


Community Detection on Social Networks With Sentimental Interaction

Bingdao Feng, Tianjin University, China

 <https://orcid.org/0000-0002-0806-6405>

Fangyu Cheng, Harbin Institute of Technology, China*

Yanfei Liu, Tianjin University, China

Xinglong Chang, Tianjin University, China

Xiaobao Wang, Tianjin University, China

Di Jin, Tianjin University, China

ABSTRACT

Many studies on community detection are mainly based on the similarity in friendship between users. Recent studies have started to explore node contents to identify semantically meaningful communities. However, the sentimental interaction information which plays an important role in community detection is often ignored. By analyzing and utilizing the abundant sentimental interaction information, one can not only more precisely identify the communities, but also discover the interesting interactions and conflicts between these communities. Based on this concept, the authors propose a new Community Sentiment Diffusion Detection Model (CSDD), which utilizes sentimental information embedded in forward posts. Furthermore, the authors present an efficient variational algorithm for model inference. The community detection results have been verified on two large Twitter datasets. It is experimentally demonstrated that we can provide a fine-grained view of sentimental interaction between communities and discover the mechanism of sentiment diffusion between communities.

KEYWORDS

Community Detection, Sentiment Diffusion, Variational Algorithm, Bayesian Model, Social Networks

In recent years, one of the hottest topics in online social networking has been community detection (Zhe et al., 2019; Zhang et al., 2020). Typically, the network's basic units are abstracted as nodes, and the mutual influence of units is abstracted as edges. *Community* is defined as a group of nodes that are closely connected (Girvan & Newman, 2002). Not only can community detection help people understand the structure of a network, but it can also be used to find functional modules in protein interactions, find a group of nodes with similar properties (Vlaic et al., 2018), and even predict the actions of nodes in complex systems—for instance, finding political factions in a blog network (Bickel & Sarkar, 2016).

DOI: 10.4018/IJSWIS.341232

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

User-defined communities are an essential component of many social networks, allowing users to express their differing perspectives on current events. According to the survey by Fortunato and Hric (2016), a large number of traditional community detections are based on network topology. We can basically infer communities by taking into account the network structure from various perspectives (Jiang et al., 2018; Cheng et al., 2018). In addition to structural information, social networks also contain a large amount of textual semantic information. By accounting for the semantics of texts and other aspects, researchers can increase the accuracy of community detection results (Neville et al., 2003). Recent studies have begun to consider the diffusion on community level (Hu et al., 2015; Tu et al., 2018). In work by Cai et al. (2017), the concept of profiling is put forth, and communities are found by comprehensively describing both their internal nature and their external behaviors. In the latest work, Wang et al. (2020a) propose the GHIPT model by integrating group homophily and individual personality traits into topics on the basis of intra-community and inter-community links. Studies (e.g., Kumar et al., 2018) show that in real networks, inter-community actions can be positive, leading to the exchange of information and ideas, or they can take a negative turn, leading to overt conflicts between community members. The work demonstrates to us that sentimental interaction plays a very important role in community detection.

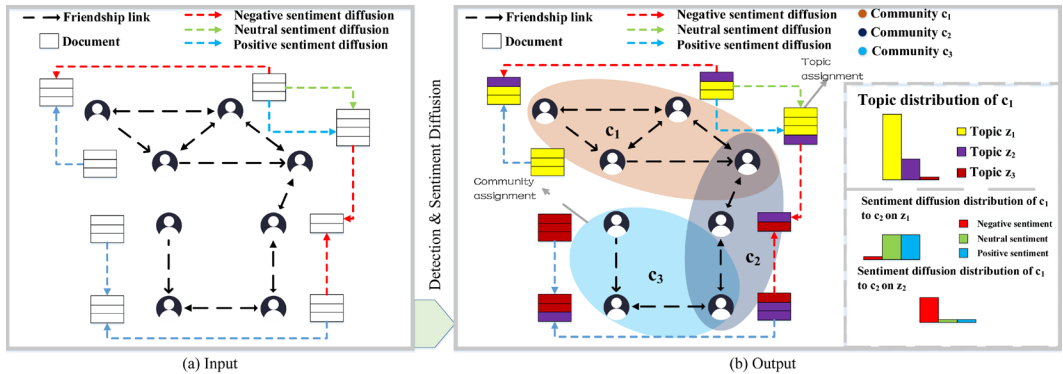
Currently, while there are various works (Chen et al., 2017; Wang et al., 2017; Feng et al., 2020) that take sentiments into account to discover users who are closely connected and highly consistent in their sentiments about one specific product or service; these works do not consider the impact of sentimental interaction on community detection. An intuitive idea is that members of a community should disseminate documents with similar sentiments on a specific topic.

Let us take the example of American political parties to illustrate sentimental interactions. Some voters support Joe Biden, while others support Donald Trump. Because they hold opposing political views, there would be negative sentimental interactions between the two communities, but positive sentimental interactions within communities. Sentimental interactions play an important role in the formation of communities. Documents are frequently disseminated with sentimental information that has an impact on the structure of communities. Understanding the generation mechanisms of sentiment diffusion can thus promote optimal community structures not only from the perspective of topology and semantics but also from the perspective of sentiment diffusion.

There is a straightforward and intuitive method for discovering communities directly through sentiment diffusion. We can define communities using existing methods, then count the sentiments of documents being spread at the community level, and finally modify the community structure on the basis of statistics about sentiments. However, this approach fails to capture our intuitive notion that sentiment diffusion may also be community-related.

Therefore, designing a reasonable and unified model is necessary. For this purpose, we propose a **Community Sentiment Diffusion Detection (CSDD)** that combines friends' links and document content (including words and sentiment diffusion). Our approach has two primary goals: 1) to identify sentiment diffusion-based communities, i.e., communities that share nearly consistent sentiments on common topics, and 2) to reveal the mechanism of sentiment diffusion at the community level. As shown in Figure 1(a), we collected a series of data, including users, their friendship links, original and forwarded documents, and the sentimental polarity of each forwarded document, with the box representing the collection of all documents sent by the user. Using Twitter as an example, the diagram denotes users and their followers, tweets and retweets, as well as retweet sentiment polarity tags. In Figure 1(b), we extract the communities to which each user belongs, the topics that each community is interested in (e.g., community c_1 tends to publish topics z_1 and z_2), and the diffusion of each sentiment polarity at the community level (e.g., on topic z_1 , community c_1 tends to diffuse neutral and positive sentiment to the community c_2). To avoid ambiguity, we use *tweet* to refer to the original document and *retweet* to refer to the forwarded document.

Figure 1. Overview of sentiment diffusion and community detection



The contributions of the paper are summarized as follows:

- We present a new method of community detection that considers sentiment diffusion for improving the accuracy of community detection. To the best of our knowledge, we are the first to investigate this critical and challenging problem.
- To perform model inference, we transformed the aforementioned generalized semantic community detection problem into a MAP (maximum a posteriori) problem and developed an efficient variational inference algorithm.
- We conducted several experiments on two real datasets to demonstrate that the accuracy of community detection will be enhanced by considering sentiment diffusion.

The remainder of this work is arranged in the following manner: The relevant work is introduced in Section 2. Section 3 delves into the specifics of our model. The stages of model inference and parameter estimation are covered in Section 4. Section 5 contains experimental data as well as a case study. Finally, in Section 6, we wrap up the paper with some discussion.

RELEVANT WORK

Many methods of community detection have been proposed in recent years. The primary theories and techniques include the following: modularity-based methods (Yang et al., 2016; Newman, 2016; Zhang & Moore, 2014), hierarchical clustering methods (Li et al., 2015), spectral algorithms (Jia et al., 2015; Ma et al., 2018), dynamic algorithms (Jiang et al., 2018), statistical inference-based methods (Xie et al., 2018), etc. Please refer to the survey by Fortunato and Hric (2016). There are also some ways to use the metadata of the node to improve community detection results; for instance, Peel (2011, 2012) extended the stochastic block model for jointly modelling relational and class label information. Peel used the Supervised Blockmodel to achieve strong link-structure-based classification results and also provided a clear summary of network interactions to aid in the comprehension of the data.

Community Detection Based on Semantics

Many traditional community detection methods ignore node content attributes in favor of network topology information alone. Nevertheless, node content is crucial and helpful for community discovery, because it divides the community in a way that is more in line with its actual circumstances and makes a bigger contribution to the community. Node characteristics are based on the inherent structure of individuals, which is another way to express the nature of the community.

