is then chosen according to the topic's behavior distribution $\lambda_z^l$.

The full generative process is summarized as follows.

—Generate topics' word and behavior distributions from their priors
For each topic $k = 1, \ldots, K$
—Sample the topic's word distribution from its Dirichlet prior: $\phi_k \sim Dirichlet(\beta)$
—For each type of behavior $l = 1, \ldots, L$, sample the topic's type-$l$ behavior distribution from its Dirichlet prior: $\lambda_k^l \sim Dirichlet(\gamma_l)$
—Generate realms' topic distributions from their prior
For each realm $r$ ($r = 1, \ldots, R$), sample the realm's topic distribution from its Dirichlet prior: $\sigma_r \sim Dirichlet(\eta)$
—Generate users' topic, realm, and dependence distributions from their priors
For each user $u$
(1) Sample $u$'s topic distribution $\theta_u \sim Dirichlet(\alpha)$
(2) Sample $u$'s realm distribution $\pi_u \sim Dirichlet(\tau)$
(3) Sample $u$'s dependence distribution $\mu_u \sim Beta(\rho)$
—Generate users' content and behavior
For each user $u$
—Generate tweets
For each tweet $t$ that $u$ posts
(1) Sample the source: $c \sim Bernoulli(\mu_u)$
(2) Sample topic:
—if $c = 0$, sample the topic from $u$'s topic distribution: $z \sim Multinomial(\theta_u)$
—If $c = 1$, sample the topic from one of the realms:
—Sample the realm: $r \sim Multinomial(\pi_u)$
—Sample the topic: $z \sim Multinomial(\sigma_r)$
(3) Sample the tweet's words: For each word of the tweet, sample the word: $w \sim Multinomial(\phi_z)$
—Generate behavior
For each behavior instance of type-$l$ that $u$ adopts:
(1) Sample the source: $c \sim Bernoulli(\mu_u)$
(2) Sample the topic:
—If $c = 0$, sample the topic from $u$'s topic distribution: $z \sim Multinomial(\theta_u)$
—If $c = 1$, sample the topic from one of the realms
—Sample the realm: $r \sim Multinomial(\pi_u)$
—Sample the topic: $z \sim Multinomial(\sigma_r)$
(3) Sample the behavior instance: $b \sim Multinomial(\phi_{z_b})$

### 3.4. Discussion

The GBT model shares the same idea with the TwitterLDA and QBLDA models (see Section 2.1) that topics are assigned to tweets instead of words. GBT differs from the two predecessors by considering realms in addition to users' personal interest. It also accommodates multiple types of user behavior simultaneously. Moreover, users' tweets and behavior are modeled by GBT using the same set of topics, thus keeping the modeling complexity unchanged when adding more user behavior types.

Also, like TwitterLDA, QBLDA, and other LDA-based models, GBT is a Bayesian clustering model. Hence, when fitting a given dataset, *likelihood* and *perplexity* of the GBT model are sensitive to its number of variables. The variables in the GBT model are:

—$K \times V_t$ variables for $K$ topics' word distributions, and $K \times \sum_{l=1}^{L} V_b^l$ variables for the topics' behavior distributions.
—$R \times K$ variables for $R$ realms' topic distributions.

—$U \times K$ variables for $U$ users' topic distributions, $U \times R$ variables for the users' realm distributions, and $U \times 2$ for the users' dependence distribution.

In total, the number of variables in the GBT model is

$$K \times V_t + K \times \sum_{l=1}^{L} V_b^l + R \times K + U \times K + U \times R + U \times 2 = K \times \left( U + V_t + \sum_{l=1}^{L} V_b^l \right) + R \times (U + K) + 2 \times U$$

Since $R \ll K \ll U \ll V_t + \sum_{l=1}^{L} V_b^l$, the number of variables in the GBT model is significantly increased when we increase $K$, but not when we increase $R$. Therefore, *likelihood* and *perplexity* of the GBT model would be significantly increased and decreased, respectively, when $K$ is increased, but not when $R$ is increased.

### 3.5. Model Learning

Due to the intractability of LDA-based models [Blei et al. 2003], we make use of a sampling method for estimating the parameters in the GBT model from a given dataset and the priors. More specifically, we first randomly initialize the latent source, latent realm, and latent topic for all tweets and behavior instances in the given dataset. We then use a collapsed Gibbs sampler ([Liu 1994]) to iteratively sample the source, realm, and latent topic of every tweet and every behavior instance to obtain a sample set to estimate the model's parameters.

We denote the bag-of-topics, bag-of-sources, and bag-of-realms of all the tweets and behavior instances in the given dataset by $\mathcal{Z}$, $\mathcal{C}$, and $\mathcal{R}$, respectively. For each tweet $t_u^j$, we use $\mathcal{C}_{-t_u^j}$, $\mathcal{R}_{-t_u^j}$, $\mathcal{Z}_{-t_u^j}$ to denote the bag-of-sources, bag-of-realms, and bag-of-topics, respectively, of all the behavior instances and all other tweets in the given dataset except $t_u^j$. To simplify the notations, we use $\mathcal{O}_{-t_u^j}$ to refer to the tuple $(\mathcal{T}, \mathcal{B}, \mathcal{C}_{-t_u^j}, \mathcal{R}_{-t_u^j}, \mathcal{Z}_{-t_u^j}, \alpha, \beta, \eta, \rho, \gamma_1, \ldots, \gamma_L)$.

Similarly, for each adopted behavior instance $b_u^{lj}$, we use $\mathcal{C}_{-b_u^{lj}}$, $\mathcal{R}_{-b_u^{lj}}$, $\mathcal{Z}_{-b_u^{lj}}$ to denote the bag-of-sources, bag-of-realms, and bag-of-topics, respectively, of all the tweets and all other behavior instances in the dataset except $b_u^{lj}$. Also, we use $\mathcal{O}_{-b_u^{lj}}$ to refer to the tuple $(\mathcal{T}, \mathcal{B}, \mathcal{C}_{-b_u^{lj}}, \mathcal{R}_{-b_u^{lj}}, \mathcal{Z}_{-b_u^{lj}}, \alpha, \beta, \eta, \rho, \gamma_1, \ldots, \gamma_L)$.

**Sampling for a tweet.** After we randomly initialize source, realm, and topic for all tweets and behavior instances, $\mathcal{C}$, $\mathcal{R}$, and $\mathcal{Z}$ are determined. Hence, for any tweet $t_u^j$, when sampling the tweet's source $c_u^j$ and realm $r_u^j$, we are given $\mathcal{O}_{-t_u^j}$ and the tweet's topic $z_u^j$. Similarly, when sampling $z_u^j$, we are given $\mathcal{O}_{-t_u^j}$ and $c_u^j$, as well as $r_u^j$ (if it exists). Thus, the source and the realm are jointly sampled according to equations in Figure 4, while the topic is sampled according to equations in Figure 5. Note that, when $c_u^j = 0$, we do not have to sample $r_u^j$, and the current $r_u^j$ (if it exists) will be discarded.

In equations in Figures 4 and 5, $\mathbf{n_c}(c, u, \mathcal{C})$ records the number of times that the source $c$ is observed in the set of tweets and behavior instances of user $u$ for the bag-of-sources $\mathcal{C}$. Similarly, $\mathbf{n_{zu}}(z, u, \mathcal{Z})$ records the number of times that the topic $z$ is observed in the set of tweets and the bag of behavior instances of user $u$ for the bag of topics $\mathcal{Z}$. $\mathbf{n_{zr}}(z, r, \mathcal{Z}, \mathcal{R})$ records the number of times that the topic $z$ is observed in the set of tweets and the bag-of-behavior instances that are tweeted/adopted based on the realm $r$ by any user for the bag-of-topics $\mathcal{Z}$ and the bag-of-realms $\mathcal{R}$. $\mathbf{n_{ru}}(r, u, \mathcal{R})$ records the number of times that the realm $r$ is observed in the set of tweets and the bag-of-behavior instances of user $u$. Last, $\mathbf{n_w}(w, z, \mathcal{T}, \mathcal{Z})$ records the number of times that the word $w$ is observed in the topic $z$ for the set of tweets $\mathcal{T}$ and the bag-of-topics $\mathcal{Z}$.

$$p(c_u^j = 0|\boldsymbol{\mathcal{O}}_{-t_u^j}, z_u^j) \propto \frac{\mathbf{n_c}(0, u, \boldsymbol{\mathcal{C}}_{-t_u^j}) + \rho_0}{\sum_{c=0}^{1}\left(\mathbf{n_c}(c, u, \boldsymbol{\mathcal{C}}_{-t_u^j}) + \rho_c\right)} \cdot \frac{\mathbf{n_{zu}}(z_u^j, u, \boldsymbol{\mathcal{Z}}_{-t_u^j}) + \alpha_{z_u^j}}{\sum_{k=1}^{K}\left(\mathbf{n_{zu}}(k, u, \boldsymbol{\mathcal{Z}}_{-t_u^j}) + \alpha_k\right)} \tag{1}$$

$$p(c_u^j = 1, r_u^j = r|\boldsymbol{\mathcal{O}}_{-t_u^j}, z_u^j) \propto \frac{\mathbf{n_c}(1, u, \boldsymbol{\mathcal{C}}_{-t_u^j}) + \rho_1}{\sum_{c=0}^{1}\left(\mathbf{n_c}(c, u, \boldsymbol{\mathcal{C}}_{-t_u^j}) + \rho_c\right)} \cdot \frac{\mathbf{n_{ru}}(r, u, \boldsymbol{\mathcal{R}}_{-t_u^j}) + \tau_r}{\sum_{r'=1}^{G}\left(\mathbf{n_{ru}}(r', u, \boldsymbol{\mathcal{R}}_{-t_u^j}) + \tau_{r'}\right)} \cdot \frac{\mathbf{n_{zr}}(z_u^j, r, \boldsymbol{\mathcal{Z}}_{-t_u^j}, \boldsymbol{\mathcal{R}}_{-t_u^j}) + \eta_{z_u^j}}{\sum_{k=1}^{K}\left(\mathbf{n_{zr}}(k, r, \boldsymbol{\mathcal{Z}}_{-t_u^j}, \boldsymbol{\mathcal{R}}_{-t_u^j}) + \eta_k\right)} \tag{2}$$

Fig. 4. Probabilities used in jointly sampling source and realm for tweet $t_u^j$ without regularization.

$$p(z_u^j = z|\boldsymbol{\mathcal{O}}_{-t_u^j}, c_u^j = 0) \propto \frac{\mathbf{n_{zu}}(z, u, \boldsymbol{\mathcal{Z}}_{-t_u^j}) + \alpha_z}{\sum_{k=1}^{K}\left(\mathbf{n_{zu}}(k, u, \boldsymbol{\mathcal{Z}}_{-t_u^j}) + \alpha_k\right)} \cdot \prod_{i=1}^{N_u^j} \frac{\mathbf{n_w}(w_u^{ji}, z, \boldsymbol{\mathcal{Z}}_{-t_u^j}) + \beta_{w_u^{ji}}}{\sum_{v=1}^{V_t}\left(\mathbf{n_w}(v, z, \boldsymbol{\mathcal{Z}}_{-t_u^j}) + \beta_v\right)} \tag{3}$$

$$p(z_u^j = z|\boldsymbol{\mathcal{O}}_{-t_u^j}, c_u^j = 1, r_u^j) \propto \frac{\mathbf{n_{zr}}(z, r_u^j, \boldsymbol{\mathcal{Z}}_{-t_u^j}, \boldsymbol{\mathcal{R}}_{-t_u^j}) + \eta_z}{\sum_{k=1}^{K}\left(\mathbf{n_{zr}}(k, r_u^j, \boldsymbol{\mathcal{Z}}_{-t_u^j}, \boldsymbol{\mathcal{R}}_{-t_u^j}) + \eta_k\right)} \cdot \prod_{i=1}^{N_u^j} \frac{\mathbf{n_w}(w_u^{ji}, z, \boldsymbol{\mathcal{T}}_{-t_u^j}, \boldsymbol{\mathcal{Z}}_{-t_u^j}) + \beta_{w_u^{ji}}}{\sum_{v=1}^{V_t}\left(\mathbf{n_w}(v, z, \boldsymbol{\mathcal{T}}_{-t_u^j}, \boldsymbol{\mathcal{Z}}_{-t_u^j}) + \beta_v\right)} \tag{4}$$

Fig. 5. Probabilities used in sampling topic for tweet $t_u^j$ without regularization.

