A Novel Generative Topic Embedding Model by Introducing Network Communities

Di Jin
College of Intelligence and Computing, Tianjin University, China
jindi@tju.edu.cn

Jiantao Huang
College of Intelligence and Computing, Tianjin University, China
huangjt@tju.edu.cn

Pengfei Jiao
Center for Biosafety Research and Strategy, Tianjin University, China
College of Intelligence and Computing, Tianjin University, China
pjiao@tju.edu.cn

Liang Yang*
School of Artificial Intelligence, Hebei University of Technology, China
yangliang@vip.qq.com

Dongxiao He
College of Intelligence and Computing, Tianjin University, China
hedongxiao@tju.edu.cn

Françoise Fogelman-Souliè
College of Intelligence and Computing, Tianjin University, China
francoise.soulie@outlook.com

Yuxiao Huang
Data Science, Columbian College of Arts & Sciences, George Washington University, USA
yuxiaohuang@gwu.edu

ABSTRACT

Topic models have many important applications in fields such as Natural Language Processing. Topic embedding modelling aims at introducing word and topic embeddings into topic models to describe correlations between topics. Existing topic embedding methods use documents alone, which suffer from the topical fuzziness problem brought by the introduction of embeddings of semantic fuzzy words, e.g. polysemous words or some misleading academic terms. Links often exist between documents which form document networks. The use of links may alleviate this semantic fuzziness, but they are sparse and noisy which may meanwhile mislead topics. In this paper, we utilize community structure to solve these problems. It can not only alleviate the topical fuzziness of topic embeddings since communities are often believed to be topic related, but also can overcome the drawbacks brought by the sparsity and noise of networks (because community is a high-order network information). We give a new generative topic embedding model which incorporates documents (with topics) and network (with communities) together, and uses probability transition to describe the relationship between topics and communities to make it robust when topics and communities do not match. An efficient variational inference algorithm is then proposed to learn the model. We validate the superiority of our new approach on two tasks, document classifications and visualization of topic embeddings, respectively.

*corresponding author. Email: yangliang@vip.qq.com

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313623

CCS CONCEPTS

• **Information systems → Document topic models**.

KEYWORDS

Document networks; Topic embedding; Community structure

ACM Reference Format:
Di Jin, Jiantao Huang, Pengfei Jiao, Liang Yang, Dongxiao He, Françoise Fogelman-Souliè, and Yuxiao Huang. 2019. A Novel Generative Topic Embedding Model by Introducing Network Communities. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313623

1 INTRODUCTION

Topic models, such as Latent Dirichlet Allocation (LDA) [3] and Probability Latent Semantic Analysis (PLSA) [11], have a variety of applications in areas such as Natural Language Processing (NLP). A limitation of traditional topic models is that, they have not considered the correlation between topics. Some studies have made effort to characterize topic correlations. For example, correlated topic models [1, 2, 5] replace Dirichlet distribution with logistic normal distribution, and topic embedding models [19] represent topics in a low-dimension embedding space.

Compared to correlated topic models, topic embedding modelling is a straightforward and efficient alternative with lower computational cost [9, 14]. In most topic embedding models, word embedding (instead of the traditional one-hot encoding for words) is applied, to describe the semantic correlations between words. Topic embedding is used by replacing the original word distribution of topics, to represent topics in a continuous low-dimensional space. A link function is then defined to connect word embeddings to topic embeddings to describe the relationship between words and

topics. The main merit of topic embedding models is their capacity for capturing words co-occurrence at different semantic levels to derive correlations between topics while they still have problems. On the one hand, there are often semantic fuzzy words in documents, which have ambiguous semantic meanings. E.g., the polysemous word 'apple' can mean a kind of fruit or a mobile phone brand, and the academic terms 'parallel' and 'efficient' belonging to high-performance computing area may be also included in a paper of artificial intelligence. The existence of these words affects the accurate semantic representation of word embeddings. On the other hand, word embeddings also influence topic representations since they are connected by link functions. Thus, the semantic fuzziness of word embeddings (brought by the existence of semantic fuzzy words) may also lead to the semantic fuzziness of topics in topic embedding models.