The topological information and content information complement each other, thus producing a more accurate result. Another advantage is that even if a single source of information is lost, a relatively stable community structure can still be built with another source of information, e.g., content analysis and personality characteristics. The idea of using a topic model-based method to unify the model and then estimating each parameter to find the community distribution was proposed by Balasubramanian and Cohen (2011) and Yang et al. (2009). Zhu et al. (2013) combined classic ideas in topic modeling with a variant of the mixed-membership block model recently developed in the statistical physics community; their model can also be used for detecting generalized communities. Block-LDA (Balasubramanian & Cohen, 2011) combines aspects of mixed membership stochastic block models and topic models to improve entity link modeling by jointly modeling links and text about the entities that are linked. Yang et al. (2009) proposed a discriminative methodology for detecting communities by integrating link and content analysis. For content analysis, they also developed a conditional model and a discriminative model. Neville et al. (2003) proposed a weighted adjacency matrix to allow a content similarity measure. The weight of each edge was defined as the number of attribute values shared by the two end vertices. They then applied three existing graph clustering algorithms to the weighted adjacency matrix to perform network clustering. Furthermore, since semantic structures are related to each other, TCCD (Wang et al., 2019) can learn the community structure and semantic interpretation of each community and search for the correlations of different topics in community detection.

Community Detection Based on Diffusion

It is important to recognize that community detection, which integrates network topology and community semantics, has advanced to the point where the outcomes can roughly mirror the community structure. However, because billions of users generate documents and disseminate a large amount of information every day, diffusion must also be considered. If we merely look at diffusion at the level of individual users, we consume a large quantity of time with getting comprehensive answers. Therefore, some researchers take the effect of community spread into account. The COLD (Hu et al., 2015) method is expressive while remaining efficient. The extracted community-level patterns enable diffusion exploration from a new perspective. Han and Tang (2015) pointed out that the results of detection and behavior prediction are still far from satisfactory, so they argued for incorporating all the information about nodes and edges such as links, communities, user attributes, roles, and behaviors. Cai et al. (2017) proposed the concept of community profiling, with rich user information such as user-published content and user diffusion links, and they characterized a community in terms of both its internal content profile and its external diffusion profile. Tu et al. (2018) incorporated community structure of network embedding methods and also proposed a new approach to learning network embeddings with regular equivalence. Wang et al. (2018) proposed a GHIPT model by integrating group homophily and individual personality of topics, relying on intracommunity and intercommunity links. Our method not only comprehensively considers topology, semantics, and diffusion, but also makes use of extremely important sentiment information, and we can detect communities more accurately using abundant information of sentiment diffusion.

Community Detection Based on Sentiment

In social networks, sentiment is tremendously significant. Each document not only contains semantic data but also a great deal of sentimental data. Communities are now frequently discovered based on sentimental information.

Recently, there have been some papers studying community detection and sentiment mining together. The JST model (Lin & He, 2009) is a directed graph model that considers the influence of sentimental factors on topic mining through the graph model for the first time. The POT model (Chen et al., 2017) introduces the concept of community detection based on opinions, which simulates the social relationships, common interests, and common opinions of users in a unified way. Wang et al.

(2017) put forward the concept of sentimental community, that is, the close linking of a group of users' sentiments about a product. This technology is suitable for the practical application of product recommendation. Feng et al. (2020) proposed an extraction model based on user group text and sentimental topics to alleviate the problem of feature sparsity in short text.

However, existing community detection algorithms based on sentiment do not take the transmission characteristics of sentiments into account. A method will not consider sentiment rendering if it relies only on the sentimental attributes of tweets to aggregate users into communities. The sentiment diffusion and discovery communities can be tracked using our method.

THE BAYESIAN MODEL

In this section, we present the novel Bayesian graphical model, denoted as **Community Sentiment Diffusion Detection (CSDD)**. The CSDD model can find the community to which users belong, the semantics to which each community belongs, and the topic to which each word belongs on the basis of tweet content, retweet content, topology, and sentimental polarity of retweets. For instance, it can determine the fraction of the polarity of sentiment diffusion within communities on a specific topic. Thus, CSDD can simulate the generation process of community, original topic, forward topic, and sentiment diffusion.

Problem Formulation

We first describe our problem formulation. The notations applied in this paper are summarized in Table 1.

Definition 1. We define a **social network** as $G = (U, D, D^*, \mathbf{f}, \mathbf{e})$, where $u \in U$ is a user, $d \in D$ is a user published tweet, and $d^* \in D^*$ is a user published retweet. There are two types of edges in G . $f_{uu^*} \in \mathbf{f}$ is a friendship link from user u to user u^* ; $e_{ud^*} \in \mathbf{e}$ is a sentiment diffusion edge on retweet d^* . Given the social network G , our objectives are to 1) recognize the words in the content text as generated by original topics and forwarded topics; 2) partition G into c communities, z original topics, and z^* forwarded topics on the basis of network topology and contents; 3) express the mechanism of sentiment diffusion at community level by forwarded topics. Our novel probability graph model can solve all three problems at the same time.

Definition 2. We define a user u 's **community membership** as being a $|C|$ -dimensional multinomial distribution π_u over communities, where each dimension $\pi_{u,c}$ denotes the probability of the user u belonging to community c . $\pi_{u,c}$ satisfies the constraints $\sum_{c=1}^C \pi_{u,c} = 1$, and $\pi_{u,c} \in [0, 1]$.

Definition 3. **Original community semantics** is a $|K|$ -dimensional vector θ_c , where each dimension $\theta_{c,k}$ is the probability of the original topic k belonging to community c .

Definition 4. **Forwarded community semantics** is a $|K|$ -dimensional vector θ_c^* , and each dimension $\theta_{c,k}^*$ is the probability of forwarded topic k belonging to community c .

Definition 5. **Topic** $k \in \{1, 2, \dots, |K|\}$ is a $|V|$ -dimensional vector ϕ_k over all words, where each dimension $\phi_{k,v}$ is the probability of a word $v \in \{1, 2, \dots, |V|\}$ belonging to topic k .

Definition 6. Given the community assignment c_u of each node u , we then generate a **topology of friendship** f_{uu^*} of the node connected to u^* , represented by a $|U|$ -dimensional vector γ_c , which denotes the probability that node u^* from community c_i selects the node u^* as another node when it generates a friendship link.

Table 1. Notations used in this paper

Signs	Descriptions
$ U $	Number of users
$ D_u , D_u^* $	Number of tweets and retweets
$ W_{u_d} , W_{u_d}^* $	Number of tweet words and retweet words
$ V $	Total number of words in vocabulary V
$ F $	Number of friend relationships
$ C , K $	Number of communities and topics
f_{uu^*}	Directed link from user u to user u^*
$w_{ud_n}, w_{ud_n}^*$	The n -th words of d -th tweet and n^* -th words of d^* -th retweet
$e_{u_d u^*}$	Sentiment diffusion from user u^* to user u via d^* -th retweet
c_u	Community assignment for u
$z_{u_d}, z_{u_d}^*$	Topic assignment for the d -th tweet published by u and d^* -th retweet forwarded by u
π_u	Multinomial distribution over communities specific to user u
θ_c, θ_c^*	Multinomial distribution over original topic and forwarded topic to the c -th community
ϕ_k	Multinomial distribution over words specific to the k -th topic
γ_c	Multinomial distribution over friendships to the c -th community
$\eta_{c,c^*,k}$	Multinomial distribution over sentiment diffusion from community c^* to community c on topic k
$\alpha, \alpha^*, \beta, \Omega, \rho, \xi$	Parameters of Dirichlet priors

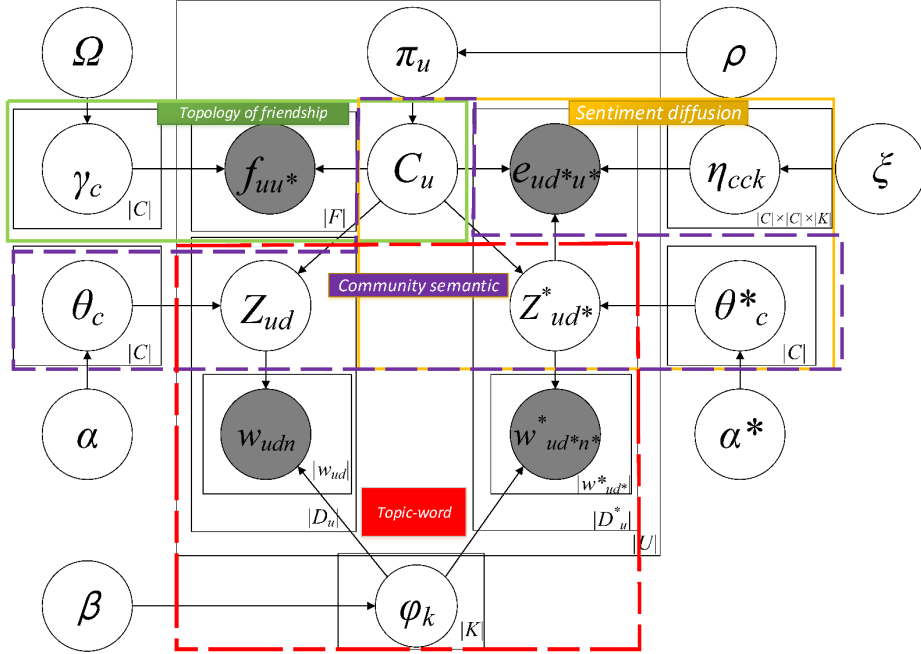
Definition 7. The **sentiment diffusion** e between two communities in the forwarded topic is a $|S|$ -dimensional vector $\eta_{c,c^*,k}$, where each dimension $\eta_{c,c^*,k,s}$ is the probability of sentiment s from the community c^* diffusing community c on the topic k .