In the right-hand side of Equation (1): (i) the first term is proportional to the probability that the source 0 is generated given the priors and (current) values of all other latent variables (i.e., the sources, realms (if they exist), and topics of all other tweets and behavior instances); and (ii) the second term is proportional to the probability that the (current) topic $z_u^j$ is generated given the priors, (current) values of all other latent variables, and the chosen source.

Similarly, in the right-hand side of Equation (2): (i) the first term is proportional to the probability that the source 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 realm $r$ is generated given the priors, (current) values of all other latent variables, and the chosen source; and (iii) the third term is proportional to the probability that the (current) topic $z_u^j$ is generated given the priors, (current) values of all other latent variables, and the chosen source as well as the chosen realm.

In the right-hand side of Equation (3): (i) the first term is proportional to the probability that the topic $z$ is generated given the priors and (current) values of all other latent variables, and the corresponding source is 0; and (ii) the second term is proportional to the probability that the tweet content is generated given the priors, (current) values of all other latent variables, and the chosen topic.

Last, in the right-hand side of Equation (4): (i) the first term is proportional to the probability that the topic $z$ is generated given the priors and (current) values of all other latent variables, and the corresponding source is 1; (ii) the second term is proportional

$$p(c_u^{lj}=0|\boldsymbol{\mathcal{O}}_{-b_u^{lj}}, z_u^{lj}) \propto \frac{\mathbf{n_c}(0,u,\boldsymbol{\mathcal{C}}_{-b_u^{lj}})+\rho_0}{\sum_{c=0}^{1}\left(\mathbf{n_c}(c,u,\boldsymbol{\mathcal{C}}_{-b_u^{lj}})+\rho_c\right)} \cdot \frac{\mathbf{n_{zu}}(z_u^{lj},u,\boldsymbol{\mathcal{Z}}_{-b_u^{lj}})+\alpha_{z_u^{lj}}}{\sum_{k=1}^{K}\left(\mathbf{n_{zu}}(k,u,\boldsymbol{\mathcal{Z}}_{-b_u^{lj}})+\alpha_k\right)} \tag{5}$$

$$p(c_u^{lj}=1, r_u^{lj}=r|\boldsymbol{\mathcal{O}}_{-b_u^{lj}}, z_u^{lj}) \propto \frac{\mathbf{n_c}(1,u,\boldsymbol{\mathcal{C}}_{-b_u^{lj}})+\rho_1}{\sum_{c=0}^{1}\left(\mathbf{n_c}(c,u,\boldsymbol{\mathcal{C}}_{-b_u^{lj}})+\rho_c\right)} \cdot \frac{\mathbf{n_{ru}}(r,u,\boldsymbol{\mathcal{R}}_{-b_u^{lj}})+\tau_g}{\sum_{r'=1}^{R}\left(\mathbf{n_{ru}}(r',u,\boldsymbol{\mathcal{R}}_{-b_u^{lj}})+\tau_{r'}\right)} \cdot \frac{\mathbf{n_{zr}}(z_u^{lj},r,\boldsymbol{\mathcal{Z}}_{-b_u^{lj}},\boldsymbol{\mathcal{R}}_{-b_u^{lj}})+\eta_{z_u^j}}{\sum_{k=1}^{K}\left(\mathbf{n_{zr}}(k,r,\boldsymbol{\mathcal{Z}}_{-b_u^{lj}},\boldsymbol{\mathcal{R}}_{-b_u^{lj}})+\eta_k\right)} \tag{6}$$

Fig. 6. Probabilities used in jointly sampling source and realm for behavior instance $b_j^{l,i}$ without regularization.

$$p(z_u^{lj}=z|\boldsymbol{\mathcal{O}}_{-b_u^{lj}}, c_u^{lj}=0) \propto \frac{\mathbf{n_{zu}}(z,u,\boldsymbol{\mathcal{Z}}_{-b_u^{lj}})+\alpha_z}{\sum_{k=1}^{K}\left(\mathbf{n_{zu}}(k,u,\boldsymbol{\mathcal{Z}}_{-b_u^{lj}})+\alpha_k\right)} \cdot \frac{\mathbf{n_b^l}(b_u^{lj},z,\boldsymbol{\mathcal{B}}_{-b_u^{lj}},\boldsymbol{\mathcal{Z}}_{-b_u^{lj}})+\gamma_{lb_u^{lj}}}{\sum_{b=1}^{V_b^l}\left(\mathbf{n_b^l}(b,z,\boldsymbol{\mathcal{B}},\boldsymbol{\mathcal{Z}}_{-b_u^{lj}})+\gamma_{lb}\right)} \tag{7}$$

$$p(z_u^{lj}=z|\boldsymbol{\mathcal{O}}_{-b_u^{lj}}, c_u^{lj}=1, r_u^{lj}) \propto \frac{\mathbf{n_{zr}}(z,r_u^{lj},\boldsymbol{\mathcal{Z}}_{-b_u^{lj}},\boldsymbol{\mathcal{R}}_{-b_u^{lj}})+\eta_z}{\sum_{k=1}^{K}\left(\mathbf{n_{zr}}(k,r_u^{lj},\boldsymbol{\mathcal{Z}}_{-b_u^{lj}},\boldsymbol{\mathcal{R}}_{-b_u^{lj}})+\eta_k\right)} \cdot \frac{\mathbf{n_b^l}(b_u^{lj},z,\boldsymbol{\mathcal{B}}_{-b_u^{lj}},\boldsymbol{\mathcal{Z}}_{-b_u^{lj}})+\gamma_{lb_u^{lj}}}{\sum_{b=1}^{V_b^l}\left(\mathbf{n_b^l}(b,z,\boldsymbol{\mathcal{B}},\boldsymbol{\mathcal{Z}}_{-b_u^{lj}})+\gamma_{lb}\right)} \tag{8}$$

Fig. 7. Probabilities used in sampling topic for behavior $b_u^{lj}$ without regularization.

to the probability that the tweet content is generated given the priors, (current) values of all other latent variables, and the chosen topic.

**Sampling for a behavior instance.** Similar to what has been described before, for any behavior instance $b_u^{lj}$, when sampling the instance's source $c_u^{lj}$ and realm $r_u^{lj}$, we are given $\boldsymbol{\mathcal{O}}_{-b_u^{lj}}$ and the instance's topic $z_u^{lj}$. Also, when sampling $z_u^{lj}$, we are given $\boldsymbol{\mathcal{O}}_{-b_u^{lj}}$, $c_u^{lj}$ and $r_u^{lj}$. Thus, the source and the realm are jointly sampled according to equations in Figure 6, while the topic is sampled according to equations in Figure 7. Again, note that, when $c_j^{i,l}=0$, we do not have to sample $r_j^{i,l}$, and the current $r_j^{i,l}$ (if it exists) will be discarded.

In equations in Figures 6 and 7, $\mathbf{n_b^l}(b,z,\boldsymbol{\mathcal{B}},\boldsymbol{\mathcal{Z}})$ records the number of times that the type-$l$ behavior $b$ is observed in the topic $z$ for the bag-of-behavior instances $\boldsymbol{\mathcal{B}}$ and the bag-of-topics $\boldsymbol{\mathcal{Z}}$. The terms in the right-hand side of Equations (5) to (8) have the same meaning as those of Equations (1) to (4), respectively.

### 3.6. Sparsity Regularization

As we want to differentiate users' tweets and behavior instances based on personal interest from those based on realms while distinguishing one realm from the others, we prefer (a) realms' topic distributions and users' topic distributions to skew on different topics, and (b) different realms' topic distributions to skew on different topics. More

$$\mathcal{R}_{\text{topicSource-Source\&Realm}}(c|z_u^j) = exp\Bigg( - \frac{\Big(H^{source}_{c_u^j=c}(z_u^j) - \mu_{\text{topicSource}}\Big)^2}{2\sigma^2_{\text{topicSource}}}\Bigg) \tag{9}$$

$$\mathcal{R}_{\text{topicSource-Source\&Realm}}(c|z_u^{lj}) = exp\Bigg( - \frac{\Big(H^{source}_{c_u^{lj}=c}(z_u^{lj}) - \mu_{\text{topicSource}}\Big)^2}{2\sigma^2_{\text{topicSource}}}\Bigg) \tag{10}$$

Fig. 8. Topic-specific source distribution regularization terms used in sampling source and/or realm for tweet $t_u^j$ and behavior instance $b_u^{lj}$.

$$\mathcal{R}_{\text{topicSource-Topic}}(z|t_u^j) = exp\Bigg( - \sum_{z'=1}^{K}\Bigg[\frac{\Big(H^{source}_{z_u^j=z}(z') - \mu_{\text{topicSource}}\Big)^2}{2\sigma^2_{\text{topicSource}}}\Bigg]\Bigg) \tag{11}$$

$$\mathcal{R}_{\text{topicSource-Topic}}(z|b_u^{lj}) = exp\Bigg( - \sum_{z'=1}^{K}\Bigg[\frac{\Big(H^{source}_{z_u^{lj}=z}(z') - \mu_{\text{topicSource}}\Big)^2}{2\sigma^2_{\text{topicSource}}}\Bigg]\Bigg) \tag{12}$$

Fig. 9. Topic-specific source distribution regularization terms used in sampling topic for tweet $t_u^j$ and behavior instance $b_u^{lj}$.

specifically, in estimating parameters in the GBT model, we need to obtain sparsity in the following distributions.

—Topic-specific source distribution $p^{source}(\cdot|z)$, where $z$ is a topic: The sparsity in this distribution is to ensure that each topic $z$ is mostly covered by either users' personal interest or realms.
—Topic-specific realm distribution $p^{realm}(\cdot|z)$, where $z$ is a topic: The sparsity in this distribution is to ensure that each topic $z$ is mostly covered by one or only a few realms.

To obtain this sparsity, we use the *pseudo-observed variable*-based regularization technique proposed by Balasubramanyan and Cohen [2013], as follows.

*3.6.1. Topic-Specific Source Distribution Regularization.* Since the topic-specific source distributions are determined by both source and realm joint sampling and topic-sampling steps, we regularize both these steps to bias the distributions to some target sparsity.

**In source & realm joint sampling steps.** In each source & realm sampling step for the tweet $t_u^j$, we multiply the right-hand side of equations in Figure 4 with a corresponding regularization term $\mathcal{R}_{\text{topicSource-Source\&Realm}}(c|z_u^j)$, which is computed based on empirical entropy of $p(c|z_u^j)$, as in Equation (9). Similarly, in each source & realm sampling step for the behavior instance $b_j^{i,j}$, we multiply the right-hand side of the equations in Figure 6 with a corresponding regularization term $\mathcal{R}_{\text{topicSource-Source\&Realm}}(c|z_u^{lj})$, which is computed based on empirical entropy of $p(c|z_u^{lj})$, as in Equation (10).

**In topic-sampling steps.** In each topic-sampling step for the tweet $t_u^j$, we multiply the right-hand side of equations in Figure 5 with a corresponding regularization term $\mathcal{R}_{\text{topicSource-Topic}}(z|t_u^j)$, which is computed based on empirical entropy of $p(c|z)$, as in Equation (15). Similarly, in each topic-sampling step for the behavior instance $b_j^{i,j}$, we multiply the right-hand side of equations in Figure 7 with a corresponding

$$\mathcal{R}_{\text{topicRealm-Source\&Realm}}(c, r|z_u^j) = exp\left( - \frac{\left( H^{realm}_{c_u^j=c, r_u^j=r}(z_u^j) - \mu_{\text{topicRealm}} \right)^2}{2\sigma^2_{\text{topicRealm}}} \right) \quad (13)$$

$$\mathcal{R}_{\text{topicRealm-Source\&Realm}}(c, r|z_u^{lj}) = exp\left( - \frac{\left( H^{realm}_{c_u^{lj}=c, r_u^{lj}=r}(z_u^{lj}) - \mu_{\text{topicRealm}} \right)^2}{2\sigma^2_{\text{topicRealm}}} \right) \quad (14)$$

Fig. 10. Topic specific realm distribution regularization terms used in sampling source and/or realm for tweet $t_u^j$ and behavior instance $b_u^{lj}$

regularization term $\mathcal{R}_{\text{topicSource-Topic}}(z|b_u^{lj})$, which is computed based on empirical entropy of $p(c|z)$, as in equations in Figure 12.