At the same time, links are ubiquitous among documents [27, 28]. E.g., papers are linked via citations in DBLP and webpages are connected via hyperlinks in Wikipedia. The linked documents can be represented as a document network. Different from documents themselves that capture objective semantics, network structure denoted by connections between documents often provides subjective semantics which may complement document contents. Then, a pair of linked documents should have similar topic distributions [13, 26]. By doing so, networks may be able to help mitigate the semantic fuzziness of topics. That is to improve topic embeddings and topic distributions of documents, which often suffer from the semantic fuzziness due to word embeddings of semantic fuzzy words. However, we also find that using links directly may not perform well on improving topics since this low-order network information is often sparse and noisy [16, 22, 24]. That is, some nearby documents may not be connected due to the sparsity of networks, and meanwhile, several uncorrelated documents may be connected due to noisy links. In special cases, this may even destroy the quality of topics.

Fortunately, community structure, as a type of high-order network information, may be suitable to solve this problem. A community in a document network consists of documents which are connected more tightly than those in different communities. Also, it is often believed that documents in the same community are often topic-related [10, 12, 30]. So, rather than using links directly, one may have a good reason to use communities, to make documents in a same community own similar topic distributions. By doing so, the problem brought by sparsity and noise of networks may be nicely alleviated.

To be specific, we present a generative topic embedding model, namely Community-Enhanced Topic Embedding (CeTe), which incorporates documents and network structure together. CeTe consists of two main components connected by a probability transition mechanism. The first is the document component which incorporates word embeddings into topic modelling to describe correlations between topics and capture the local word co-occurrence. The second is the network component in which communities are described based on stochastic blockmodels [17]. It plays the role of alleviating the problem of semantic fuzziness of topics with the help of topic-related network communities. We further use probability transition to describe the intrinsic relationship between topics and communities. In this way, the model will work robustly even when the topics (from documents) and communities (from networks) are not well matched. We finally give an efficient variational inference algorithm to learn the model.

The contributions of this paper are as follows:

- Topic embedding models suffer from the semantic fuzziness of topics brought by the semantic fuzziness of word embeddings caused by the existence of semantic fuzzy words. The low-order link information is not suitable to be used directly for this problem due to the sparsity and noise of networks. To the best of our knowledge, this is the first time high-order network information (i.e., communities that are topic related) is used to solve this problem.
- We give a novel generative topic embedding model by using documents with topics and network structure with communities together. Their relationship is further described by using probability transition to make this new model robust even when topics and communities do not match well. We derive an efficient variational inference algorithm to learn this model.
- The superior performance of CeTe is tested on two tasks, i.e. document classification by comparing with seven state-of-the-art methods and a topic visualization application.

2 THE MODEL

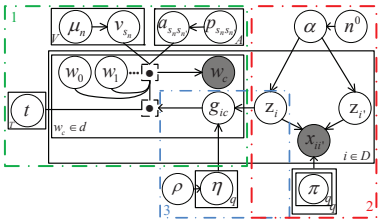

Figure 1: Graphical representation of CeTe. Part 1 in the green box (top left) denotes a document component describing topics. Part 2 in the red box (on the right) denotes a topological component describing network communities. Part 3 in the blue box (on the bottom) denotes the probabilistic transition mechanism connecting these two parts.

In this section, we give a formal description of the proposed model, i.e., Community Enhanced Topic Embedding (CeTe), with the purpose of improving document representation (topic distributions of documents) and topic embedding via introducing communities in networks. The graphical representation of this CeTe model is shown in Figure 1.

As shown in Figure 1, CeTe contains three components:

Document component with topics. This component depicts the generative process of each word given its topic and context words. In a set of documents $D = \{d_1, ..., d_N\}$ with K topics, each document d_i has L_i words, and the words in all documents form the vocabulary set $S = \{s_1, ..., s_W\}$. Each word w_{ij} in document d_i is assigned to a topic indexed by g_{ij}, where $\mathbf{G} = (g_{ij})_{N \times L_i}$ denotes the matrix of topics. Besides topics, the observed word is also affected by its context words. The context of a focus word w_c includes the c words before w_c in the document, denoted by $w_0 : w_{c-1}$. Both the topic and context words are represented in the