Model Structure

We propose a probabilistic generative model that consists of four components: community semantics, topic-word, sentiment diffusion, and friendship topology. This model's probabilistic graphical representation is depicted in Figure 2.

Figure 2. Diagram of the generative model of CSDD

Note. Part 1 (purple box) denotes community semantics component. Part 2 (red box) denotes topic-word component. Part 3 (green box) denotes heterogeneous edge component. Part 4 (yellow box) denotes sentiment diffusion component.



Community semantics. Users who write documents on the same topic are more likely to belong to the same community, resulting in the community’s semantic nature. We categorize semantics into two types: original semantics and forwarded semantics. The topic of tweets written by users for the first time is considered the original semantics, and the topic of retweets is considered the forwarded semantics. Given the community label c_u of the user u , we need to sample the original topic label z_{ud} of the d -th tweet posted by the user u from a multinomial distribution; then we sample the forwarded topic label $z_{ud^*}^*$ of d^* -th retweet posted by user u from a multinomial distribution.

Topic-word. Each tweet d_u contains a bag of words $\{w_{ud_1}, \dots, w_{ud_n}\}$, and each retweet d_u^* contains a bag of words $\{w_{ud_1}^*, \dots, w_{ud_n^*}^*\}$, where n and n^* represent the length of the tweet and retweet. We generate these words on the basis of several latent factors, i.e., original topic-word distribution θ_c and forwarded topic-word distribution θ_c^* . Its generation process is similar to that of LDA (Blei et al., 2003), except that we think the topic generates documents rather than words, because the meaning of words in short texts is not clear.

Topology of friendship. This part works by dividing the nodes of a network into classes such that the members of each class have similar patterns of connection to other nodes (Newman & Leicht, 2007), which can reflect the topological relationship between two users. Given the community labels c_u of user u , we sample all edges connected to nodes u from a multinomial distribution. In essence γ_c denotes the “preferences” in community c_u .

Sentiment diffusion. We need to determine how to generate sentiment diffusion, so we get the sentimental polarity of the retweet $e_{u,d^*,s}$, which denotes the sentimental polarity of the d^* 's document forwarded by user u is s . The multinomial distribution is defined as

$$p(e_{u,d^*,s} = s | \eta_{c_u, c_u^*, z_{ud^*}^*}) = \eta_{c_u, c_u^*, z_{ud^*}^*, s}, \quad (1)$$

where $\eta_{c_u, c_u^*, z_{ud^*}^*}$ denotes the probability of polarity of sentiment diffusion that the community c^* of user u^* diffuses sentiments to the community c of user u on forwarded topic z^* and is subject to $\eta_{c_u, c_u^*, z_{ud^*}^*} = 1$. In our experiment, the dimension of s is set to 3, including negative, neutral, and positive. Next, we take a Bayesian approach to learning these parameters to find suitable parameters η instead of fixing each parameter randomly in advance. Because the Dirichlet distribution is a conjugate prior of multinomial likelihood, we use the Dirichlet distribution as the super parameter of the multinomial distribution, which is assumed to be given first and fixed to predetermined values. The Dirichlet hyperparameter form of the multinomial distribution of η is defined as

$$p(\eta_{cck} | \xi) = \frac{\Gamma\left(\sum_{s=1}^S \xi_s\right)}{\prod_{s=1}^S \Gamma(\xi_s)} \prod_{s=1}^S \eta_{c,c,k,s}^{\xi_s - 1}, \quad (2)$$

where $\Gamma(\cdot)$ is the Gamma function. The distribution is parameterized by a positive-real S -dimension vector; $\xi = (\xi_1, \xi_2, \xi_3)$ is a hyperparameter of this generative process.

Generative Process

We summarize all the generative processes as follows:

1. For each user $u \in \{1, 2, \dots, |U|\}$:
 - (a) Choose $\pi_u \sim Dir(\rho)$
2. For each community $c \in \{1, 2, \dots, |C|\}$:
 - (b) Choose $\gamma_c \sim Dir(\Omega)$
 - (c) Choose $\theta_c \sim Dir(\alpha)$
 - (d) Choose $\theta_c^* \sim Dir(\alpha^*)$
3. For each topic $k \in \{1, 2, \dots, |K|\}$:
 - (e) Choose $\phi_k \sim Dir(\beta)$
4. For each community sentiment diffusion link at topic z :
 - (f) Choose $\eta_{cck} \sim Dir(\xi)u$
5. For each user $u \in \{1, 2, \dots, |U|\}$:
 - (a) Choose community $c_u \sim Mul(\pi_u)$
 - (i) For user form an edge with user u^* , $f_{uu^*} = 1$

Draw friendship link u and u^* , $f_{uu^*} \sim Mul(\gamma_{cu})$

6. For the d -th tweet of user u :
 - (a) Choose community $z_{ud} \sim \text{Mul}(\theta_{c_u})$
 - (b) For the n -th word of the tweet d :
 - (i) Draw each word $w_{ud} \sim \text{Mul}(\varphi_{z_{ud}})$
7. For the d^* -th retweet of user u :
 - (a) Choose community $z_{ud}^* \sim \text{Mul}(\theta_{c_u}^*)$
 - (i) Draw each sentiment $e_{ud} \sim \text{Mul}(\eta_{c_u, c_u^*, z_{ud}^*})$
 - (b) For the n^* -th word of the tweet d :
 - (i) Draw each word $w_{ud}^* \sim \text{Mul}(\varphi_{z_{ud}^*})$

The Joint Distribution

With the above description, we have described the complete generative process of this model. We not only generate the parameters $\pi, \gamma, \theta, \theta^*, \phi, \eta$, but also generate hidden variables, i.e., C, Z, Z^* . We formally define a Bayesian model that denotes the underlying joint probability distribution.

Given the Dirichlet hyperparameters $\alpha, \alpha^*, \beta, \Omega, \rho, \xi$, we decompose the joint distribution over $\pi, \gamma, \theta, \theta^*, \phi, \eta, C, Z, Z^*$. The joint distribution was defined as

$$\begin{aligned}
 & p(\pi, \theta, \theta^*, \varphi, \gamma, \eta, C, Z, Z^*, \mathbf{f}, \mathbf{e}, \mathbf{w}, \mathbf{w}^* \mid \rho, \alpha, \alpha^*, \beta, \Omega, \xi) \\
 &= p(\pi \mid \rho) p(\theta \mid \alpha) p(\theta^* \mid \alpha^*) p(\varphi \mid \beta) p(\gamma \mid \Omega) \\
 &\cdot p(\eta \mid \xi) p(C \mid \pi) p(Z \mid \theta, C) p(Z^* \mid \theta^*, C) p(\mathbf{f} \mid \gamma, C) \\
 &\cdot p(\mathbf{w} \mid \varphi, Z) p(\mathbf{w}^* \mid \varphi, Z^*) p(\mathbf{e} \mid \eta, C, Z^*)
 \end{aligned} \tag{3}$$

where

$$\begin{aligned}
 & p(\pi \mid \rho) = \prod_{u=1}^U p(\pi_u \mid \rho), \quad p(\gamma \mid \Omega) = \prod_{c=1}^C p(\gamma_c \mid \Omega), \quad p(\theta \mid \alpha) = \prod_{c=1}^C (\theta_c \mid \alpha), \\
 & p(\theta^* \mid \alpha^*) = \prod_{c=1}^C (\theta_c^* \mid \alpha^*), \quad p(\varphi \mid \beta) = \prod_{k=1}^K (\varphi_k \mid \beta), \quad p(\eta \mid \xi) = \prod_{k=1}^K \prod_{c=1}^C \prod_{c^*=1}^C (\eta_{cc^*k} \mid \xi), \\
 & p(C \mid \pi) = \prod_{u=1}^U p(c_u \mid \pi_u), \quad p(Z \mid \theta, C) = \prod_{u=1}^U \prod_{d=1}^{D_u} p(z_{ud} \mid \theta_{c_u}), \quad p(Z^* \mid \theta^*, C) = \prod_{u=1}^U \prod_{d^*=1}^{D_u^*} p(z_{ud^*}^* \mid \theta_{c_u^*}^*), \\
 & p(\mathbf{f} \mid \gamma, C) = \prod_{u=1}^U \prod_{u^*=1}^U \prod_{u^*=1}^U \gamma_{u, u^*}^f, \quad p(\mathbf{e} \mid C, Z^*, \eta) = \prod_{u=1}^U \prod_{u^*=1}^U \prod_{d^*=1}^{D_u^*} \prod_{s=1}^S \eta_{c_u, c_u^*, z_{ud^*}^*, s}^e, \\
 & p(\mathbf{w} \mid \varphi, Z) = \prod_{u=1}^U \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} \varphi_{z_{ud}, n}^{w_{ud, n}}, \quad p(\mathbf{w}^* \mid \varphi, Z^*) = \prod_{u=1}^U \prod_{d^*=1}^{D_u^*} \prod_{n^*=1}^{N_{ud^*}^*} \varphi_{z_{ud^*}^*, n^*}^{w_{ud^*}^*, n^*}.
 \end{aligned}$$

Because the Dirichlet distribution is a conjugate multinomial likelihood prior, we set hyperparameters $\alpha, \alpha^*, \beta, \Omega, \rho, \xi$ to generate the above multinomial distribution. When we derive inference, the posterior will have a closed-form expression that can be used in math.