In Equation (9), $H^{source}_{c_u^j=c}(z_u^j)$ is the empirical entropy of $p^{source}(\cdot|z_u^j)$ when $c_u^j = c$; in Equation (10), $H^{source}_{c_u^{lj}=c}(z_u^{lj})$ is the empirical entropy of $p^{source}(\cdot|z_u^{lj})$ when $c_u^{lj} = c$. Similarly, for each topic $z'$, in Equation (11), $H^{source}_{z_u^j=z}(z')$ is the empirical entropy of $p^{source}(\cdot|z')$ when $z_u^j = z$; in Equation (12), $H^{source}_{z_u^{lj}=z}(z')$ is the empirical entropy of $p^{source}(\cdot|z')$ when $z_u^{lj} = z$. The two parameters $\mu_{\text{topicSource}}$ and $\sigma_{\text{topicSource}}$ are the target mean and target variance of the entropy of $p(c|z)$, respectively. These target mean and target variances are predefined parameters. Obviously, these regularization terms (1) increase weight for values of $c$, $r$, and $z$ that give lower empirical entropy of $p(c|z)$, thus increasing the sparsity of these distributions; but (2) decrease weight for values of $c$, $r$, and $z$, which give higher empirical entropy of $p(c|z)$, thus decreasing the sparsity of these distributions.

*3.6.2. Topic-Specific Realm Distribution Regularization.* Similarly, since the topic-specific realm distributions are determined by both source & realm joint sampling and topic sampling steps, we regularize both these steps to bias the distributions to some target sparsity.

**In source & realm joint sampling steps.** In each source & realm sampling step for the tweet $t_u^j$, we also multiply the right-hand side of the equations in Figure 4 with a corresponding regularization term $\mathcal{R}_{\text{topicRealm-Source\&Realm}}(c, r|z_u^j)$, which is computed based on the empirical entropy of $p(r'|z_u^j)$, as in Equation (13). Similarly, in each source & realm sampling step for the behavior instance $b_j^{i,j}$, we also multiply the right-hand side of the equations in Figure 6 with a corresponding regularization term $\mathcal{R}_{\text{topicRealm-Source\&Realm}}(c, r|z_u^{lj})$, which is computed based on the empirical entropy of $p(r'|z_u^{lj})$, as in Equation (14).

**In topic sampling steps.** In each topic sampling step for the tweet $t_u^j$, we also multiply the right hand side of equations in Figure 5 with a corresponding regularization term $\mathcal{R}_{\text{topicRealm-Topic}}(z|t_u^j)$ which is computed based on empirical entropy of $p(r|z)$ as in Equation (15). Similarly, in each topic sampling step for the behavior instance $b_j^{i,j}$, we multiply the right hand side of equations in Figure 7 with a corresponding regularization term $\mathcal{R}_{\text{topicReaml-Topic}}(z|b_u^{lj})$ which is computed based on empirical entropy of $p(c|z)$ as in equations in Figure 16.

In Equation (13), $H^{realm}_{c_u^j=c, r_u^j=r}(z_u^j)$ is the empirical entropy of $p^{realm}(\cdot|z_u^j)$ when $c_u^j = c$ & $r_u^j = r$; in Equation (14), $H^{realm}_{c_u^{lj}=c, r_u^{lj}=r}(z_u^{lj})$ is the empirical entropy of $p^{realm}(\cdot|z_u^{lj})$ when

$$\mathcal{R}_{\text{topicRealm-Topic}}(z|t_u^j) = exp\Bigg( - \sum_{z'=1}^{K} \Bigg[ \frac{\Big( H^{realm}_{z_u^j=z}(z') - \mu_{\text{topicRealm}} \Big)^2}{2\sigma^2_{\text{topicRealm}}} \Bigg] \Bigg) \tag{15}$$

$$\mathcal{R}_{\text{topicRealm-Topic}}(z|b_u^{lj}) = exp\Bigg( - \sum_{z'=1}^{K} \Bigg[ \frac{\Big( H^{realm}_{z_u^{lj}=z}(z') - \mu_{\text{topicRealm}} \Big)^2}{2\sigma^2_{\text{topicRealm}}} \Bigg] \Bigg) \tag{16}$$

Fig. 11. Topic-specific realm distribution regularization terms used in sampling topic for tweet $t_u^j$ and behavior instance $b_u^{lj}$.

$c_u^{lj} = c$ & $r_u^{lj} = r$. Similarly, for each topic $z'$, in Equation (15), $H^{realm}_{z_u^j=z}(z')$ is the empirical entropy of $p^{realm}(\cdot|z')$ when $z_u^j = z$; in Equation (16), $H^{realm}_{z_u^{lj}=z}(z')$ is the empirical entropy of $p^{realm}(\cdot|z')$ when $z_u^{lj} = z$. The two parameters $\mu_{\text{topicRealm}}$ and $\sigma_{\text{topicRealm}}$ are the target mean and target variance of the entropy of $p(r|z)$, respectively. These target mean and target variances are predefined parameters. Obviously, these regularization terms (1) increase weight for values of $c$, $r$, and $z$ that give lower empirical entropy of $p(r|z)$, thus increasing the sparsity of these distributions; but (2) decrease weight for values of $c$, $r$, and $z$, which give higher empirical entropy of $p(r|z)$, thus decreasing the sparsity of these distributions.

### 3.7. Implementation and Complexity

We use two-dimensional tables for keeping the counts $\mathbf{n_c}(c, u, \mathcal{C})$, $\mathbf{n_{zu}}(z, u, \mathcal{Z})$, $\mathbf{n_{zr}}(z, r, \mathcal{Z}, \mathcal{R})$, $\mathbf{n_w}(w, z, \mathcal{T}, \mathcal{Z})$, and $\mathbf{n_b^l}(b, z, \mathcal{B}, \mathcal{Z})$ and call them counting tables. We use one-dimensional tables for keeping row and column sums of the counting tables and call them sum tables. Also, we use one-dimensional tables for keeping the empirical entropies of $p(c|z)$ and $p(r|z)$, and call them entropy tables.

In each sampling step, only constant time updates on some counting table(s) and sum table(s) are made. For each topic $z$, the empirical entropies of $p(c|z)$ and $p(r|z)$ are computed based on the row/column $z$ of one of the counting tables. Thus, in each sampling step, the entropy tables can also be updated in constant time as follows. Let $E_{\text{current}}$ be the current empirical entropy of $p^{realm}(\cdot|z)$. $E_{\text{current}}$ is computed from the array $n_1, \ldots, n_R$, which is the row/column $z$ of one of the counting tables, that is,

$$E_{\text{current}} = -\sum_{r=1}^{R} \frac{n_r}{\sum_{r=1}^{R} n_r} \log\left( \frac{n_r}{\sum_{r=1}^{R} n_r} \right).$$

Now, assume that $n_g$ is changed to $n_g + \Delta$; then, the new empirical entropy $E_{\text{new}}$ of $p(r|z)$ can be computed from $E_{\text{current}}$ as follows:

$$\begin{aligned} E_{\text{new}} \;=\; & \frac{1}{\Delta + \sum_{r=1}^{R} n_r} \Bigg[ E_{\text{current}} \sum_{r=1}^{R} n_r + \big( n_g \log(n_g) - (n_g + \Delta)\log(n_g + \Delta) \big) \\ & + \log\left( \Delta + \sum_{r=1}^{R} n_r \right) \left( \Delta + \sum_{r=1}^{R} n_r \right) - \left( \sum_{r=1}^{R} n_r \right) \log\left( \sum_{r=1}^{R} n_r \right) \Bigg]. \end{aligned}$$

Given that the sum $\sum_{r=1}^{R} n_r$ is kept in a cell of one of the sum tables, the cost of updating the empirical entropy $p^{realm}(\cdot|z)$ is therefore constant. Similarly, in each sampling step,

we can update any entropy table in constant time. Thus, in total, a single iteration of the collapsed Gibbs sampler performs $\mathcal{O}((|\mathcal{W}| + |\mathcal{B}|) \times (K + R))$ computationsm where $|\mathcal{W}|$ is the number of observed words and $|\mathcal{B}|$ is the number of observed behavior instances in the dataset [Heinrich 2009]. We provide a JAVA implementation of the GBT model at https://github.com/smutahoang/ttm.

In our experiments, we used a sampling method with the sparsity regularization presented earlier, setting $\mu_{\text{topicSource}} = \mu_{\text{topicRealm}} = 0$, $\sigma_{\text{topicSource}} = 0.3$, and $\sigma_{\text{topicRealm}} = 0.5$. This corresponds to the case in which every topic is assigned to either realms or users' personal interests, and every topic is also assigned to at most one realm. $\sigma_{\text{topicSource}}$ is set smaller than $\sigma_{\text{topicRealm}}$ so that, for each topic, the topic's source distribution is more strictly regularized than its realm distribution. We also used conventional symmetric Dirichlet hyperparameters, which are used in previous works (e.g., Blei et al. [2003], Zhao et al. [2011], and Qiu et al. [2013]). That is, $\alpha = 50/K$, $\beta = 0.01$, $\rho = 2$, $\tau = 1/R$, $\eta = 50/K$, and $\gamma_l = 0.01$ for all $l = 1, \ldots, L$. 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

**Data collection.** In order to evaluate the GBT model properly, we need experimental datasets to be domain specific, and have full content and behavior of users over a long period of time. We have specially selected two domains, software engineering and politics, in which there are realms with distinctive topic distributions. This further helps us to empirically evaluate the results. The second requirement is necessary so that we can learn the model accurately. It cannot be easily met by simply collecting data from sampled tweet streams, however, which offer a small proportion of all Twitter data. We instead use snowball a sampling method to collect data. Given a domain, we first manually select a set of *seed users* who are experts in the domain. We then expand the set by adding the seed users' followers and/or followees. Last, we crawl the content and behavior of users in the expanded set.

Based on this approach, we constructed the following two datasets:

—**SE Dataset.** This dataset contains tweets and behavior of a large set of Twitter users who are interested in software engineering. To construct this dataset, we first utilized 100 of the most influential software developers on Twitter provided in Jurgen [2009] as the seed users. These are highly followed users who actively tweet about software engineering topics, including *Jeff Atwood*[8], *Jason Fried*[9], and *John Resig*[10]. We further expanded the user set by adding all users following at least five seed users to get more technology-savvy users. Last, we took all tweets posted by these users from August 1 to October 31, 2011 to form the first dataset, called the SE dataset.

—**Two-Week Dataset.** This dataset is in the politics domain, and was collected from Twitter 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, politically oriented Twitter users, including major US politicians, for example, Barack Obama, Mitt Romney, and Newt Gingrich; well-known political bloggers, for example, America Blog, Red State, and Daily Kos; and political sections of the US news media, for example, CNN Politics, and Huffington Post Politics. The set of users was then expanded by adding all users following at least three seed users to get more

[8]http://en.wikipedia.org/wiki/Jeff_Atwood.
[9]http://www.hanselman.com/blog/AboutMe.aspx.
[10]http://en.wikipedia.org/wiki/John_Resig.

Table II. Statistics of the Experimental Datasets

| | | SE dataset | Two-Week dataset |
|---|---|---|---|
| #user | | 14,595 | 24,046 |
| #tweets | | 3,030,734 | 3,181,583 |
| #behavior instances | user mention | 354,463 (with 2,337 adopters) | 653,758 (with 4,628 adopters) |
| | hashtag usage | 894,619 (with 3,992 adopters) | 1,820,824 (with 9,288 adopters) |
| | retweeting | 909,272 (with 5,324 adopters) | 2,396,100 (with 10,576 adopters) |

politics-savvy users. Last, we used all the tweets posted by these users during the 2 weeks from August 25 to September 7, 2012 to form the second dataset, known as the Two-Week dataset.