embedding vector form. That is, the embedding of word $w_{c'}$ and topic $g_{c'}$ are represented as $v_{c'}$ and $t_{g_{c'}}$, and we assume that their dimensions are the same. Then, matrices of all word embeddings and topic embeddings are $\mathbf{V} = (\mathbf{v}_{s_1},...,\mathbf{v}_{s_W})$ and $\mathbf{T} = (\mathbf{t}_1,...,\mathbf{t}_K)$. Thus the conditional distribution of a word given its context and topic can be factorized into two parts. The first is the distribution of this word and its context $P(w_c|w_0 : w_{c-1})$. It corresponds to the link function of a word embedding method such as PSDVec [20]. The distribution is:

$$P(w_c|w_0 : w_{c-1}) \approx P(w_c) \cdot exp\{\mathbf{v}_{w_c}^T \cdot \sum_{c'=0}^{c-1} \mathbf{v}_{w_{c'}} + \sum_{c'=0}^{c-1} a_{w_{c'}w_c}\}, \tag{1}$$

where $a_{w_{c'}w_c}$ is the bigram residual, which is non-linear and cannot be captured by $\mathbf{v}_{w_c}^T \mathbf{v}_{w_{c'}}$, and $P(w_c)$ is the probability of w_c appearing in the word corpus.

The second part is the distribution of the word and its topic. The topic can be also taken as a latent word, so that we can get the following distribution:

$$P(w_c|g_c) \approx P(w_c) \cdot exp\{\mathbf{v}_{w_c}^T \mathbf{t}_{g_c} + r_{g_c}\}, \tag{2}$$

where r_{g_c} is the logarithm of the normalized constant, which we call the topic residual. All topic residuals form a matrix $\mathbf{r}$=(r_1,...,r_K).

Since $\sum_{w_c \in S} P(w_c|k) = 1$ we can substitute it into Eq. (2) to approximate the r_k as:

$$r_k = -log\Big(\sum\nolimits_{s_j \in S} P(s_j) exp\{\mathbf{v}_{s_j}^T \mathbf{t}_k\}\Big). \tag{3}$$

Eq. (3) can be also represented in matrix form:

$$\mathbf{r} = -log\Big(\mathbf{u} exp\{\mathbf{V}^T \mathbf{T}\}\Big), \tag{4}$$

where $\mathbf{u}$ is the row vector of unigram probabilities.

Finally, according to Eqs. (1) and (2), we give the following link distribution for a document d_i:

$$\begin{aligned} &P(w_c|w_0 : w_{c-1}, g_c, d_i) \\ &\approx P(w_c) \cdot exp\{\mathbf{v}_{w_c}^T \cdot (\sum_{c'=0}^{c-1} \mathbf{v}_{w_{c'}} + \mathbf{t}_{g_c}) + \sum_{c'=0}^{c-1} a_{w_{c'}w_c} + r_{g_c}\}, \end{aligned} \tag{5}$$

where we simplify w_{ij} as w_c and $\mathbf{t}_{g_{ij}}$ as $\mathbf{t}_{g_c}$. Note that in order to avoid overfitting and alleviate the negative effect on topics, we constrain the magnitudes of all topic embeddings in a hyperball of radius λ.

Topological component with communities. This component describes the community structure that is often topic related to help model topics. Let a network have N documents which consist of Q communities. The adjacency matrix $\mathbf{X} = (x_{ii'})_{N\times N}$ is used to represent connections between documents. First, we generate $\boldsymbol{\alpha}$=(α_1,...,α_Q) by Dirichlet distribution, where α_q denotes the proportion of documents belonging to community q, and then get the following distribution:

$$P(\boldsymbol{\alpha}|\mathbf{n}^0) = [\Gamma(\sum\nolimits_{q=1}^{Q} n_q^0)/\prod\nolimits_{q=1}^{Q} \Gamma(n_q^0)] \prod\nolimits_{q=1}^{Q} \alpha^{n_q^0 - 1},$$

where $\mathbf{n}^0$=(n_1^0,...,n_Q^0) is the hyper-parameter and $\Gamma(\cdot)$ denotes the Gamma function.