Model Optimization

First, we figure out the maximum a posteriori (MAP) approach to infer this model, which requires us to determine the equation shown below.

$$\hat{C}, \hat{Z}, \hat{Z}^* = \arg \max_{C, Z, Z^*} P(C, Z, Z^* | \mathbf{f}, \mathbf{e}, \mathbf{w}, \mathbf{w}^*),$$

where $P(C, Z, Z^* | \mathbf{f}, \mathbf{e}, \mathbf{w}, \mathbf{w}^*)$ is the posterior distribution of C , Z and Z^* given \mathbf{f} , \mathbf{e} , \mathbf{w} and \mathbf{w}^* . \hat{C} , \hat{Z} , and \hat{Z}^* provide the most probable clustering of users and tweets retweets independently and provides the best explanation for the given explicit variables' matrices \mathbf{f} , \mathbf{e} , \mathbf{w} , \mathbf{w}^* .

Despite its conceptual simplicity, the probabilistic inference problem is difficult to solve. A major difficulty is the computation of the posterior distribution of Z , Z^* and C .

$$\begin{aligned} p(C, Z, Z^* | \mathbf{f}, \mathbf{e}, \mathbf{w}, \mathbf{w}^*) \\ = \int p(\Delta, C, Z, Z^* | \mathbf{f}, \mathbf{e}, \mathbf{w}, \mathbf{w}^*) d\Delta \end{aligned} \quad (4)$$

where

$$\begin{aligned} p(\Delta, C, Z, Z^* | \mathbf{f}, \mathbf{e}, \mathbf{w}, \mathbf{w}^*) \\ = \frac{p(\Delta, C, Z, Z^*)}{\sum_{C, Z, Z^*} \int p(\Delta, C, Z, Z^*) d\Delta} \end{aligned} \quad (5)$$

where Δ denotes a collection of all parameters without hyperparameters. We create variational inference to address this issue because it is impossible to find a closed-form solution because of the integrals over the parameters.

The Main Idea

The main idea behind our variational inference method is to use Jensen's inequality to find the tightest lower bound of the original expressions.

We approximate the posterior by a novel variational distribution function q . According to the theory of variational optimization, the variational distribution q can be defined as

$$\begin{aligned} q(\pi, \rho, \theta, \eta, \varphi, \Omega, C, Z, Z^* | \tilde{\rho}, \tilde{\Omega}, \tilde{\alpha}, \tilde{\alpha}^*, \tilde{\beta}, \tilde{\xi}, \tilde{\pi}, \tilde{\theta}, \tilde{\theta}^*) \\ = \prod_{u=1}^U q(\pi_u | \tilde{\rho}_u) \prod_{c=1}^C q(\rho_c | \tilde{\Omega}_c) \prod_{c=1}^C q(\theta_c | \tilde{\alpha}_c) \prod_{c=1}^C q(\theta_c^* | \tilde{\alpha}_c^*) \prod_{k=1}^K q(\varphi_k | \tilde{\beta}_k) \\ \cdot \prod_{c=1}^C \prod_{c^*=1}^K q(\eta_{c,c^*} | \tilde{\xi}_{c,c^*}) \prod_{u=1}^U q(e_u | \tilde{\pi}_u) \prod_{u=1}^U \prod_{d=1}^{D_u} q(z_{u_d} | \tilde{\theta}_{u_d}) \prod_{u=1}^U \prod_{d^*=1}^{D_u^*} q(z_{u_d^*}^* | \tilde{\theta}_{u_d^*}^*) \end{aligned}$$

where $\tilde{\rho}, \tilde{\Omega}, \tilde{\alpha}, \tilde{\alpha}^*, \tilde{\beta}, \tilde{\xi}, \tilde{\pi}, \tilde{\theta}$ and $\tilde{\theta}^*$ are variational parameters. In order to obtain the closest variational distribution, we first need to find a way to measure the distance between the variational

distribution q and the posterior distribution p . We adopt the Kullback-Leibler (KL) divergence (Daminelli et al., 2015), which is defined as

$$\begin{aligned} & \text{KL}(q \parallel p) \\ &= E_q \log q(\Delta, C, Z, Z^* | \rho, \tilde{\Omega}, \tilde{\alpha}, \tilde{\alpha}^*, \tilde{\beta}, \tilde{\xi}, \tilde{\pi}, \tilde{\theta}, \tilde{\theta}^*) \\ & - E_q \log p(\Delta, C, Z, Z^* | \rho, \Omega, \alpha, \alpha^*, \beta, \xi, \mathbf{f}, \mathbf{w}, \mathbf{w}^*, \mathbf{e}) \end{aligned}$$

We note that model parameters and variational parameters make up the KL divergence, and our objective is to reduce it. Although the true posterior probability involved in the KL divergence cannot be measured directly, it is comparable to determining the greatest value of ELBO (Evidence Lower Bound). In order to address the computational challenge, we use the variational distribution to roughly approximate the optimal distribution. ELBO is therefore defined as

$$\begin{aligned} & \tilde{L}(\rho, \tilde{\Omega}, \tilde{\alpha}, \tilde{\alpha}^*, \tilde{\beta}, \tilde{\xi}, \tilde{\pi}, \tilde{\theta}, \tilde{\theta}^*; \rho, \Omega, \alpha, \alpha^*, \beta, \xi) \\ &= E_q \log p(\Delta, C, Z, Z^*, \mathbf{f}, \mathbf{w}, \mathbf{w}^*, \mathbf{e} | \rho, \Omega, \alpha, \alpha^*, \beta, \xi) \\ & - E_q \log q(\Delta, C, Z, Z^* | \rho, \tilde{\Omega}, \tilde{\alpha}, \tilde{\alpha}^*, \tilde{\beta}, \tilde{\xi}, \tilde{\pi}, \tilde{\theta}, \tilde{\theta}^*) \end{aligned} \quad (8)$$

The equivalence between these two optimization problems can then be easily derived, since these two objectives sum up to a constant for the given data. That is,

$$\text{KL}(q \parallel p) + \tilde{L} = \log p(\mathbf{f}, \mathbf{e}, \mathbf{w}, \mathbf{w}^*).$$

The derivatives of the variational parameters are taken and set to 0, which is characterized as

$$\nabla \tilde{L}(q) = \left(\frac{\partial \tilde{L}}{\partial \rho}, \frac{\partial \tilde{L}}{\partial \tilde{\Omega}}, \frac{\partial \tilde{L}}{\partial \tilde{\alpha}}, \frac{\partial \tilde{L}}{\partial \tilde{\alpha}^*}, \frac{\partial \tilde{L}}{\partial \tilde{\beta}}, \frac{\partial \tilde{L}}{\partial \tilde{\xi}}, \frac{\partial \tilde{L}}{\partial \tilde{\pi}}, \frac{\partial \tilde{L}}{\partial \tilde{\theta}}, \frac{\partial \tilde{L}}{\partial \tilde{\theta}^*} \right) = 0. \quad (9)$$

Ultimately all the parameters that need to be updated can be expressed as follows:

$$\tilde{\rho}_{uc} \propto \rho_c + \tilde{\pi}_{uc}, \quad (10)$$

$$\tilde{\Omega}_{cu^*} \propto \Omega_{u^*} + \sum_{u=1}^U \tilde{\pi}_{uc} f_{uu^*}, \quad (11)$$

$$\tilde{\alpha}_{ck} \propto \alpha_k + \sum_{u=1}^U \sum_{d=1}^{D_n} \tilde{\pi}_{uc} \theta_{udk}, \quad (12)$$

$$\tilde{\alpha}_{ck}^* \propto \alpha_k^* + \sum_{u=1}^U \sum_{d^*=1}^{D_n^*} \tilde{\pi}_{uc} \theta_{ud^*k}^*, \quad (13)$$

$$\tilde{\beta}_{ki} \propto \beta_i + \sum_{u=1}^U \sum_{d=1}^{D_n} \sum_{n=1}^{N_{ud}} \theta_{udk} w_{udn}^i + \sum_{u=1}^U \sum_{d^*=1}^{D_n^*} \sum_{n^*=1}^{N_{ud^*}} \theta_{udk} w_{ud^*n^*}^{*i}, \quad (14)$$

where w denotes the number of times that this word occurs in the document D_{ud} , corresponding to the i -th word in the word list, and w^* denotes the number of times that the word occurs in the document D_{ud^*} .