**Data preprocessing.** We employed the following preprocessing steps to clean both datasets.

—**Tweet selection.** We first removed stopwords from the tweets. Then, we filtered out tweets with less than 3 nonstopwords. Next, we excluded users with less than 50 (remaining) tweets to focus on users with sufficient data.

—**Behavior instance selection.** In both datasets, we consider instances of following-behavior types: (1) *user mention*, (2) *hashtag usage*, and (3) *retweeting*. These are behavior instances beyond content generation that users may adopt multiple times. More precisely, for each time that user $u$ mentions user $v$ in user $u$'s tweets, we consider $v$ as a behavior instance of user mention type of $u$. Similarly, for each time that user $u$ retweets a tweet originally posted by $v$, we also consider $v$ as a behavior instance of retweeting type of $u$. Last, for each time that user $u$ uses hashtag $h$ in user $u$'s tweets, we consider $h$ as a behavior instance of hashtag usage type of $u$.

Similar to tweets' words, for each behavior instance, we filtered away those with less than 10 adopting users. Also, for each user $u$ and each behavior type, we filtered out all of $u$'s behavior instances of the type if $u$ adopted less than 50 instances of the type. These minimum thresholds are necessary so that, for each behavior instance and each user, we have enough adoption observations for learning both the influence of the user's personal interest and that of the realms on the instance's adoption.

Table II shows the statistics of the two datasets after the preprocessing steps. As shown in the table, the two datasets after the filtering are still large. In the SE dataset, there are about 200 tweets, 150 user-mention instances, 225 hashtag-usage instances, and 170 retweeting instances per user. In the Two-Week dataset, there are about 120 tweets, 140 user-mention instances, 195 hashtag-usage instances, and 225 retweeting instances per user. This large size allows us to learn the latent factors accurately.

### 4.2. Content Modeling

We first evaluate the ability of the GBT model in modeling topics of content. To do this, we compare the GBT model with two state-of-the-art topic models for Twitter data: the TwitterLDA model [Zhao et al. 2011], and QBLDA model [Qiu et al. 2013]. We briefly reviewed these models in Section 2.

**Evaluation metrics.** We adopt *likelihood* and *perplexity* for evaluating the resultant topics. For each user, we randomly selected 90% of tweets of the user to form a training set, and the remaining 10% of the tweets as the test set. Then, 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.

**Performance comparison.** Figures 12(a) and (b) show the performance of the TwitterLDA, QBLDA, and GBT models in content modeling on the SE dataset by varying the number of topics $K$ and the number of realms $R$. Figures 12(c) and (d) show

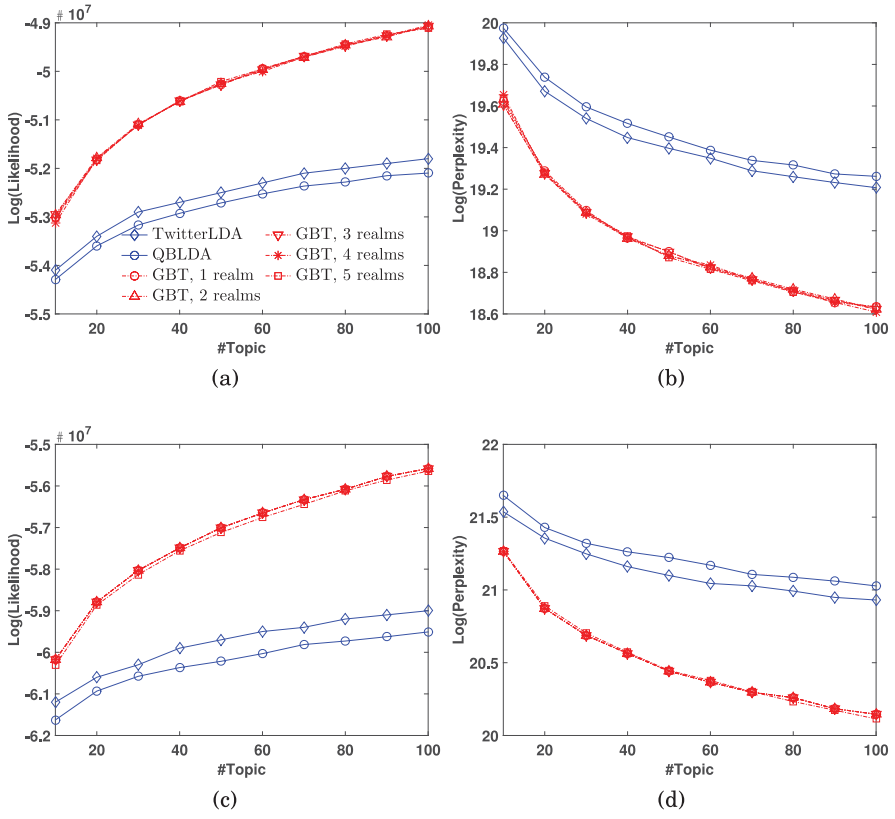



Fig. 12. Loglikelihood and perplexity of different models in: (a) and (b) SE dataset, and (c) and (d) Two-Week dataset.

the similar results on the Two-Week dataset. As expected, a larger number of topics $K$ results in a larger likelihood and smaller perplexity, and the amount of improvement diminishes as $K$ increases. The figures show that: (1) the GBT model significantly outperforms both the TwitterLDA and QBLDA models in the content modeling task; and (2), as we expected, GBT's likelihood and perplexity do not significantly change as we increase the number of the realms from 1 to 5.

**Setting the numbers of topics and realms.** We further look into the realms returned by the **GBT** model with a different number of realms and found that there is a semantically hierarchical structure among the realms. That is, when the number of realms is increased, the realms are divided into more semantically distinctive realms. For example, Figure 13 shows the top topics of the realm(s) found in the SE dataset when the number of realms varied from 1 to 3. Here, the labels of the topics are manually assigned after examining the topics' top words and top tweets. For each topic, the topic's top words are the words having the highest probabilities given the topic, and the topic's top tweets are the tweets having the lowest perplexities given the topic. The figure clearly shows that the unique realm in the case $R = 1$ is divided into two semantically clearer realms when $R = 2$. These two realms divided into three realms with even clearer semantics when $R = 3$. We also have similar qualitative findings from the Two-Week dataset. This suggests that the GBT model can recover the more detailed realms by increasing the number of realms, even though the quantitative performance does not significantly improve.

**1 Realm**

| | Unique realm | | |
|---|---|---|---|
| **Topic Id** | **41** | **67** | **52** |
| **Topic label** | **Daily stuffs** | **Programming** | **Smart devices** |
| **Probability** | **0.510** | **0.085** | **0.068** |

**2 Realms**

| | Realm 0 | | | Realm 1 | | |
|---|---|---|---|---|---|---|
| **Topic Id** | **4** | **7** | **28** | **67** | **14** | **34** |
| **Topic label** | **Daily works** | **Children** | **Networking services** | **Programming** | **Operating systems** | **Project management** |
| **Probability** | **0.297** | **0.291** | **0.126** | **0.216** | **0.213** | **0.191** |

**3 Realms**

| | Realm 0 | | | Realm 1 | | | Realm 2 | | |
|---|---|---|---|---|---|---|---|---|---|
| **Topic Id** | **44** | **66** | **26** | **38** | **22** | **66** | **76** | **43** | **26** |
| **Topic label** | **Scripting programming languages** | **Email & social networking services** | **Readings** | **iOS** | **iPhone & iPad** | **Email & social networking services** | **Daily stuffs** | **Foods & drinks** | **Readings** |
| **Probability** | **0.760** | **0.044** | **0.043** | **0.369** | **0.231** | **0.102** | **0.536** | **0.098** | **0.089** |

Fig. 13. Top topics of **realm(s)** found in SE dataset when the number of realms varies from 1 to 3.

Table III. Top Words of Background Topic Found in SE Dataset by TwitterLDA and QBLDA Models

| Model | Top words of background topic |
|---|---|
| TwitterLDA | life,making,video,blog,change,reading,job,home,thought,line team,power,game,business,money,friends,talking,starting,month,company |
| QBLDA | video,life,blog,change,job,game,reading,business,power,making thought,line,home,#fb,giving,friends,team,money,talking,running |

Table IV. Top Topics of Realms Found in SE Dataset

| Realm Id | Realm Label | Top topics | | |
|---|---|---|---|---|
| | | Topic Id | Topic Label | Probability |
| 0 | Software development | 44 | Scripting programming languages | 0.760 |
| | | 66 | Email & social networking services | 0.044 |
| | | 26 | Readings | 0.043 |
| 1 | Apple's products | 38 | iOS | 0.369 |
| | | 22 | iPhone & iPad | 0.231 |
| | | 66 | Email & social networking services | 0.102 |
| 2 | Daily life | 76 | Daily stuffs | 0.536 |
| | | 43 | Foods & drinks | 0.098 |
| | | 26 | Readings | 0.089 |

Considering both time and space complexities, and that it is not practical to expect a large number of topics falling in realm(s), we set the number of topics to 80 and set the number of the realms to 3 for the experiments presented in the following sections.

### 4.3. Background Topics & Realms Analysis

We now examine the background topics found by the TwitterLDA and QBLDA models, and realms found by the GBT model.

Table III shows the top words of the background topics found by the TwitterLDA model and QBLDA model in the SE dataset, while Table IV shows the top topics for each realm found in the same dataset. Note that, other than the background topics in the TwitterLDA and QBLDA models, the labels of other topics are also manually assigned after examining the topics' top words (shown in Table X) and top tweets. The label of each realm is also manually assigned after examining the realm's top topics. The tables show that the background topics found by the TwitterLDA and QBLDA models are not semantically clear, while the realms and their extreme topics found by the GBT model are both semantically clear and reasonable. In the SE dataset,

Table V. Top Words of Background Topic Found in Two-Week Dataset by TwitterLDA and QBLDA Models

| Model | Top words of background topic |
|---|---|
| TwitterLDA | life,making,home,america,called,house,change,thought,video,talking line,american,money,country,job,obama,friends,fact,lost,hell |
| QBLDA | video,making,american,called,obama,america,talking,thought,house,country president,job,line,giving,home,life,lost,fact,#dnc2012,change |

Table VI. Top Topics of Realms Found in Two-Week Dataset

| Realm Id | Realm Label | Top topics | | |
|---|---|---|---|---|
| | | Topic Id | Topic Label | Probability |
| 0 | Responses to DNC & RNC 2012 | 5 | Responses to speeches at DNC 2012 | 0.624 |
| | | 17 | Clint Eastwood's empty chair[13] | 0.105 |
| | | 28 | Economics issues | 0.072 |
| 1 | Republicans opposing | 8 | Criticizing Obama | 0.347 |
| | | 65 | Government & people's rights | 0.138 |
| | | 3 | Criticizing Chris Matthews's comments on Republicans | 0.098 |
| 2 | DNC & RNC 2012 | 31 | Speeches at RNC 2012 | 0.353 |
| | | 54 | Media reports on DNC & RNC 2012 | 0.174 |
| | | 77 | Speeches at DNC 2012 | 0.152 |

other than the *Daily Life* realm, as reported in Java et al. [2007], it is expected that professional realms *Software Development* and *Apple's product* exist in the dataset as most of its users are working in the IT industry. This agrees with the findings by Zhao and Rosson [2009], that people also use Twitter for gathering and sharing useful information for their profession.

Similarly, Table V shows the top words of the background topics found by the TwitterLDA and QBLDA models in the Two-Week dataset, while Table VI shows the top topics for each realm found in the same dataset. Again, the topics' labels are manually assigned based on examining the topics' top words (shown in Table XV) and top tweets; the realms' labels are also manually assigned based on examining the realm's top topics. Also, the tables show that the background topics found by the TwitterLDA and QBLDA models are not semantically clear, while the realms and their extreme topics found by the GBT model are both semantically clear and reasonable. In the Two-Week dataset, it is expected that political realms *Responses to DNC & RNC 2012*, *Republicans Opposing*, and *DNC and RNC 2012* exist in the dataset, as it was collected during the 2012 US presidential election, including the national conventions of both the Democratic[11] and Republican[12] parties.

In summary, the empirical content analysis results look reasonable when our proposed GBT model is applied on the two datasets. We now turn our focus to behavior modeling results.