Next, we use the Multinomial distribution to sample z_i in $\mathbf{Z} = (z_1, ..., z_N)$, where z_i is the community label of document d_i. Then, the distribution is defined as:

$$P(\mathbf{Z}|\boldsymbol{\alpha}) = \prod\nolimits_{i=1}^{N} P(z_i|\boldsymbol{\alpha}) = \prod\nolimits_{i=1}^{N} \alpha_{z_i}.$$

Finally, given the sampled community labels of documents d_i and d_j, i.e., z_i and $z_{i'}$, we sample $x_{ii'}$ from a Bernoulli distribution, defined as:

$$P(\mathbf{X}|\boldsymbol{\pi}, \mathbf{Z}) = \prod\nolimits_{i<i'} (deg_i deg_{i'} \pi_{z_i z_{i'}})^{x_{ii'}} (1 - deg_i deg_{i'} \pi_{z_i z_{i'}})^{1-x_{ii'}}.$$

The distribution above is based on the degree-corrected stochastic block model [17]. Here, $\boldsymbol{\pi} = (\pi_{ql})_{Q\times Q}$ is the block matrix where π_{ql} denotes the connection probability between nodes from communities q and l, and deg_i is the degree of document d_i.

Probability transition connecting topics and communities. This component describes the transition process from communities to topics to make the model work even when communities and topics do not match well. First, the probability transition matrix $\mathbf{H} = (\eta_{qk})_{Q\times K}$ is generated by a Dirichlet distribution, where η_{qk} represents the probability that document d_i is in the k-th semantic topic given it belongs to the q-th community. This distribution is defined as:

$$P(\mathbf{H}|\boldsymbol{\rho}) = \prod\nolimits_{q=1}^{Q} [\Gamma(\sum\nolimits_{k=1}^{K} \rho_k)/\prod\nolimits_{k=1}^{K} \Gamma(\rho_k)] \prod\nolimits_{k=1}^{K} \eta_{ik}^{\rho_k - 1},$$

where $\boldsymbol{\rho}$ is the hyper-parameter.

Next, given the community label z_i of document d_i, we use a Multinomial distribution to sample g_{ij}, defined as:

$$P(\mathbf{G}|\mathbf{H}, \mathbf{Z}) = \prod\nolimits_{i=1}^{N} \prod\nolimits_{j=1}^{L_i} \eta_{z_i g_{ij}}.$$

After finishing the definition of these three components, a complete generative process of this model is shown in the following. Note that, word embedding $\mathbf{v}_{s_n}$ is drawn from a Gaussian distribution as well as bigram residuals $a_{s_n s_{n'}}$. But for clarity, here we only focus on modelling topic embeddings while ignoring the generative process of word embeddings.

1. Choose $\boldsymbol{\alpha} \sim Dir(\boldsymbol{n^0})$
2. For each community q: Choose $\boldsymbol{\eta_q} \sim Dir(\boldsymbol{\rho})$
3. For the k-th topic: Choose $\mathbf{t}_k \sim Unif(B_\lambda)$
4. For each document d_i:
   (a) Draw community assignment $z_i \sim Mult(\boldsymbol{\alpha})$
   (b) For each document $d_{i'}$ with $i' > i$:
   Draw link $x_{ii'} \sim Bernoulli(deg_i deg_{i'} \pi_{z_i z_{i'}})$
   (c) For the j-th word:
   (i) Draw topic assignment $g_{ij} \sim Mult(\boldsymbol{\eta}_{z_i})$
   (ii) Draw word w_{ij} from S according to $P(w_{ij}|w_{i,j-c}, w_{i,j-1}, g_{ij}, d_i)$

where Dir, $Unif$ and $Mult$ are the Dirichlet, Uniform and Multinomial distributions respectively.