$$\widetilde{\xi}_{cc^*ks}^* \propto \xi_{cs} + \sum_{u=1}^U \sum_{d^*=1}^{D_u^*} \sum_{u^*=1}^U \widetilde{\pi}_{uc} \widetilde{\pi}_{u^*c^*} \theta_{ud^*k}^* e^{s_{ud^*u^*}}. \quad (15)$$

There is only one case where the e value is 1, that is, sentimental polarity of $D_{ud^*}^*$ is s .

$$\widetilde{\theta}_{udk} \propto \exp \left\{ \sum_{n=1}^{N_{ud}} \sum_{i=1}^V w_{udn}^i [\Psi(\widetilde{\beta}_{ki}) - \Psi \sum_{i'=1}^V (\widetilde{\beta}_{ki'})] + \sum_{c=1}^C \widetilde{\pi}_{uc} [\Psi(\widetilde{\alpha}_{ck}) - \Psi \sum_{k'=1}^K (\widetilde{\alpha}_{ck'})] \right\}, \quad (16)$$

$$\theta_{ud^*k}^* \propto \exp \left\{ \sum_{n^*=1}^{N_{ud^*}^*} \sum_{i=1}^V w_{ud^*n^*}^{*i} [\Psi(\widetilde{\beta}_{ki}) - \Psi \sum_{i'=1}^V (\widetilde{\beta}_{ki'})] + \sum_{c=1}^C \widetilde{\pi}_{uc} [\Psi(\widetilde{\alpha}_{ck}^*) - \Psi \sum_{k'=1}^K (\widetilde{\alpha}_{ck'}^*)] \right\}, \quad (17)$$

$$\widetilde{\pi}_{uc} \propto \exp \left[\begin{aligned} & \sum_{u^*=1}^U \widetilde{f}_{uu^*} [\Psi(\widetilde{\Omega}_{cu^*}) - \Psi \sum_{u'=1}^U (\widetilde{\Omega}_{cu'})] + [\Psi(\widetilde{\rho}_{uc}) - \Psi \sum_{c'=1}^C (\widetilde{\rho}_{uc'})] \\ & + \sum_{d=1}^{D_u} \sum_{k=1}^K \widetilde{\theta}_{udk} [\Psi(\widetilde{\alpha}_{ck}) - \Psi \sum_{k'=1}^K (\widetilde{\alpha}_{ck'})] + \sum_{d^*=1}^{D_u^*} \sum_{k=1}^K \theta_{ud^*k}^* [\Psi(\widetilde{\alpha}_{ck}^*) - \Psi \sum_{k'=1}^K (\widetilde{\alpha}_{ck'}^*)] \\ & + \sum_{d^*=1}^{D_u^*} \sum_{u^*=1}^U \sum_{c=1}^C \sum_{k=1}^K \sum_{s=1}^S \widetilde{\pi}_{u^*c^*} \theta_{ud^*k}^* e^{s_{ud^*u^*}} [\Psi(\widetilde{\xi}_{cc^*ks}^*) - \Psi \sum_{s'=1}^S (\widetilde{\xi}_{cc^*ks'}^*)] \end{aligned} \right], \quad (18)$$

where $\psi(\cdot)$ is the Digamma function, which is the logarithmic derivative of the Gamma function $\Gamma(\cdot)$, defined as

$$\psi(x) = \frac{d \log \Gamma(x)}{dx} = \frac{\Gamma'(x)}{\Gamma(x)}.$$

Algorithm Summary and Time Complexity

Finally, we summarize the above optimization process in Algorithm 1. For step 3 the times to update $\widetilde{\rho}, \widetilde{\Omega}, \widetilde{\alpha}, \widetilde{\alpha}^*, \widetilde{\beta}, \widetilde{\xi}$ are $O(|U||C|)$, $O(|U||F||C|)$, $O(|C||K||D|)$, $O(|C||K||D^*|)$, $O(|K||W| + |K||W^*|)$, and $O(|C^2||K||D^*|)$. Then, the times to compute $\widetilde{\theta}, \theta^*, \widetilde{\pi}$ in step 4 are $O(|K||W| + |U||C||K|)$, $O(|K||W^*| + |U||C||K|)$ and

$$O(|U||F||C| + |U||C| + |C||K||D| + |C^2||K||D^*|).$$

Thus, the overall time complexity is

$$O(|C^2||K||D^*| + |U||F||C| + |U||C||K| + |C||K||D| + |K||W| + |K||W^*|).$$

EXPERIMENTS

We evaluated the CSDD model on two large-scale real-world datasets and compared it with nine state-of-the-art baselines on a PC with Intel(R) 4.2GHz CPUs and 32GB RAM.

Algorithm 1. Inference algorithm of CSDD

Input: $U, D, D^*, \mathbf{w}, \mathbf{w}^*, \mathbf{f}, \mathbf{e}$, a threshold ε , $countmax$
Output: $\rho, \Omega, \alpha, \alpha^*, \beta, \xi, \theta, \theta^*, \pi$

1. Initialize θ, θ^*, π randomly and $t = 1$
2. **Repeat:**
3. (a). Update $\rho, \Omega, \alpha, \alpha^*, \beta, \xi$, via (10) - (15)
4. (b). Update and normalize θ, θ^*, π , via (16) - (18)
5. (c). Computing \tilde{L}^{count} and $t = t + 1$
6. **Until** $\tilde{L}(q^{count}) - \tilde{L}(q^{count-1}) < \varepsilon$ or $t > countmax$

Datasets

Our study used two Twitter datasets (Wang et al., 2020b), which were scraped from 147,909 users and spanned the months from December 2010 to July 2011. Only 147,909 users were crawled, since certain users’ privacy settings were unavailable.

We extracted two time periods from the long time span: March 1, 2011 to March 10, 2011 and May 1, 2011 to May 2, 2011. Among them, the iPad 2 was released on March 3, 2011, and Bin Laden was assassinated on May 1, 2011. We removed all stop words, non-English phrases, and repeated tweets written by the same user. Each word had to appear in at least three additional tweets, and each tweet had to contain at least three additional words. The statistical information from the two datasets is summarized in Table 2.

Sentimental Status Discovery

Thanks to the sentiment analysis method in Xiong et al. (2018), we could divide the large number of tweets into three different polar sentiments automatically, i.e., negative, neutral, and positive. Then each tweet was assigned a positive $R^+(w)$ and negative $R^-(w)$ sentimental score. Both scores were on a scale ranging between 1 (neutral) and 5 (strongly positive or negative). We defined the polarity score $R(w)$ as the tweet’s sentiment score, defined as

$$R(w) = R^+(w) - R^-(w).$$

If $R^+(w) = R^-(w)$, we considered the sentimental polarity of the tweet to be neutral. When $R(w) < 0$ or $R(w) > 0$, we considered the sentimental polarity of the tweet to be negative or positive, respectively.

Baselines

We chose nine state-of-the-art baselines to evaluate the community results, which can be divided into five categories, as listed below.

Table 2. Datasets

Datasets	No. of users	No. of tweets	No. of retweets	No. of words
Twitter 3	60588	234410	121659	138969
Twitter 5	13657	42541	22660	22806

Topology-Based Community Detection

These methods consider only the topology information and ignore the semantic, sentiment, and document diffusion information.

BIGCLAM (Yang & Lescovec, 2013): builds mainly on a novel observation that overlaps between communities are densely connected, which is especially effective for dense overlapping communities in large networks.

DCSBM (Karrer & Newman, 2011): detects community structure in social networks, considering variation in vertex degree, and uses only topology information.

Semantics-Based Community Detection

These methods consider topology and semantics information. They believe that communities should have semantic information, so users who write the same types of documents are more likely to be in the same community.

PMTLM (Zhu et al., 2013): generates topics using the Poisson distribution and takes into account only semantic information. We integrate the topic classification of each user's documents as the community this user belongs to.

GUCD (He et al., 2021): combines Graph Convolutional Networks and Markov Random Fields to generate community structure.

Sentiment-Based Community Detection

This method considers sentimental information. It considers users who have the same sentimental polarity about the same topic to be more likely to be in the same community.

JST (Lin & He, 2009): can learn the topic of the document. We use this model to get the topic according to the sentimental information that we use the same sentimental software to generate. For the results, we integrate the topic classification of each user's documents as the community this user belongs to.

Diffusion-Based Community Detection

These methods not only consider the relationships between friends but also the document diffusion edges. They also ignore sentiment-type information in the inputs.

COLD (Hu et al., 2015): uncovers temporal diffusion and extracts inter-community influence dynamics.

CPD (Cai et al., 2017): identifies a community in terms of both its internal content profile and its external diffusion profile and formalizes the concept of community profiling.

GHIPT (Wang et al., 2020a): can integrate group homophile and individual personality topics and generate the topic with hierarchical structure through document diffusion.

Ablation of CSDD

We investigated the effect of the sentiment diffusion component. We believe it is important in determining the performance of our approach on community detection and semantics.

CSDD without sentiment diffusion (CSDD-NS): removes the polarity of sentiment diffusion and considers the sentiment of all retweets to be neutral. The method ignores sentiment diffusion between communities.