### 4.4. User Behavior Analysis

Last, we examine the user behavior instances associated with the result topics. Tables X and XV show some of representative topics found in the SE and Two-Week datasets, respectively, together with the topics' top behavior instances. For each topic, and each behavior type, similar to the topic's top words, the topic's top behavior instances are the instances having the highest probabilities given the topic. The tables show that the key

[11]http://en.wikipedia.org/wiki/2012_Democratic_National_Convention.
[12]http://en.wikipedia.org/wiki/2012_Republican_National_Convention.
[13]http://en.wikipedia.org/wiki/Clint_Eastwood_at_the_2012_Republican_National_Convention.

behavior instances for each topic are reasonable. For example, in the SE dataset, we observe for topic *Scripting programming languages* (topic 44) that people use scripting languages–related hashtags (#*javascript*, #*ruby*, #*nodejs*, #*php*, and so on), mention and retweet from software-project hosting services and scripting language builder & developers (@*github*, @*heroku*, @*rubyrogues*, @*steveklabnik*, *garybernhardt*, *tenderlove*, *dhh*, and so on). We also observe for topic *iPhone & iPad* (topic 22) that people use iPhone- and iPad-related hashtags (#*iphone*, #*iphone5*, #*apple*, and so on), mention big IT companies and phone and tablet producers (*branch*, @*twitter*, @*google*, @*amazon*, @*att*, and), and retweet from iOS developers and IT bloggers (*marcoarment*, *John Gruber*, *dcurtis*).

Similarly, in the Two-Week dataset, we observe for topic *Responses to DNC & RNC 2012* (topic 5) that people use DNC & RNC 2012–related hashtags (#*dnc2012*, #*rnc2012*, #*literally*, and so on), mention key persons in the two conventions (e.g., *dwstweets*, *stefcutter*, *reince*, and so on), and retweet from political bloggers and commentators (*guypbenson*, *jimgeraghty*, *iowahawkblog*, *jonahnro*, and so on). We also observe for topic *Criticizing Obama* (topic 8) that people use negative hashtags related to Obama and DNC 2012 (#*dncin4words*, #*howtopissoffademocrat*, #*overheardatdnc2012*, #*obamatvshows*, and so on), and mention and retweet from Republican politicians and media (e.g., @*jjauthor*, @*klsouth*, *slone*, *polarcoug*, and so on). A qualitatively similar result holds for the remaining topics as well as topics that are not shown in the two tables.

On the whole, the user behavior analysis results are pretty consistent with that of content analysis. Now that the topics learned by the GBT model are reasonable, they can be used in the user-profiling experiments.

## 5. UTILITY OF USER TOPICS IN USER-PROFILING TASKS

In this section, we compare and contrast topics and users' personal topical interests uncovered by the GBT model with those uncovered by the TwitterLDA and LDA models in some user-profiling tasks for Twitter. Our aim here is not to propose any new user-profiling models. Instead, we want to evaluate the utility of different topic models in the user-profiling tasks that differentiate users with different user labels. Here, the user labels are the professional and political preferences of the users. Since the background topics are shared by users of all classes, they are the least discriminative topics. Thus, a model better at modeling users' personal topics and identifying the background topics would result in a better performance in the user-profiling tasks.

### 5.1. Profiling Tasks

We consider the following tasks.

—**User clustering.** In this task, we use the K-mean method with Euclidean similarity to cluster a set of users.
—**User classification.** In this task, we use the SVM method with linear kernel to classify a set of users into classes corresponding to different user labels.

### 5.2. User Representation

We represent each user by the user's topic distribution(s) learned from the user's content and behavior using a topic model. More precisely, for each model, each topic is a feature to represent users, and the feature vector of a user is the user's topic distribution(s) learned by the model. We examine the following topic models.

—**TwitterLDA.** In this model, each user $u$ is represented by $\theta_u^{TwitterLDA}$, where $\theta_u^{TwitterLDA}$ is the topic distribution of $u$ learned by the TwitterLDA model. That

means that each user is represented by personal interests learned from that user's content only.

—**QBLDA.** In this model, each user $u$ is represented by $\theta_u^{QBLDA}$, where $\theta_u^{QBLDA}$ is the topic distribution of $u$ learned by the QBLDA model. Each user is represented by personal interests learned from that user's content and user behavior types associated with the content (see Section 2).

—**TwitterLDA+behaviorLDA.** In this model, we consider both the user's personal interest learned from the user's content and the user's personal interests independently learned from the user's behavior. That means that each user $u$ is represented by a vector feature $\vec{u}$, where $\vec{u}$ is formed by concatenating $\theta_u^{TwitterLDA}$ and $\theta_u^1, \ldots, \theta_u^L$. Here, for every behavior type $l$ ($l = 1, \ldots, L$), if $u$ has type-$l$ behavior instances, $\theta_u^l$ is the topic distribution of $u$ learned by applying LDA [Blei et al. 2003] on the bags-of-behavior instances of type $l$ of all the adopting users. Otherwise, $\theta_u^l$ is a zero vector. We suppose that adding latent factors learned from behavior to the TwitterLDA model will improve performance in user-profiling tasks.

—**GBT-noBehavior.** For this model, we represent each user $u$ by $\theta_u^{GBT-noBehavior}$, where $\theta_u^{GBT-noBehavior}$ is the topic distribution of $u$ learned by running the GBT model only on the dataset, excluding user behavior. With this model, we want to evaluate the effectiveness of user behavior in profiling a user.

—**GBT-noRegularization.** For this model, we represent each user $u$ by $\theta_u^{GBT-noRegularization}$ where $\theta_u^{GBT-noRegularization}$ is the topic distribution of $u$ learnt by running the GBT model on the full dataset (both user content and user behavior), but without any sparsity regularization. With this model, we want to evaluate the effectiveness of the proposed sparsity regularization technique in learning clearer user interests.

—**GBT.** For this model, we represent each user $u$ by $\theta_u^{GBT}$, where $\theta_u^{GBT}$ is topic distribution of $u$ learned by running the GBT model on the full dataset and with the regularization technique used. We expect the GBT model to outperform all previous models. This improvement is attributed to joint modeling of user interest from both user content and user behavior and more accurate measuring of users' personal interests after filtering out their dependency on realms.

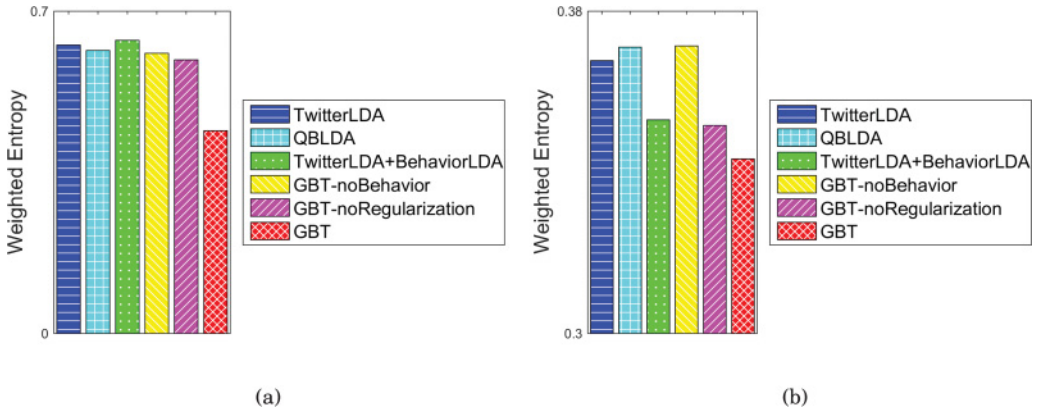
Similar to the previous experiments, in all these models, we set the number of topics to 80; in GBT-noBehavior, GBT-noRegularization and GBT models, we set the number of realms to 3.

### 5.3. Experimental Datasets

To evaluate the performance of these topic models in user-profiling tasks, we need some datasets with ground-truth labels for all users. Since we do not have ground-truth labels for all users in SE and Two-Week datasets, we derived the following (sub)datasets.

—**Developer** dataset: From the users' *self-descriptions*, we were able to manually label 691 users in the SE dataset as developers. Among these users, 328 users declare .NET-based programming languages (e.g., C#, Visual Basic) as their preferential languages, and 363 users declare other languages (e.g., Java, PhP, Python). We respectively denote the label for the former and latter set of these users by **.NET** and **non-.NET**. Then, for the clustering task, we cluster the developers into two clusters. For the classification task, we performed a binary classification.

—**Political affiliation** dataset: Similarly, from users' *self-descriptions*, we were able to manually label 186 users in the Two-Week dataset as Democrat and 1288 users as Republican. Again, for the clustering task, we cluster these manually labeled users into two clusters; for the classification task, we also performed a binary classification.




(a) (b)

Fig. 14. Performance of different models in user clustering task in (a) the Developer dataset and (b) the Political affiliation dataset.

### 5.4. Evaluation Metrics

For convenience, in Developer dataset, we call .NET user label 1 and non-.NET user label 2. Also, in the Political affiliation dataset, we call Democrats user label 1 and Republicans user label 2.

For the user-clustering task, we adopt *weighted entropy* as the performance metric. After running the K-means method with the number of clusters set to 2, we computed the weighted entropy of the resultant clusters as follows:

$$E = -\sum_{c=1}^{2} \frac{n_c}{n} * \left[ \frac{n_c^1}{n_c} * log \frac{n_c^1}{n_c} + \frac{n_c^2}{n_c} * log \frac{n_c^2}{n_c} \right], \tag{17}$$

where $n_c$ is the number of users assigned to cluster $c$ and $n$ is total number of users. $n_c^1$ and $n_c^2$ are, respectively, the number of users having user label 1 and user label 2 that are assigned to clustering $c$, that is, $n_c = n_c^1 + n_c^2$. The model with a lower entropy is the winner in the task.

For the user-classification task, we adopt the *average* $F1$ *score* as the performance metric. To do this on a dataset, we first evenly distributed the set of all users in the dataset into 10 folds such that, for each user label, the folds have the same fraction of users having the label. Then, for each model, we use 9 folds to train an SVM classifier using SVMlight toolbox[14], and use the remaining fold to test the learned classifier. We then compute the average $F1$ score obtained by each model with respect to the two user labels. The model with a higher score is the winner in the task.

### 5.5. Performance Comparison

Figure 14 shows the weighted entropy of the various models in the user clustering task for the Developer and Political affiliation datasets. Figure 15 shows the average $F1$ scores for the user classification task. The figures show that adding the behavior topic distributions improves the performance in user profiling. The TwitterLDA+ behaviorLDA model has lower weighted entropies and higher average $F1$ scores than the TwitterLDA model in both cases. Similarly, the GBT-noRegularization and GBT models also have lower weighted entropies and higher average $F1$ scores than the GBT-noBehavior model in both cases. However, the QBLDA model does not always

[14]http://svmlight.joachims.org/.

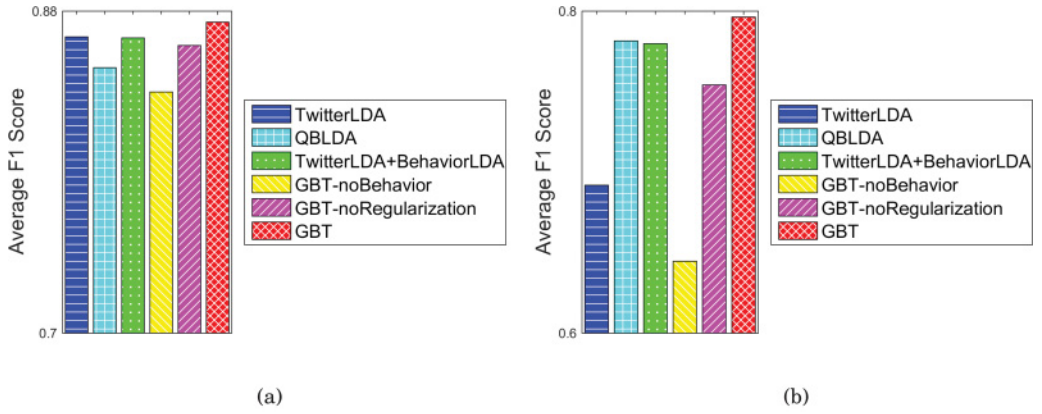



Fig. 15. Performance of different models in user classification task in (a) the Developer dataset and (b) the Political affiliation dataset.

outperform the TwitterLDA and GBT-noBehavior models. This suggests that, by aggregating user-behavior instances to their types, as in the QBLDA model, we may loss useful information for deriving user interest. Last, the figures clearly show that the GBT model significantly improves the performance over the GBT-noRegularization model, and also significantly outperforms all other models. This implies the effectiveness of the proposed sparsity regularization technique, and the GBT model provides a better way for representing users to more accurately differentiate users having different preferences.