Then, the complete data likelihood is:

$$\begin{aligned} &P(D, \mathbf{X}, \mathbf{A}, \mathbf{V}, \mathbf{T}, \mathbf{G}, \mathbf{Z}, \mathbf{H}, \boldsymbol{\alpha}|\boldsymbol{\pi}, \mathbf{n}^0, \boldsymbol{\rho}, \lambda, \boldsymbol{\mu}) \\ &= \prod_{n=1}^{W} P(\mathbf{v}_{s_n}; \mu_n) \prod_{n,n'=1}^{W,W} P\Big(a_{s_n s_{n'}}; f(p_{s_n s_{n'}})\Big) \prod_{k}^{K} Unif(B_\lambda) \\ &\cdot \prod_{i=1}^{N} \prod_{j=1}^{L_i} \Big(Mult(\boldsymbol{\eta}_{z_i}) \cdot P(d_i|\mathbf{V}, \mathbf{A}, \mathbf{T}, \mathbf{G})\Big) \prod_{i=1}^{N} Mult(z_i|\boldsymbol{\alpha}) \\ &\cdot \prod_{q=1}^{Q} Dir(\boldsymbol{\eta}_q|\boldsymbol{\rho}) \cdot Dir(\boldsymbol{\alpha}|\mathbf{n}^0) \prod_{i<i'} P(x_{ii'}|z_i, z_{i'}, deg_i deg_{i'} \pi_{z_i z_{i'}}). \end{aligned}$$

In the steps above, $P(\mathbf{v}_{s_n}; \mu_n)$ and $P(a_{s_n s_{n'}}; f(p_{s_n s_{n'}}))$ are both defined as Gaussian priors, and $P(d_i|\mathbf{V}, \mathbf{A}, \mathbf{T}, \mathbf{G})$ is a simple expression of $P(w_{ij}|w_{i,j-c}, w_{i,j-1}, g_{ij}, d_i)$.

3 MODEL OPTIMIZATION

In this section, we aim at learning the embeddings of words and topics $\{\mathbf{V}, \mathbf{T}\}$, model parameters $\{\mathbf{H}, \boldsymbol{\alpha}\}$, and latent variables $\{\mathbf{G}, \mathbf{Z}\}$. Due to the coupling between word-related variables $\{\mathbf{V}, \mathbf{A}\}$ and topic-related variables $\{\mathbf{T}, \mathbf{G}, \mathbf{Z}, \mathbf{H}\}$, their simultaneous optimization is too time consuming in general. So as was done by other topic embedding methods [19], we also use a word embedding method, i.e. PSDVec [20], to learn $\{\mathbf{V}, \mathbf{A}\}$ first, and then optimize $\{\mathbf{T}, \mathbf{G}, \mathbf{Z}, \mathbf{H}, \boldsymbol{\alpha}\}$ with the fixed $\{\mathbf{V}^*, \mathbf{A}^*\}$. But this is also a non-trivial task. For this optimization problem, we propose a variational expectation-maximization (EM) algorithm that includes variational inference and parameter maximization, respectively.

3.1 Variational Inference

In this step, we first fix $\mathbf{T}$ as the constant $\mathbf{T}^*$ and then maximize the posterior $P(\mathbf{G}, \mathbf{H}, \mathbf{Z}, \boldsymbol{\alpha}|D, \mathbf{X}, \boldsymbol{\pi}, \mathbf{T}^*, \mathbf{A}^*, \mathbf{V}^*)$ via variational inference. First, we define the following variational distributions:

$$q(\boldsymbol{\alpha}, \mathbf{H}, \mathbf{Z}, \mathbf{G}) = q(\boldsymbol{\alpha})q(\mathbf{H})q(\mathbf{Z})q(\mathbf{G}),$$

where the variables are mutually independent and their distributions are specified as follows:

$$q(\boldsymbol{\alpha}) = \prod_{q=1}^{Q} Dir(\alpha_q|n_q), \quad q(\mathbf{H}) = \prod_{q=1}^{Q}\prod_{k=1}^{K} Dir(\eta_{qk}|\gamma_{qk}),$$
$$q(\mathbf{Z}) = \prod_{i=1}^{N}\prod_{q=1}^{Q} Mult(z_{iq}|\tau_{iq}), \quad q(\mathbf{G}) = \prod_{i=1}^{N}\prod_{j=1}^{L_i} Mult(g_{ij}|\boldsymbol{\beta}_{ij}).$$