Metrics

In light of the fact that no ground-truth is known regarding user communities in these two networks, in order to validate the community detection quality, we adopt *conductance* and *expansion* (Leskovec et al., 2010; Kloster & Gleich, 2014) as the metrics. The lower value of the score signifies a more community-like set of nodes. Fortunato and Hric (2016) defined a strong community as a subgraph

such that the internal degree of each vertex is greater than its external degree. This is why we chose conductivity and expansion as the metrics.

Conductance measures the fraction of total edge volume that points outside the community, defined as

$$f(S) = B_c / (2E_c + B_c).$$

Expansion measures the number of edges per node that point outside the community, defined as

$$f(S) = B_c / n_c.$$

In the two equations above, we let C be the set of nodes in the community, where n_c is the number of nodes in C , $n_c = |C|$; $E_c = |\{(v_i, v_j) : v_i \in C, v_j \in C\}|$ is the number of edges in C ; and $B_c = |\{(v_i, v_j) : v_i \in C, v_j \notin C\}|$ is the number of edges on the boundary of C .

Community Detection

Table 3 shows the *conductance* and *expansion* results of nine baselines and CSDD on the Twitter3 and Twitter5 datasets. The programs of all the methods compared were obtained from their authors, and we used their default.

- (1) Our approach (i.e., CSDD) outperforms the nine baselines, with the exception of CPD on *expansion*, because, while the method does not consider sentiments, it does consider the impact of different time periods on community detection. This impact will grow over a long period of time. Furthermore, expansion does not consider edges within the community, whereas CPD is better at identifying edges between communities because it pays attention to the user's personality, which is another reason.
- (2) Our approach outperforms JST. The reason for this is that the JST model divides communities in social networks only on the basis of sentiment, ignoring the fact that different communities diffuse different polarities of sentiment. Furthermore, because it relies solely on sentiment division, which ignores document content structure, its classification results are inferior to those obtained solely through semantic classification.
- (3) The ablation studies were carried out to answer the following questions: The effect of sentiment diffusion on the detection of communities; In *conductance* and *expansion*, the model's performance decreased by an average of 4.7% and 4.6%, demonstrating the importance of sentiment diffusion.
- (4) Results show that the CSDD model achieves 3.5% *conductance* decrease over the second-best baseline on Twitter 3, 1.8% *conductance* decrease, and 1.5% *expansion* decrease over the second-best baseline on Twitter5. By mining conflict and promotion at the community level, our algorithm outperforms baselines in most cases. This is largely due to the fact that our method takes into account not only the influence of sentiments within the community, but also sentiment diffusion across communities. Furthermore, CSDD is better at detecting sentiment-rich networks, such as networks created in response to specific events, especially sentiments that are easily aroused, making it easier to track sentimental resonance and capture real communities.

Table 3. Comparison in terms of conductance and expansion on 50 communities

Methods	Twitter3		Twitter5	
	Conductance	Expansion	Conductance	Expansion
BIGCLAM	0.9510	0.9836	0.9550	0.1864
DCSBM	N/A	N/A	0.9416	0.1795
JST	0.9152	0.9309	0.9171	0.1841
GUCD	0.8453	0.8597	0.8512	0.1552
PMTLM	N/A	N/A	0.8650	0.1627
COLD	0.9378	0.9930	0.9379	0.1876
CPD	0.8436	0.7928	0.8986	0.1656
GHIPT	0.8263	0.8311	0.8647	0.1603
CSDD-NS	0.8337	0.8365	0.8800	0.1647
CSDD	0.7976	0.8193	0.8358	0.1529

Note. Lower value of score signifies a more community-like set of nodes. N/A means out of memory or run times > 48 hours.

Finding Semantics

We extracted the model’s latent factors, showing its accurate modeling ability in both topics. We quantitatively evaluated CSDD’s topic extraction capacity by *perplexity* (Blei et al., 2003), which is a probability distribution over entire sentences or text. A low *perplexity* indicates that the model distribution is closer to the real distribution. *Perplexity* is defined as follows:

$$Perplexity(D) = \exp\left(-\frac{\sum_{d=1}^D \log(p(w_d))}{\sum_{d=1}^D N_d}\right), \quad (21)$$

where N_d is the length of each post d , and $p(w_d)$ is the probability of a series of words in each tweet: for the CSDD model, $p(w_d)$ can be denoted as

$$\log(p(w_d)) = \log\left(\sum_c \left(\sum_z (p(c|u))(p(z|c))(p(w_d|z))\right)\right). \quad (22)$$

We tested the result of word clustering under a different number of topics on the Twitter5 dataset. Table 4 reveals that CSDD has the lowest *perplexity* on 20 topics. Our method is consistently better than COLD and CPD across a number of topics. This is for two reasons: (1) It considers the homogeneity of diffusion, which makes it easier for the forwarded topic to segment sentiment words with different sentimental polarities; (2) The distribution of topic-word is applicable not only to the original topics but also to the forwarded topics.

A Case Study

We analyzed parameters modeled in CSDD on the Twitter5 dataset. They were community-topic distribution, topic-word distribution, and especially the distribution of sentiment diffusion, respectively (i.e., $\{\theta, \theta^*, \phi, \eta\}$).

Table 4. Perplexity values

Methods	No. of topics			
	5	10	20	50
COLD	24804	19673	19500	18067
CPD	198616	60012	54445	45161
CSDD-NS	22907	15330	15639	16088
CSDD	19136	12899	12713	13091

Note. Lower value of score signifies a better community-semantics.

Sentiment Diffusion Visualization

We visualized sentiment diffusion at the community level. After that, we used the default hyperparameters and set up 10 communities and 10 topics. We could now visualize not only how communities feature distinct content (e.g., what a community tweets), but also how the emotionalist interacts (e.g., how a community retweets others), which was often overlooked before.

In order to facilitate the analysis, we selected the top 6 communities and top 6 topics, and then drew a diffusion picture, selecting the top 13 sentimental edges. As shown in Figure 3, we can clearly see some laws of sentiment diffusion: (1) Compared with neutral and positive sentiment, the diffusion of negative sentiment tends to gather in the same community. (2) For different communities, the sentiment diffusion may be different. For example, the sentimental polarity of community c_4 diffuses to community c_6 is positive, but it diffuses to community c_3 is neutral. (3) The topic of diffusion depends mainly on the topic that the receiver cares about, followed by the topic that the disseminator cares about. (4) Both communities c_1 and c_6 are good at diffusing negative sentiment, and community c_4 is good at diffusing positive sentiment; e.g., a user with ID 15742760 in community c_6 retweeted the following from community c_4 : "fans and Mets fans rejoice together. Osama Bin Laden is dead." (5) We found that sharing the same topic makes it easier for communities to diffuse negative sentiments. As shown in Figure 3, both c_1 and c_6 tend to diffuse documents on social topics to c_2 , but the communication between them is negative.

Unbalance of Sentiment Diffusion

We find that sentiment diffusion is unbalanced for different communities; i.e., a lot of the sentiment diffusion occurred in a small fraction of communities. In the 100 negative sentiment diffusion edges at the community level, 20% of community interaction accounted for 55% of the total negative sentiment diffusion. 20% of community interaction accounted for 39% of the total neutral sentiment diffusion, and 20% of community interaction accounted for 47% of the total positive sentiment diffusion (see Table 5). This shows that certain categories of communities also contributed far more to sentiment diffusion of a certain polarity. In addition, our results can perceive the imbalance of sentiment diffusion, leading to almost outperforming nine baselines at community detection and topic extraction.

Word Clouds of Topics

Word clouds of four topics are illustrated in Figure 4, showing that each topic has practical meaning. We extracted the most frequently occurring words in each topic and adjusted the number of words according to the frequency of occurrence by analyzing the probability of topic-word distribution. We can deduce from Figure 4(a) that the words *blog*, *friend*, *facebook* and *twitter* are related to the topic of *social*, which is the main topic of community c_2 . Topic *terrorist*, as shown in Figure 4(b),