### 5.6. Feature Analysis

Finally, we examine the most representative topic features for each user label learned by the SVM-based classifiers in the user classification tasks. For each model, we first normalize the topic features' weight returned by the classifiers (in the training phase) by the maximum weight of all the topic features associated with the same model. Thus, for each model, the normalized weight of each topic feature in the model represents the topic's relative importance in the model. As we run 10-fold cross-validation, for each model, we compute the average normalized weight of every topic across the 10 folds. The topics with highest and lowest average normalized weights are then the most representative for the two user labels, respectively.

Table VII shows the most representative topics for the two user labels in the Developer dataset learned by the comparative models. Again, we manually labeled the topics by examining their top words (as shown in Tables VIII, IX, and X) and top tweets. The table clearly shows that the most representative topics learned by the GBT model are more reasonable than those learned by the other models. The most representative topics learned by the GBT model are related to the two programming frameworks (*Microsoft Visual Studio Windows 8* and *Windows Tablets & Phones* for the .NET label; and *Scripting programming languages*, *Java software development*, and *Open-source data management systems* for the non-.NET label). On the other hand, the most representative topics for the two user labels learned by the other models are not always related to the two programming frameworks (e.g., *Entertainment* (TwitterLDA model), *Happenings in London* (QBLDA model), and *Readings* (TwitterLDA+behaviorLDA model)), or semantically discriminative for the frameworks (e.g., *Data management* (TwitterLDA and TwitterLDA+behaviorLDA models) and *HTML & Web* (QBLDA model)).

Similarly, Table XI shows the most representative topics for the two user labels in the Political affiliation dataset learned by the comparative models. Also, we manually

Table VII. Top Representative Topics for User Label in **Developer** Dataset Learnt by Comparative Models

| User label | TwitterLDA | | QBLDA | | TwitterLDA+behaviorLDA | | GBT | |
|---|---|---|---|---|---|---|---|---|
| | Topic | Topic Label | Topic | Topic Label | Topic | Topic Label | Topic | Topic Label |
| .NET | 66 | Microsoft Visual Studio | 5 | Microsoft Visual Studio | tweet topic 66 | Microsoft Visual Studio | 69 | Microsoft Visual Studio |
| | 7 | Windows Tablets & Phones | 47 | Windows Tablets Phones | tweet topic 7 | Windows Tablets & Phones | 35 | Windows 8 |
| | 40 | Lance Armstrong | 58 | Happenings in London | retweet topic 27 | Windows developers | 65 | Windows Tablets & Phones |
| non-.NET | 75 | Data management | 79 | HTML & Web | tweet topic 75 | Data management | 44 | Scripting programming languages |
| | 47 | iOS & iPhone | 52 | Internet & Media | tweet topic 47 | iOS & iPhone | 71 | Java software development |
| | 64 | Entertainment | 62 | Web Browsers | tweet topic 9 | Readings | 48 | Open-source data management systems |

Table VIII. Top Words of Topics Discovered by TwitterLDA and QBLDA Models From SE Dataset

| Model | Topic | Top words |
|---|---|---|
| **TwitterLDA** | 7 | windows,microsoft,surface,#windows8,#win8,metro,nokia,xbox,#bldwin,tablet |
| | 9 | reading,life,internet,book,language,person,english,thought,article,code |
| | 40 | armstrong,lance,bbc,riot,pussy,police,tour,jones,david,cameron |
| | 47 | ios,google,iphone,apple,maps,mac,android,ipad,facebook,chrome |
| | 64 | star,wars,disney,trek,graphics,episode,angry,birds,blog,lucasfilm |
| | 66 | windows,studio,visual,sharepoint,server,dotnet,sql,#sharepoint,microsoft,azure |
| | 75 | data,java,node,api,cloud,blog,database,server,code,performance |
| **QBLDA** | 5 | windows,studio,visual,microsoft,azure,sharepoint,#windows8,server,#win8,blog |
| | 47 | windows,microsoft,surface,nokia,lumia,tablet,xbox,#windows8,tablets,#surface |
| | 52 | media,science,internet,human,article,reading,data,journalism,change,book |
| | 58 | bbc,london,police,train,david,british,olympics,boris,olympic,cameron |
| | 62 | google,maps,ios,apple,chrome,internet,firefox,explorer,microsoft,safari |
| | 79 | mobile,responsive,content,html5,css,#rwd,device,images,presentation,media |

Table IX. Top Retweeted Users of Topics Discovered by **LDA** Model From **SE** Dataset

| Topic | Top users |
|---|---|
| 27 | hmemcpy,hhariri,markrendle,jbogard,adymitruk,gregyoung,troyhunt kellabyte,demisbellot,jeremydmiller |

labeled the topics by examining their top words (as shown in Tables XII, XIII, XIV, and XV) and top tweets. Again, the table clearly shows that the most representative topics learned by the GBT model are more reasonable than those learned by the two other models. All the representative topics learned by the GBT model are related to the two political affiliation labels (*Romney's tax policy* and *Romney's policies on same sex marriage*, in which Democrats criticize Romney for his proposed tax policy and his opposing to same-sex marriage, and *Speeches at DNC 2012* for the Democrats label; *Republicans on Sandra Fluke's speech at DNC 2012*, in which Republicans angrily react to Sandra Fluke's speech at the DNC 2012[15], *Religion issues*, and *Ron Paul* for the Republicans label). On the other hand, the most representative topics for the two user labels learned by the two other models are not always representative, for example, *Romney's taxes and religion* and *Obama's private life* (TwitterLDA model), in which

[15]http://www.slate.com/blogs/xx_factor/2012/09/06/sandra_fluke_at_the_dnc_angry_reaction_from_the_right_wing_is_good_for_obama_.html.

Table X. Top Words and Top Behavior Instances of Topics Discovered by GBT Model from SE Dataset

| Topic | Top words | Top hashtags | Top mentions | Top retweeted |
|---|---|---|---|---|
| 22 | iphone,apple,ipad<br>internet,data,wifi<br>home,battery,macbook | #iphone,#fail,#iphone5<br>#apple,#win,#appleevent<br>#uxaustralia,#iphon,#keynote | @branch,@google,@twitter<br>@kickstarter,@amazon,@att<br>@apple,@dropbox,@turf | marcoarment,gruber,dcurtis<br>siracusa,wilshipley,danielpunkass<br>mrgan,rands,jsnell |
| 26 | life,internet,human<br>problem,media,money<br>article,reading,thought | #a11y,#a11,#fai<br>#heweb12,#heweb1,#audio<br>#fail,#accessibility,#facepal | @prismati,@hnycombinator,@prismatic<br>@leolaporte,@danbenjamin,@doctorow<br>@kevinmarks,@jeffjarvis,@t | davewiner,umairh,timoreilly<br>anildash,pinboard,0xabad1dea<br>cstross,mralancooper,rands |
| 35 | windows,microsoft,#windows8<br>#win8,hosting,metro<br>surface,win8,#bldwin | #windows8,#win8,#windows<br>#surface,#bldwi,#wp8<br>#win,#windowsphone,#wp | @microsoft,@surface,@windowsphone<br>@ch9,@maryjofoley,@windows<br>@winobs,@windowsazure,@nokia | maryjofoley,shanselman,windowsphone<br>thurrott,benthepcguy,gcaughey<br>everythingms,windows,visualstudio |
| 38 | ios,mac,iphone<br>apple,google,chrome<br>windows,lion,mountain | #ios,#android,#ios6<br>#tb,#in,#apple<br>#androi,#chrome,#fb | @pocket,@appdotnet,@tweetbot<br>@marcoarment,@gruber,@tapbots<br>@hotdogsladies,@instagram,@jdalrymple | flyosity,stevestreza,mattgemmell<br>stroughtonsmith,mantia,panzer<br>viticci,sdw,joshhelfferich |
| 43 | coffee,beer,eating<br>dinner,lunch,ice<br>wine,cream,bacon | #yelp,#sf,#getgluehd<br>#opportunity,#sanfrancisco,#chicago<br>#austin,#career,#designer | @google,@starbucks,@jason<br>@instagram,@foursquare,@jezebel<br>@gawker,@mike,@kickstarter | mike_ftw,paulryangosling,anildash<br>pres_bartlet,beep,fakegrimlock<br>joelhousman,pourmecoffee,kissane |
| 44 | code,ruby,javascript<br>git,rails,github<br>python,data,php | #javascript,#ruby,#strangeloo<br>#nodejs,#php,#python<br>#github,#git,#rails | @github,@heroku,@rubyrogues<br>@steveklabnik,@travisci,@madisonruby<br>@ashedryden,@simplify,@tenderlove | steveklabnik,garybernhardt,tenderlove<br>dhh,github,roidrage<br>shit_hn_says,zedshaw,mfeathers |
| 48 | data,#bigdata,analytics<br>business,google,hadoop<br>information,database,analysis | #bigdata,#bigdat,#data<br>#ibm,#analytics,#hadoop<br>#ibmiod,#bi,#strataconf | @siliconbea,@timoreilly,@harvardbiz<br>@whitehouse,@nytimes,@radar<br>@digiphile,@wired,@slideshare | moonpolysoft,shanley,alex_gaynor<br>argv0,joedamato,pharkmillups<br>rickasaurus,jrecursive,cscotta |
| 65 | windows,microsoft,nokia<br>surface,android,tablet<br>samsung,nexus,lumia | #tech,#technology,#windowsphon<br>#switchtolumi,#smallbiz,#wincha<br>#technews,#microsof,#htc | @engadge,@verg,@cne<br>@sharethi,@io,@mashabl<br>@youtub,@verge,@rw | edbott,verge,tomwarren<br>drpizza,joshuatopolsky,theromit<br>ckindel,stroughtonsmith,bdsams |
| 66 | email,facebook,google<br>service,spam,emails<br>password,page,gmail | #facebook,#youtube,#blog<br>#howto,#twitte,#vide<br>#google,#lol,#fai | @twitter,@commun,@dropbox<br>@facebook,@bufferapp,@nealschaffer<br>@linkedin,@customerthink,@hootsuite | codinghorror,shanselman,rickygervais<br>levie,codepo8,mattcutts<br>marscuriosity,morgonfreeman,troyhunt |
| 69 | windows,studio,server<br>visual,sql,dotnet<br>azure,microsoft,blog | #windowsazure,#vs2012,#azure<br>#sqlserver,#microsoft,#powershell<br>#sqlserve,#sql,#mvpbuz | @pluralsight,@shanselman,@john<br>@shanselma,@codemash,@telerik<br>@julielerman,@ch,@oreillymedia | shanselman,pluralsight,kellabyte<br>jongalloway,haacked,migueldeicaza<br>elijahmanor,windowsazure,chrislove |
| 71 | sharepoint,java,programming<br>#sharepoint,code,blog<br>language,#java,scala | #sharepoint,#java,#fe<br>#javaone,#sp2013,#sharepoin<br>#scala,#sharepoint2013,#javaon | @thefanc,@skillsmatter,@jenkinsci<br>@dzone,@newsycombinator,@java<br>@infoq,@gregyoung,@kevlinhenney | debasishg,wfaler,dzone<br>fogus,java,psnively<br>jboner,jamesiry,typesafe |
| 76 | home,kids,house<br>#fb,life,car<br>dog,room,playing | #runkeepe,#wtf,#debat<br>#debate,#justsayin,#awesome<br>#awesom,#wt,#winnin | @klout,@twitter,@runkeeper<br>@pinteres,@marscuriosity,@kickstarter<br>@jack,@theonion,@oatmeal | neiltyson,sarcasticrover,theonion<br>robdelaney,wilw,honesttoddler<br>hotdogsladies,marscuriosity,chrisrockoz |