Variational inference here is to minimize the KL divergence between the true posterior distribution $P(\mathbf{G}, \mathbf{H}, \mathbf{Z}, \boldsymbol{\alpha}|D, \mathbf{X}, \boldsymbol{\pi}, \mathbf{T}^*, \mathbf{A}^*, \mathbf{V}^*)$ and $q(\boldsymbol{\alpha}, \mathbf{H}, \mathbf{Z}, \mathbf{G})$, so that the posterior of $\{\boldsymbol{\alpha}, \mathbf{G}, \mathbf{Z}, \mathbf{H}\}$ can be nicely approximated. This is equivalent to maximizing the evidence lower bound (ELBO) $L(q, \mathbf{T})$:

$$L(q, \mathbf{T}) \quad = E_q log\left(\frac{P(D, \mathbf{X}, \mathbf{A}, \mathbf{V}, \mathbf{G}, \mathbf{Z}, \mathbf{H}, \boldsymbol{\alpha}|\mathbf{T}, \boldsymbol{\pi}, \mathbf{n^0}, \boldsymbol{\rho}, \lambda, \boldsymbol{\mu})}{\mathrm{q}(\boldsymbol{\alpha}, \mathbf{H}, \mathbf{Z}, \mathbf{G})}\right).$$

Next, in order to maximize the ELBO $L(q, \mathbf{T})$, we calculate its partial derivative with respect to $n_q, \tau_{iq}, \gamma_{qk}$ and $\beta_{ij,k}$, respectively, and set these derivatives to 0. That is,

$$\nabla L(q, \mathbf{T}) = (\frac{\partial L}{\partial n}, \frac{\partial L}{\partial \tau}, \frac{\partial L}{\partial \gamma}, \frac{\partial L}{\partial \beta}) = 0.$$

After this inference, we can collect the optimal terms:

$$\mathrm{n}_q = n_q^0 + \sum_{i=1}^{N} \tau_{iq}, \tag{6}$$

$$\tau_{iq} \propto exp\Big\{\psi(n_q) - \psi(\sum_{q=1}^{Q} n_q) + \sum_{j=1}^{L_i}\sum_{k=1}^{K} \beta_{ij,k}\Big(\psi(\gamma_{qk}) - \psi(\sum_{k=1}^{K}\gamma_{qk})\Big) + \sum_{i'\neq i}^{N}\sum_{l=1}^{Q}\tau_{i'l}\Big(x_{ii'}log(deg_i deg_{i'}\pi_{ql}) + (1-x_{ii'})log(1-deg_i deg_{i'}\pi_{ql})\Big)\Big\}, \tag{7}$$

$$\beta_{ij,k} \propto exp\Big\{\sum_{q=1}^{Q}\tau_{iq}\Big(\psi(\gamma_{qk}) - \psi(\sum_{k=1}^{K}\gamma_{qk})\Big) + \mathbf{v}_{w_{ij}}^T\mathbf{t}_k + r_{ik}\Big\}, \tag{8}$$

$$\gamma_{qk} = \rho_k + \sum_{i=1}^{N}\sum_{j=1}^{L_i}\tau_{iq}\beta_{ij,k}, \tag{9}$$

where $\psi(\cdot)$ denotes the Digamma function.

3.2 Parameter Maximization

In this step, $\mathbf{n}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \boldsymbol{\beta}$ are fixed, and our purpose is to optimize $\mathbf{T}$. First, we take the derivatives of $L(q, \mathbf{T})$ with respect to $\mathbf{T}$ as:

$$\frac{\partial L}{\partial \mathbf{T}} = \sum_{i=1}^{N}\sum_{j=1}^{L_i}(\mathbf{v}_{w_{ij}}\boldsymbol{\beta}_{ij}^T + \sum_{k=1}^{K}\beta_{ij,k}\frac{\partial r_k}{\partial \mathbf{T}}). \tag{10}$$

In the following, we set $\sum_{j=1}^{L_i}\beta_{ij,k} = \overline{m}_{ik}$, which is the sum of variational probabilities of every word in the document d_i being assigned into the k-th topic. Then we use the Gradient Descent method [19]. The k'-th column of the gradient matrix $\partial r_k/\partial \mathbf{T}$ is $\partial r_k/\partial \mathbf{t}_{k'}$. According to Eq. (3), when $k' = k$, we can derive the following equation:

$$\frac{\partial r_k}{\partial \mathbf{t}_k} = -exp\{r_k\}\cdot exp\{\mathbf{t}_k^T\mathbf{V}\}(\mathbf{u}\circ\mathbf{V}),$$

where $\mathbf{u}\circ\mathbf{V}$ denotes the element of each column of $\mathbf{V}$ multiplied by the corresponding element of $\mathbf{u}$. When $k' \neq k$, $\partial r_k/\partial \mathbf{t}_{k'} = \mathbf{0}$ can also be easily derived.