Figure 3. Visualization about sentiment diffusion

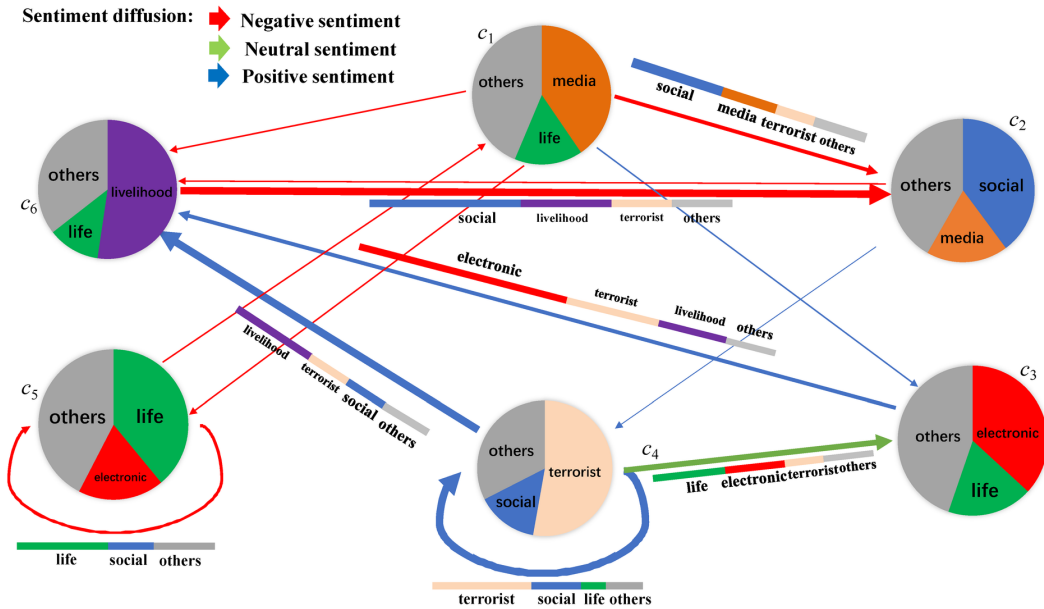
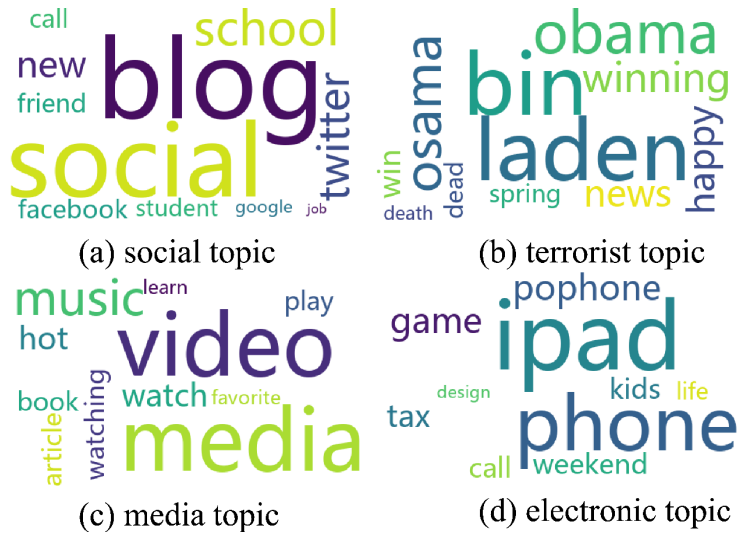


Table 5. Sentiment diffusion statistics at the community level

Sentimental polarity	Community interaction (%)								
	10	20	30	40	50	60	70	80	90
Negative	38	55	65	76	82	87	90	93	97
Neutral	21	39	51	67	81	89	92	97	98
Positive	30	47	62	76	85	90	95	98	99

is the main topic of the internal discussion of community c_4 and the second topic of the communication from community c_4 to community c_6 . Words *obama*, *bin*, *laden*, and *osama* are the names of people who are related to terrorists. The words *winning* and *happy* usually appear at the same time to describe the general views of social networks after the elimination of the Bin Laden group. *Dead* and *death* usually appear in neutral reports and positive reports and usually appear at the same time as words *news*. Similarly, because of the words *video*, *music*, and *watch*. Figure 4(c) is considered as a media topic, which is the main topic of community c_1 . Among these words, *video* and *music* are the specific forms of social media. Interestingly, *book* as a form of paper media appeared in this topic many times. The last example is shown in Figure 4(d), which is the main topic of community c_3 and the main topic of community c_3 diffusing to c_6 . Among the topics here, the topic of the *iPad*, *iPhone*, and *PoPhone* are electronic specific content. As a kind of products, electronic products are closely related to *tax*; at the same time, as a new technology product, electronic products are often used by students and children for entertainment, so we have mined the words *kids*, *life*, and *game* in *electronic* topic.

Figure 4. Word clouds of four topics: Social, terrorist, media, electronic



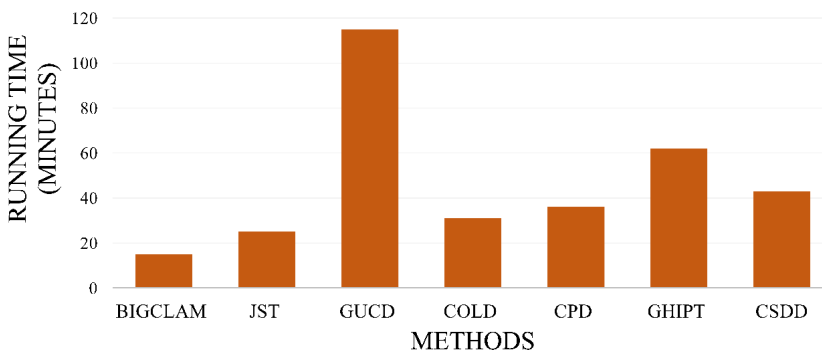
Parameter Initiation

Our method can choose any number of communities c and topics k . However, for clarity and simplicity, we focused on the case of $c = k$. All the hyperparameters were fixed at a predefined value. In this research, we followed the convention in Bayesian statistics and set the value as ($\alpha = 0.01$, $\alpha^* = 0.05$, $\beta = 0.001$, $\Omega = 0.05$, $\rho = 0.05$, $\xi = 0.001$; Wang et al., 2019).

Running Time

The actual running time is shown in Figure 5. In the Twitter5 dataset, we recorded the running time of CSDD as well as baselines. Our method isn't the quickest, because it takes into account not just the topology and content of the documents, but also sentimental interaction factors. Because we consider three-dimensional sentiment diffusion on the basis of diffusion, the model's running time is slightly faster than that of CPD. The model's execution time is slightly longer than GHIPT, because, while it does not consider sentimental diffusion, it does examine the topic's vertical relevance. Furthermore, we employed Gibbs sampling for approximate inference, which takes longer than variational inference.

Figure 5. Comparison in terms of running time



CONCLUSION AND DISCUSSIONS

In this paper, we studied the problem of discovering communities that are more accurate and explainable. We further studied the regulation of sentiment diffusion based on communities. However, the difficulty of model configuration in this situation is largely underestimated. Here, we have solved three aspects of this problem, namely the following: (1) clearly distinguishing sentiment polarity; (2) heterogeneity in social networks; and (3) the uniqueness of the topic. Our work explains for the first time the mechanism of sentiment diffusion at the community level and investigates the sentiment diffusion mechanism in real social networks, revealing the impact of topics on different sentiments. We also discovered an important phenomenon: that the diffusion of negative sentiments tends to concentrate in a small number of communities.

The limitation of our model is that it must obtain information on sentiment diffusion. Although considering the factors of emotional communication will more accurately detect communities, many network structures do not contain such information—for example, citation networks.

Graph neural networks (GNN) have been widely used in recent years because of their powerful representation ability. GNN is also used to discover communities (Bruna & Li, 2017; Ma et al., 2020; Sun et al., 2021). However, because their resulting representations are difficult to interpret, we will next combine GNN and topic models for community detection.

ACKNOWLEDGMENT

This research is supported by the Natural Science Foundation of Heilongjiang Province of China (Grant No. LH2023E049) and the School of Architecture and Design, Harbin Institute of Technology, Key Laboratory of Cold Region Urban and Rural Human Settlement Environment Science and Technology, Ministry of Industry and Information Technology.

REFERENCES

- Balasubramanyan, R., & Cohen, W. W. (2011, April). Block-LDA: Jointly modeling entity-annotated text and entity-entity links. In B. Liu, H. Liu, C. Clifton, T. Washio, & C. Kamath (Eds.), *Proceedings of the 2011 SIAM International Conference on Data Mining* (pp. 450-461). Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611972818.39
- Bickel, P. J., & Sarkar, P. (2016). Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 78(1), 253–273. doi:10.1111/rssb.12117
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.
- Bruna, J., & Li, X. (2017). Community detection with graph neural networks. *Stat*, 1050, 27.
- Cai, H., Zheng, V. W., Zhu, F., Chang, K. C. C., & Huang, Z. (2017). From community detection to community profiling. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 10(7), 817–828. doi:10.14778/3067421.3067430
- Chen, H., Yin, H., Li, X., Wang, M., Chen, W., & Chen, T. (2017). People opinion topic model: Opinion based user clustering in social networks. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 1353–1359). International World Wide Web Conferences Steering Committee doi:10.1145/3041021.3051159
- Cheng, J., Li, L., Yang, H., Li, Q., & Chen, X. (2018, July). A hybrid spectral method for network community detection. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data* (pp. 90-104). Springer, Cham. doi:10.1007/978-3-319-96890-2_8
- Daminelli, S., Thomas, J. M., Durán, C., & Cannistraci, C. V. (2015). Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks. *New Journal of Physics*, 17(11), 113037. doi:10.1088/1367-2630/17/11/113037
- Feng, J., Rao, Y., Xie, H., Wang, F. L., & Li, Q. (2020). User group based emotion detection and topic discovery over short text. *World Wide Web (Bussum)*, 23(3), 1553–1587. doi:10.1007/s11280-019-00760-3
- Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659, 1–44. doi:10.1016/j.physrep.2016.09.002
- Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12), 7821–7826. doi:10.1073/pnas.122653799 PMID:12060727
- Han, Y., & Tang, J. (2015). Probabilistic community and role model for social networks. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 407–416). Association for Computing Machinery. doi:10.1145/2783258.2783274
- He, D., Song, Y., Jin, D., Feng, Z., Zhang, B., Yu, Z., & Zhang, W. (2021). Community-centric graph convolutional network for unsupervised community detection. In C. Bessiere, *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on artificial intelligence* (pp. 3515–3521).
- Hu, Z., Yao, J., Cui, B., & Xing, E. (2015). Community level diffusion extraction. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (pp. 1555–1569). Association for Computing Machinery. doi:10.1145/2723372.2723373
- Jia, S., Gao, L., Gao, Y., Nastos, J., Wang, Y., Zhang, X., & Wang, H. (2015). Defining and identifying cograph communities in complex networks. *New Journal of Physics*, 17(1), 013044. doi:10.1088/1367-2630/17/1/013044
- Jiang, Y., Huang, X., Cheng, H., & Yu, J. X. (2018). Vizcs: Online searching and visualizing communities in dynamic graphs. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)* (pp. 1585–1588). IEEE.
- Karrer, B., & Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review. E*, 83(1), 016107. doi:10.1103/PhysRevE.83.016107 PMID:21405744
- Kloster, K., & Gleich, D. F. (2014). Heat kernel based community detection. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge discovery and Data Mining* (pp. 1386–1395). Association for Computing Machinery. doi:10.1145/2623330.2623706

- Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference* (pp. 933–943). International World Wide Web Conferences Steering Committee. doi:10.1145/3178876.3186141
- Leskovec, J., Lang, K. J., & Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 631–640). Association for Computing Machinery. doi:10.1145/1772690.1772755
- Li, Y., He, K., Bindel, D., & Hopcroft, J. E. (2015). Uncovering the small community structure in large networks: A local spectral approach. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 658–668). International World Wide Web Conferences Steering Committee. doi:10.1145/2736277.2741676
- Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 375–384). Association for Computing Machinery. doi:10.1145/1645953.1646003
- Ma, J., Cui, P., Wang, X., & Zhu, W. (2018). Hierarchical taxonomy aware network embedding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1920–1929). Association for Computing Machinery. doi:10.1145/3219819.3220062
- Ma, Y., Guo, Z., Ren, Z., Tang, J., & Yin, D. (2020). Streaming graph neural networks. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 719–728). Association for Computing Machinery.
- Neville, J., Adler, M., & Jensen, D. (2003, August 9–15). *Clustering relational data using attribute and link information* [conference presentation]. Text Mining and Link Analysis Workshop, 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico. <https://www.cs.cmu.edu/~dunja/TextLink2003/Papers/NevilleTextLink03.pdf>
- Newman, M. E. (2016). Equivalence between modularity optimization and maximum likelihood methods for community detection. *Physical Review. E*, 94(5), 052315. doi:10.1103/PhysRevE.94.052315 PMID:27967199
- Newman, M. E., & Leicht, E. A. (2007). Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(23), 9564–9569. doi:10.1073/pnas.0610537104 PMID:17525150
- Peel, L. (2011). Topological feature based classification. In *14th International Conference on Information Fusion* (pp. 1–8). IEEE.
- Peel, L. (2012). Supervised blockmodelling. <https://arxiv.org/abs/1209.5561>
- Sun, J., Zheng, W., Zhang, Q., & Xu, Z. (2021). Graph neural network encoding for community detection in attribute networks. *IEEE Transactions on Cybernetics*, 52(8), 7791–7804. doi:10.1109/TCYB.2021.3051021 PMID:33566785
- Tu, K., Cui, P., Wang, X., Yu, P. S., & Zhu, W. (2018). Deep recursive network embedding with regular equivalence. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2357–2366). Association for Computing Machinery. doi:10.1145/3219819.3220068
- Vlaic, S., Conrad, T., Tokarski-Schnelle, C., Gustafsson, M., Dahmen, U., Guthke, R., & Schuster, S. (2018). ModuleDiscoverer: Identification of regulatory modules in protein-protein interaction networks. *Scientific Reports*, 8(1), 1–11. doi:10.1038/s41598-017-18370-2 PMID:29323246
- Wang, D., Li, J., Xu, K., & Wu, Y. (2017). Sentiment community detection: Exploring sentiments and relationships in social networks. *Electronic Commerce Research*, 17(1), 103–132. doi:10.1007/s10660-016-9233-8
- Wang, X., Jin, D., Musial, K., & Dang, J. (2020b, April). Topic enhanced sentiment spreading model in social networks considering user interest. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(1), 989–996. doi:10.1609/aaai.v34i01.5447
- Wang, Y., Jin, D., Musial, K., & Dang, J. (2019). Community detection in social networks considering topic correlations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 321–328. doi:10.1609/aaai.v33i01.3301321 PMID:32219010
- Wang, Y., Jin, D., Yang, C., & Dang, J. (2020a). Integrating group homophily and individual personality of topics can better model network communities. In *2020 IEEE International Conference on Data Mining (ICDM)* (pp. 611–620). IEEE. doi:10.1109/ICDM50108.2020.00070

- Xie, Y., Gong, M., Wang, S., & Yu, B. (2018, September). Community discovery in networks with deep sparse filtering. *Pattern Recognition, 81*, 50–59. doi:10.1016/j.patcog.2018.03.026
- Xiong, X., Li, Y., Qiao, S., Han, N., Wu, Y., Peng, J., & Li, B. (2018, January 15). An emotional contagion model for heterogeneous social media with multiple behaviors. *Physica A, 490*, 185–202. doi:10.1016/j.physa.2017.08.025
- Yang, J., & Leskovec, J. (2013). Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (pp. 587–596). Association for Computing Machinery. doi:10.1145/2433396.2433471
- Yang, L., Cao, X., He, D., Wang, C., Wang, X., & Zhang, W. (2016). Modularity based community detection with deep learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 2252–2258). AAAI Press.
- Yang, T., Jin, R., Chi, Y., & Zhu, S. (2009). Combining link and content for community detection: A discriminative approach. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 927–936). Association for Computing Machinery. doi:10.1145/1557019.1557120
- Zhang, P., & Moore, C. (2014). Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. *Proceedings of the National Academy of Sciences of the United States of America, 111*(51), 18144–18149. doi:10.1073/pnas.1409770111 PMID:25489096
- Zhang, Y., Xiong, Y., Ye, Y., Liu, T., Wang, W., Zhu, Y., & Yu, P. S. (2020). SEAL: Learning heuristics for community detection with generative adversarial networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1103–1113). Association for Computing Machinery. doi:10.1145/3394486.3403154
- Zhe, C., Sun, A., & Xiao, X. (2019). Community detection on large complex attribute network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery. doi:10.1145/3292500.3330721
- Zhu, Y., Yan, X., Getoor, L., & Moore, C. (2013, August). Scalable text and link analysis with mixed-topic link models. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge discovery and Data Mining* (pp. 473–481). Association for Computing Machinery. doi:10.1145/2487575.2487693

Bingdao Feng is currently an Eng. D student of the College of Intelligence and Computing, Tianjin University, China. He received his M. S. degrees from Tianjin University, China. His research interests are mainly related to community detection, data mining and graph neural networks.

Fangyu Cheng is a lecturer at the School of Architecture and Design, Harbin Institute of Technology. His academic interests are rooted in spatial phenomenology, exploring reception aesthetics and embodied experiences, and the broad media application of environmental narrative regarding the regional cultural context. He holds an MA with distinction and a PhD in spatial design, obtained by Chelsea College of Arts, University of the Arts London.

Yanfei Liu is currently an Eng. D student of the College of Intelligence and Computing, Tianjin University, China. He received his M. S. degrees from ChongQing University of Technology, China. His research interests are mainly related to community detection, social network analysis and machine learning.

Xinglong Chang received his M. S. degrees from Tianjin University, China. His research interests are mainly related to pattern recognition and machine learning.

Xiaobao Wang received his Ph.D. degree in computer science from Tianjin University, China, in 2022. He is an assistant professor with the College of Intelligence and Computing, at Tianjin University, China. He has authored or co-authored 20 top-tier journal and conference papers. His current research interests include data mining and network analysis.

Di Jin received the Ph.D. degree in computer science from Jilin University, Changchun, China, in 2012. He was a research scholar in DMG at UIUC from 2019 to 2020. He is currently a Professor at the College of Intelligence and Computing, Tianjin University, Tianjin, China. His research interests include graph data mining and graph machine learning, especially on community detection, network embedding, and GNNs. To date, he has published more than 100 research papers in top-tier journals and conferences, including the TKDE, TNNLS, TYCB, AAAI, IJCAI, NeurIPS, and WWW. He was the recipient of the Best Paper Award Runner-up of WWW 2021, the Best Student Paper Award Runner-up of ICDM 2021, and the Rising Star Award of ACM Tianjin in 2018.