Table XI. Top Representative Topics for User Labels in Political Affiliation Dataset Learned by Comparative Models

| User label | TwitterLDA | | QBLDA | | TwitterLDA+behaviorLDA | | GBT | |
|---|---|---|---|---|---|---|---|---|
| | Topic | Topic Label | Topic | Topic Label | Topic | Topic Label | Topic | Topic Label |
| Democrats | 35 | Romney's taxes and religion | 5 | Romney's policies on same-sex marriage | retweet topic 37 | Left-leaning political bloggers | 7 | Romney's tax policy |
| | 2 | Democrats on RNC 2012 | 79 | Romney's tax policy | retweet topic 34 | Democrat politicians & pro-Democrat organizations | 76 | Romney's policies on same-sex marriage |
| | 30 | DNC 2012 | 36 | Voting issues | tweet topic 30 | DNC 2012 | 67 | Speeches at DNC 2012 |
| Republicans | 10 | Republicans on Obama's speech at DNC 2012 | 16 | Republicans on Sandra Fluke's speech at DNC 2012 | retweet topic 31 | Republican politicians & pro-Republican organizations | 57 | Republicans on Sandra Fluke's speech at DNC 2012 |
| | 21 | Obama's private life | 26 | Public debt | tweet topic 10 | Republicans on Obama's speech at DNC 2012 | 39 | Religion issues |
| | 67 | Religion issues speech at DNC 2012 | 15 | Ron Paul | hashtag topic 22 | Living status | 40 | Ron Paul |

Table XII. Top Words of Topics Discovered by TwitterLDA and QBLDA Models from Two-Week Dataset

| Model | Topic | Top words |
|---|---|---|
| **TwitterLDA** | 2 | #p2,#gop,#tcot,#rnc,#dnc2012,romney,#gop2012,#romney,#rnc2012,#obama2012 |
| | 10 | obama,speech,dnc,#dnc2012,stadium,convention,charlotte,#tcot,debt,dems |
| | 21 | obama,michelle,college,#dnc2012,barack,money,#tcot,president,kids,romney |
| | 30 | #dnc2012,obama,charlotte,convention,dnc,president,tampa,#dnc,delegates,speech |
| | 35 | romney,mitt,tax,bain,capital,#romney,taxes,money,mormon,#p2 |
| | 67 | god,platform,jerusalem,dnc,party,democrats,#dnc2012,israel,obama,dems |
| **QBLDA** | 5 | gay,marriage,labor,romney,rights,#p2,union,workers,#lgbt |
| | 15 | paul,ron,romney,gop,#ronpaul,convention,supporters,delegates,rnc |
| | 16 | fluke,sandra,#dnc2012,bill,jason,clinton,biggs,birth,dnc |
| | 26 | debt,obama,trillion,#tcot,#dnc2012,#obama,unemployment,budget,#romneyryan2012 |
| | 36 | voter,voting,law,federal,ohio,election,texas,gop,voters |
| | 79 | romney,tax,mitt,bain,taxes,money,rich,cuts,capital |

Table XIII. Top Retweeted Users of Topics Discovered by LDA Model from Two-Week Dataset

| Topic | Top words |
|---|---|
| 34 | obama2012,barackobama,truthteam2012,thedemocrats,demconvention michelleobama,donnabrazile,edshow,ofa_nc,jameshaning |
| 37 | angryblacklady,otoolefan,gottalaff,shoq,karoli,jeffersonobama,steveweinstein,owillis eclecticbrotha,bobcesca_go |

Table XIV. Top Hashtags of Topics Discovered by LDA Model from Two-Week Dataset

| Topic | Top words |
|---|---|
| 22 | #areyoubetteroff,#failingagenda,#16trillionfail,#areyoubetterof #failingagend,#forward2012,#wirigh,#wiright,#arithmetic,#16trillionfai |

people talk about Romney and Obama both positively and negatively; *Voting issues* and *Public debt* (QBLDA model), a topic that was actively talked about by users of both parties[16]; and *Living status*, a controversial topic that was first raised by Republicans followed by many opposing responses, even from the Republicans[17].

## 6. CONCLUSION

In this article, we propose a novel topic model for simultaneously modeling realms and users' topical interest in microblogging data. Our model associates user behavior with the latent topics as well as to model multiple types of behavior in a common framework. To learn the model's parameters, we propose an efficient Gibbs sampling method. We further develop a regularization technique incorporated with the sampling method so that the proposed model is biased to learn more semantically clear realms. We also report experiments on two Twitter datasets showing the effectiveness of the proposed model in topic modeling, as well as its improvement over other state-of-the-art topic models in some user-profiling tasks.

This work can be extended in several directions. First, 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 [Newman et al. 2009], and stale synchronous parallel implementation of variational inference procedures [Ho et al. 2013]. Second, a user may adopt a behavior because the user is socially or topically motivated [Prentice et al. 1994]. Distinguishing between these two

[16]http://www.huffingtonpost.com/2012/08/27/womens-vote-2012-election_n_1832825.html?
[17]http://thecaucus.blogs.nytimes.com/2012/09/04/republicans-ask-are-you-better-off-and-many-reply-yes/.

Table XV. Top Words and Top Behavior Instances of Topics Discovered by GBT from Two-Week Dataset

| Topic | Top words | Top hashtags | Top mentions | Top retweeted |
|---|---|---|---|---|
| 3 | #tcot,chris,obama<br>msnbc,racist,matthews<br>media,romney,liberal | #tco,#tcot,#twisters<br>#p2,#twister,#caring<br>#carin,#racist,#lapdogmedia | @msnbc,@jasonbiggs,@barackobama<br>@nickelodeontv,@hardball,@sandrafluke<br>@dwstweets,@truthteam2012,@noltenc | kesgardner,keder,noltenc<br>rbpundit,twitchyteam,iowahawkblog<br>toddkincannon,soopermexican,jtlol |
| 5 | obama,romney,speech<br>#dnc2012,clinton,convention<br>mitt,president,gop | #dnc201,#dnc2012,#rnc201<br>#dnc101,#literally,#factcheck<br>#drink,#dcn201,#factchec | @dwstweets,@chucktodd,@stefcutter<br>@reince,@davidaxelrod,@msnbc<br>@jaketapper,@jonahnro,@sandrafluke | guypbenson,jimgeraghty,iowahawkblog<br>jonahnro,noltenc,melissatweets<br>ewerickson,michellemalkin,freddoso |
| 7 | romney,mitt,tax<br>bain,#p2,capital<br>money,#tcot,taxes | #p2,#topprog,#ctl<br>#p2b,#p21,#toppro<br>#ct,#connectthelef,#p3 | @thinkprogres,@dailyko,@dailykos<br>@tp,@thinkprogress,@politicusus<br>@motherjone,@salo,@tpm | mattison,bluedupage,thenewdeal<br>gottalaff,rcooley123,factsaboutmitt<br>thedailyedge,sunshineejc,watchdogsniffer |
| 8 | obama,america,president<br>#tcot,romney,barack<br>#dnc2012,#dncin4words,country | #dncin4words,#howtopissoffademocrat<br>#overheardatdnc2012,#obamatvshows<br>#iamnotademocratbecause | @jjauthor,@klsouth,@slone<br>@katyinindy,@irritatedwoman,@chucknellis<br>@conservative,@usahipster,@patdollard | jjauthor,klsouth,nathanhale1775<br>slone,polarcoug,chucknellis<br>gregwhoward,katyinindy,tricianc1 |
| 17 | obama,#dnc2012,biden<br>joe,chair,romney<br>#tcot,dog,speech | #dnc2012,#rnc2012,#insertchai<br>#insertchair,#overheardatdnc201,#fai<br>#chairmovies,#eastwooding,#justsayin | @dloesch,@chrisloesch,@dloesc<br>@clayaiken,@michellemalkin,@katiepavlich<br>@melissatweets,@codepink,@gaypatriot | dloesch,soopermexican,gaypatriot<br>cnservativepunk,kurtschlichter,melissatweets<br>red_red_head,tabithahale,chrisloesch |
| 28 | obama,debt,jobs<br>#dnc2012,tax,trillion<br>bush,#tcot,economy | #dnc2012,#gop2012,#areyoubetteroff<br>#forward2012,#obamaisntworking<br>#areyoubetterof | @mittromney,@barackobama,@paulryanvp<br>@mittromne,@gop,@thedemocrats<br>@paulryanv,@reince,@reinc | mittromney,paulryanvp,romneyresponse<br>keder,gop,romneycentral<br>reince,kesgardner,teamromney |
| 31 | #gop2012,romney,#rnc2012<br>speech,mitt,#rnc<br>christie,ryan,ann | #gop201,#gop2012,#condi<br>#tampa201,#webuiltit,#g0p201<br>#chrischristie,#webuilti,#2012gop | @mittromney,@paulryanvp,@anndromney<br>@gopconvention,@govchristie,@marcorubio<br>@condoleezzarice,@ricksantorum,@reince | buzzfeedandrew,thefix,zekejmiller<br>chucktodd,daveweigel,ezraklein<br>buzzfeedben,jaketapper,larrysabato |
| 39 | god,platform,jerusalem<br>dnc,democrats,party<br>#dnc2012,obama,israel | #tcot,#mitt2012,#mitt201<br>#liblies,#catco,#jcot<br>#liblie,#uselection,#2016movie | @sharethi,@times247,@michellemalki<br>@townhallco,@politic,@hotairblo<br>@nypostopinio,@newyorkpos,@theblaz | dickmorristweet,davidlimbaugh,ingrahamangle<br>michellemalkin,monicacrowley,dennisdmz<br>gop,arifleischer,theblaze |
| 40 | paul,ron,romney<br>rnc,gop,#ronpaul<br>convention,party,#tcot | #ronpaul,#tlot,#rnc<br>#romney,#gogaryjohnson,#ronpau<br>#rncpowergrab,#gop,#libertarian | @youtub,@govgaryjohnson,@youtube<br>@ronpaul,@dailypau,@cbsradionews<br>@govgaryjohnso,@i,@examinerco | 1marchella,govgaryjohnson,tweetamiracle<br>i_am_change_usa,iworkiron,cblacktx<br>libertarian_76,julieborowski,j3vol |

(Continued)

Table XV. Continued

| Topic | Top words | Top hashtags | Top mentions | Top retweeted |
|---|---|---|---|---|
| 54 | convention,#dnc2012,tampa<br>#gop2012,charlotte,rnc<br>#rnc2012,dnc,gop | #rnc201,#rnc2012,#tampa<br>#clt,#tampabay,#tampaba<br>#politicolive,#carolinafest,#cl | @politico,@nytimes,@ron<br>@washingtonpost,@newtgingrich<br>@nationaljournal | antderosa,buzzfeedben,nytjim<br>buzzfeedandrew,gov,daveweigel<br>politifact,rollcall,dylanbyers |
| 57 | fluke,clinton,sandra<br>bill,#dnc2012,#tcot<br>dnc,convention,#dnc | #dnc2012,#waronwomen,#tiot<br>#gop2012,#istandwithann,#breitbartnet<br>#tio,#waronwome | @shareaholi,@newsninja2012<br>@2016themovie,@sharethis<br>@drudge,@gopblackchick,@therightsarah | newsninja2012,tmims50,shaughn_a<br>conservative_vw,thesavvy,becca51178<br>lfoshie,la_writerchick,themotleymind |
| 65 | government,america,obama<br>party,god,freedom<br>#tcot,american,country | #obama,#mittromney,#democrats<br>#gop,#romney,#barackobama<br>#msm,#paulryan,#clinteastwood | @barackobama,@foxnews,@jjauthor<br>@blackrepublican,@dineshdsouza,@obama<br>@2016themovie,@blazingcatfur,@daggy1 | jjauthor,newsninja2012,conservative_vw<br>prfekrdumbrella,2016themovie,pac43<br>elishanews,drmartyfox,daggy1 |
| 67 | clinton,#dnc2012,bill<br>obama,speech,president<br>#dnc,michelle,convention | #dnc2012,#billclinton,#clinton<br>#dnc,#flotus,#michelleobama<br>#potus,#biden,#joebiden | @barackobama,@mittromney<br>@michelleobama,@barackobam<br>@obama2012,@joebiden,@huffpostpol | obama2012,barackobama,truthteam2012<br>thedemocrats,edshow,demconvention<br>moveon,donnabrazile,michelleobama |
| 76 | rape,gop,gay<br>marriage,platform,akin<br>#p2,abortion,#gop | #gop,#lgbt,#republican<br>#republicans,#romney,#mitt<br>#go,#ryan,#waronwomen | @mittromney,@cspanw,@paulryanvp<br>@cspanwj,@anndromney,@reppaulryan<br>@dccc,@natlwow,@speakerboehner | thedailyedge,thenewdeal,barackobama<br>chrisrockoz,rcdewinter,sheshego<br>teapartycat,sunshineejc,p0tus |
| 77 | #dnc2012,#dnc,speech<br>obama,biden,convention<br>dnc,#rnc,joe | #rn,#rnc,#dn<br>#dnc,#dnc1,#rnc12<br>#rnc1,#dnc12,#santorum | @thefix,@ezraklein,@daveweigel<br>@realdonaldtrump,@buzzfeedandrew<br>@samsteinhp | mattyglesias,ezraklein,drgrist<br>daveweigel,brianbeutler,pourmecoffee<br>jbouie,joshtpm,joshgreenman |

types of motivation is important to many applications, but still a challenging problem. Third, it is potentially helpful to incorporate prior knowledge into the proposed model. Examples of the prior knowledge are topic-indicative features [Balasubramanyan et al. 2013], and social community labels for some users [Hoang et al. 2014]. Last, the types of behavior that we model in this work are of a general nature and can be addressed to users, items, or some groups of users/items. In the future, we would like to extend the proposed model to incorporate social communities, topical interests of social communities, and behavior of users addressed to other users within their social communities.