Based on the description above we have $\partial r_k/\partial \mathbf{T} = (\mathbf{0}, \ldots \partial r_k/\partial \mathbf{t}_k, \ldots, \mathbf{0})$. By substituting it into Eq. (10), we can obtain:

$$\frac{\partial L}{\partial \mathbf{T}} = \sum_{i=1}^{N}\Big(\sum_{j=1}^{L_i}\mathbf{v}_{w_{ij}}\boldsymbol{\beta}_{ij}^T + (\overline{m}_{i1}\frac{\partial r_1}{\partial \mathbf{t}_1}, \ldots, \overline{m}_{iK}\frac{\partial r_K}{\partial \mathbf{t}_K})\Big).$$

At this time, $\mathbf{T}$ can be optimized via a gradient descent process:

$$\mathbf{T}^m = \mathbf{T}^{m-1} + \Big(\sum_{i=1}^{N}\lambda(m, L_i)\Big)(\frac{\partial L}{\partial \mathbf{T}}), \tag{11}$$

where m is the ongoing iteration step during the update process, $\lambda(m, L_i) = \frac{L_0\lambda_0}{m\cdot\max\{L_i, L_0\}}$ is a function of the learning rate, λ_0 is the initial learning rate, L_i is the length of the i-th document, and L_0 is the threshold of document length. Besides, since the constraint $||\mathbf{t}_{ik}^{(m)}|| \leq \lambda$, we normalize it by $\lambda/||\mathbf{t}_k^{(m)}||$ when $||\mathbf{t}_k^{(m)}|| > \lambda$.

Also of note, after updating $\mathbf{T}_i$, we can further update the topic residuals $\mathbf{r}$ according to Eq. (4) to refine this process.

3.3 Algorithm and Complexity Analysis

The process of CeTe is shown in Algorithm 1, where we omit the word embedding process for clarity. In the step of variational inference, the time complexity for updating n_q, τ_{iq}, γ_{qk} and $\beta_{ij,k}$ is $O(N)$, $O(K + NQ)$, $O(N\bar{L})$ and $O(QK)$, respectively, where N,

K and Q are the number of documents, topics and communities respectively in the document network and $\bar{L}$ is the average number of words in all documents. Taking advantage of the sparsity of document networks, the above time complexities can be further reduced to $O(NQ)$, $O(K+NQ)$, $O(N\bar{L})$, $O(NQK+MQ^2)$, $O(QKN\bar{L})$ and $O(N\bar{L}QK^2)$, respectively, where M is the number of links in the network. In addition, the time complexity for the step of parameter maximization is $O(N(K+\bar{L}))$. Thus, the overall time complexity of this algorithm is $O(N\bar{L}QK^2+MQ^2)$, which is near linear on the scale of networks (N or M).

Algorithm 1 The Process of CeTe

Input: $\mathbf{X}$, D, Q, K, a threshold ε, $count_{max}$
Output: $\mathbf{T}, \mathbf{H}, \boldsymbol{\alpha}, \mathbf{G}, \mathbf{Z}$
1. Initialize $\boldsymbol{\pi}$, $\mathbf{T}$, $\mathbf{r}$ and variational parameters randomly
2. count=1
3. **repeat:**
   (a) Update $n_q, \tau_{iq}, \beta_{ij,k}, \gamma_{qk}$ via (6)-(9)
   (b) Update $\mathbf{T}$ via (11) and $\mathbf{r}$ via (4)
   (c) Calculate ELBO L^{count} and count=count+1
   Until $L^{count} - L^{count-1} < \varepsilon$ or $count > count_{max}$

4 EXPERIMENTS

We first introduce the experiment setup. We then use document classification, the gold method to validate topic embedding models, to evaluate the quality of document representations (the topic distribution of documents) derived by our new approach CeTe. Next, we demonstrate the visualization of topic embeddings to further validate its superiority. We finally use a case study to show why CeTe works.