## REFERENCES


Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. 2008. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* 9, 1981–2014.

Ramnath Balasubramanyan and William W. Cohen. 2013. Regularization of latent variable models to obtain sparsity. In *SDM*.

Ramnath Balasubramanyan, Bhavana Bharat Dalvi, and William W. Cohen. 2013. From topic models to semi-supervised learning: Biasing mixed-membership models to exploit topic-indicative features in entity clustering. In *ECML/PKDD*.

Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2014. Who to follow and why: Link prediction with explanations. In *KDD*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022.

Antoine Boutet, Hyoungshick Kim, and Eiko Yoneki. 2012. What's in your tweets? I know who you supported in the UK 2010 general election. In *ICWSM*.

Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. 2012. Collaborative personalized tweet recommendation. In *SIGIR*.

Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering* 26, 12.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Political polarization on twitter. In *ICWSM*.

Peng Cui, Fei Wang, Shaowei Liu, Mingdong Ou, Shiqiang Yang, and Lifeng Sun. 2011. Who should share what? Item-level social influence prediction for users and posts ranking. In *SIGIR*.

Onkar Dabeer, Prachi Mehendale, Aditya Karnik, and Atul Saroop. 2011. Timing tweets to increase effectiveness of information campaigns. In *ICWSM*.

Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. 2012. From chatter to headlines: Harnessing the real-time web for personalized news recommendation. In *WSDM*.

Qiming Diao and Jing Jiang. 2013. A unified model for topics, events and users on twitter. In *EMNLP*.

Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. 2012. Finding bursty topics from microblogs. In *ACL*.

Ying Ding. 2011. Community detection: Topological vs. topical. *Journal of Informetrics* 5, 4, 498–514.

Elena Erosheva, Stephen Fienberg, and John Lafferty. 2004. Mixed-membership models of scientific publications. *PNAS*.

Albert Feller, Matthias Kuhnert, Timm Oliver Sprenger, and Isabell M. Welpe. 2011. Divided they tweet: The network structure of political microbloggers and discussion topics. In *ICWSM*.

Wei Gao, Peng Li, and Kareem Darwish. 2012. Joint topic modeling for event summarization across news and social media streams. In *CIKM*.

Przemyslaw A. Grabowicz, Luca Maria Aiello, Victor M. Eguiluz, and Alejandro Jaimes. 2013. Distinguishing topical and social groups based on common identity and bond theory. In *WSDM*.

John Hannon, Mike Bennett, and Barry Smyth. 2010. Recommending Twitter users to follow using content and collaborative filtering approaches. In *RecSys*.

Gregor Heinrich. 2009. *Parameter Estimation for Text Analysis*. Technical Report.

Qirong Ho, James Cipar, Henggang Cui, Seunghak Lee, Jin Kyu Kim, Phillip B. Gibbons, Garth A. Gibson, Greg Ganger, and Eric Xing. 2013. More effective distributed ml via a stale synchronous parallel parameter server. In *NIPS*.

Tuan-Anh Hoang, William W. Cohen, and Ee-Peng Lim. 2014. On modeling community behaviors and sentiments in microblogging. In *SDM*.

Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. In *SOMA*.

Liangjie Hong, Byron Dom, Siva Gurumurthy, and Kostas Tsioutsiouliklis. 2011. A time-dependent topic model for multiple text streams. In *KDD*.

Yuheng Hu, Ajita John, Fei Wang, Doree Duncan Seligmann, and Subbarao Kambhampati. 2012. ET-LDA: Joint topic modeling for aligning, analyzing and sensemaking of public events and their Twitter feeds. In *AAAI*.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we twitter: Understanding microblogging usage and communities. In *WebKDD/SNA-KDD'07*.

Appelo Jurgen. 2009. Twitter top 100 for software Developers. Retrieved December 5, 2016 from http://www.noop.nl/2009/02/twitter-top-100-for-software-developers.html.

Farshad Kooti, Haeryun Yang, Meeyoung Cha, P. Krishna Gummadi, and Winter A. Mason. 2012. The emergence of conventions in online social networks. In *ICWSM*.

Haewoon Kwak, Hyunwoo Chun, and Sue Moon. 2011. Fragile online relationship: A first look at unfollow dynamics in Twitter. In *CHI*.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *WWW*.

Daifeng Li, Bing He, Ying Ding, Jie Tang, Cassidy Sugimoto, Zheng Qin, Erjia Yan, Juanzi Li, and Tianxi Dong. 2010. Community-based topic modeling for social tagging. In *CIKM*.

Kwan Hui Lim and Amitava Datta. 2012. Following the follower: Detecting communities with common interests on Twitter. In *HT*.

Kar Wai Lim and Wray Buntine. 2014. Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In *CIKM*.

Tianyi Lin, Wentao Tian, Qiaozhu Mei, and Hong Cheng. 2014. The dual-sparse topic model: Mining focused topics and focused terms in short text. In *WWW*.

Jun S. Liu. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association* 89, 427, 958–966.

Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. 2009. Topic-link LDA: Joint models of topic and author community. In *ICML*.

Zhunchen Luo, Miles Osborne, Jintao Tang, and Ting Wang. 2013. Who will retweet me? Finding retweeters in Twitter. In *SIGIR2013*.

Zhiqiang Ma, Wenwen Dou, Xiaoyu Wang, and Srinivas Akella. 2013. Tag-latent Dirichlet allocation: Understanding hashtags and their relationships. In *WI/IAT 2013*.

Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. 2015. A tri-role topic model for domain-specific question answering. In *AAAI*. 224–230.

Andrew McCallum, Andres Corrada-Emmanuel, and Xuerui Wang. 2005. Topic and role discovery in social networks. *Computer Science Department Faculty Publication Series* 3.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *SIGIR*.

Matthew Michelson and Sofus A. Macskassy. 2010. Discovering users' topics of interest on Twitter: A first look. In *AND*.

Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. 2008. Joint latent topic models for text and citations. In *KDD*.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. Distributed algorithms for topic models. *Journal of Machine Learning Research* 10, 1801–1828.

M. E. J. Newman. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103, 23, 8577–8582.

Ye Pan, Feng Cong, Kailong Chen, and Yong Yu. 2013. Diffusion-aware personalized social update recommendation. In *RecSys*.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, Republicans and Starbucks afficionados: User classification in Twitter. In *KDD*.

Deborah A. Prentice, Dale T. Miller, and Jenifer R. Lightdale. 1994. Asymmetries in attachments to groups and to their members: Distinguishing between common-identity and common-bond groups. *Key Readings in Social Psychology*. Psychology Press, 83.

Minghui Qiu, Jing Jiang, and Feida Zhu. 2013. It is not just what we say, but how we say them: LDA-based behavior-topic model. In *SDM*.

Daniel Ramage, Susan T. Dumais, and Daniel J. Liebling. 2010. Characterizing microblogs with topic models. In *ICWSM*.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *ECML*.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *UAI*.

Mrinmaya Sachan, Danish Contractor, Tanveer A. Faruquie, and L. Venkata Subramaniam. 2012. Using content and interactions for discovering communities in social networks. In *WWW*.

Mrinmaya Sachan, Avinava Dubey, Shashank Srivastava, Eric P. Xing, and Eduard Hovy. 2014. Spatial compactness meets topical consistency: Jointly modeling links and content for community detection. In *WSDM*.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *WWW*.

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *SocialCom*.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014a. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *ACL*.

Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen, and Xiaofei He. 2014b. Interpreting the public sentiment variations on Twitter. *TKDE* (2014).

Jan Vosecky, Di Jiang, Kenneth Wai-Ting Leung, Kai Xing, and Wilfred Ng. 2014. Integrating social and auxiliary semantics for multifaceted topic modeling in Twitter. *ACM Transactions on Internet Technology* 14, 4, 27.

Jinpeng Wang, Wayne Xin Zhao, Yulan He, and Xiaoming Li. 2014. Infer user interests via link structure regularization. *ACM Transactions on Intelligent Systems and Technology* 5, 2.

Michael J. Welch, Uri Schonfeld, Dan He, and Junghoo Cho. 2011. Topical semantics of Twitter links. In *WSDM*.

Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Who says what to whom on Twitter. In *WWW*.

Pengtao Xie and Eric P. Xing. 2013. Integrating document clustering and topic modeling. In *UAI*.

Rui Yan, Mirella Lapata, and Xiaoming Li. 2012. Tweet recommendation with graph co-ranking. In *ACL*.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *WWW*.

Jiang Yang and Scott Counts. 2010. Predicting the speed, scale, and range of information diffusion in Twitter. In *ICWSM*.

Jaewon Yang and Jure Leskovec. 2012. Community-affiliation graph model for overlapping network community detection. In *ICDM*.

Jaewon Yang, Julian McAuley, and Jure Leskovec. 2013. Community detection in networks with node attributes. In *ICDM*.

Jaewon Yang, Julian McAuley, and Jure Leskovec. 2014. Detecting cohesive and 2-mode communities indirected and undirected networks. In *WSDM*.

Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. 2012. We know what@ you# tag: Does the dual role affect hashtag adoption? In *WWW*.

Shuang-Hong Yang, Alek Kolcz, Andy Schlaikjer, and Pankaj Gupta. 2014. Large-scale high-precision topic modeling on Twitter. In *KDD*.

Tae Yano, William W. Cohen, and Noah A. Smith. 2009. Predicting response to political blog posts with topic models. In *NAACL*.

Dawei Yin, Liangjie Hong, and Brian D. Davison. 2011. Structural link analysis and prediction in microblogs. In *CIKM*.

Hongzhi Yin, Bin Cui, Hua Lu, Yuxin Huang, and Junjie Yao. 2013. A unified model for stable and temporal topic detection from social media data. In *ICDE*.

Zhijun Yin, Liangliang Cao, Quanquan Gu, and Jiawei Han. 2012. Latent community topic analysis: Integration of community discovery with topic modeling. *ACM Transactions on Intelligent Systems and Technology* 3, 4, 63.

Michele Zappavigna. 2011. Ambient affiliation: A linguistic perspective on Twitter. *New Media & Society* 13, 5.

Dejin Zhao and Mary Beth Rosson. 2009. How and why people twitter: The role that micro-blogging plays in informal communication at work. In *GROUP'09*.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing Twitter and traditional media using topic models. In *ECIR*.

Ding Zhou, Eren Manavoglu, Jia Li, C. Lee Giles, and Hongyuan Zha. 2006. Probabilistic models for discovering e-communities. In *WWW*.