4.1 Experiment Setup

4.1.1 Datasets. We conducted the experiments on three public datasets, including one DBLP dataset [25] and two hep-th datasets[1]. DBLP includes a collection of papers in computer field. In DBLP, the title and abstract are extracted as texts for each paper and the citation relationships are used to form links between papers. We extract its largest connected component as done by other related works [6, 17, 29]. The ground truth of the papers is fixed according to the CCF (China Computer Federation) classification. The hep-th includes a large corpus of physics-related papers. Its data processing method is similar to that used by DBLP, except that here we use Journals to form the ground truth. In addition, due to the high imbalance of document quantities in different categories for both DBLP and hep-th, we choose the five largest categories to form the DBLP dataset; and for hep-th, we use the four largest categories to form the first subdataset called large-hep, and the three smaller categories to form the other subdataset called small-hep. The datasets details are shown in Table 1.

4.1.2 Baselines. We test our approach against seven state-of-the-art methods, which can be divided into five categories. The first includes three topic models, i.e. Latent Dirichlet Allocation (LDA) [3], Laten Feature Topic Modelling (LFTM) [23] and GibbsLDA [7].

[1]https://www.cs.cornell.edu/projects/kddcup/datasets.html

Table 1: Datasets statistics.

Datasets	Documents	Edges	Words	Categories
DBLP	6,936	12,353	506,269	5
small-hep	397	812	18,718	3
large-hep	11,752	134,956	622,642	4

The second is a topic model that also incorporates network links, named Relational Topic Model (RTM) [4]. The third is a topic embedding method called TopicVec [19]. The fourth is a community detection method named Degree-Corrected Stochastic Block Model (DCSBM) [17], in which we use the community memberships of documents as their representations. The fifth is a document representation method called Doc2Vec [18].

4.1.3 Parameter Setting. As done by TopicVec [19], word embeddings are pre-trained by PSDVec [20] on the latest Wikipedia snapshot[2]. The dimensions of topic embeddings and word embeddings are both set to 500. For datasets DBLP, large-hep and small-hep, we set the number of topics to 500, 200 and 30, respectively. We set the hyper-parameters $n^0 = \rho = 1$ and $\lambda = 7$, which are similar with those used in baselines.

4.2 Test on Document Classification

On this test, according to [19], we first randomly select 80% of documents as the training set and 20% as the test set for each dataset. We then use ℓ-1 regularized SVM as classifier. We use the scikit-learn library[3] to train an ℓ-1 regularized linear SVM one-vs-all classifier for fairness. We adopt macro-averaged Precision, Recall and F1-Score as the evaluation indices. The document representations derived from each method compared are used as input instances to the classifier.

Based on the results shown in Table 2, our new approach CeTe performs the best on all datasets. Neither the topic models (LDA, GibbsLDA and LFTM) using documents alone, nor the community detection method (DCSBM) only using network topology performs well. While RTM (a topic model incorporating links) performs better in most cases, it is still not competitive with our approach CeTe. The reason should be that we introduce the words co-occurrence information at different levels and correlations between topics by using topic embedding to better model topic information. In addition, compared to the state-of-the-art topic embedding method TopicVec, our CeTe still has an obvious improvement. This should be because, by introducing the link information (especially the high-order community structure), the problem of semantic fuzziness of topics suffered by topic embedding methods has been alleviated by our approach. These all validate the effectiveness of this new approach.

4.3 Test by Visualization of Topic Embeddings

We illustrate the visualization of topic embeddings derived by our CeTe method to further validate its performance. The results on DBLP with 500 topics are shown in Figure 2. We find that some closely connected articles correspond to each topic. For example, in the upper-left set, the words 'wireless', 'communication', 'network' and 'sensor' are all related to a field of computer network.

[2]https://dumps.wikimedia.org/enwiki/
[3]http://scikit-learn.org/stable/modules/svm.html

Table 2: Performance of different methods on document classification. Bold indicates the best score.

	DBLP			small-hep			large-hep		
	Precison	Recall	F1-Score	Precison	Recall	F1-Score	Precison	Recall	F1-Score
LDA	0.700	0.457	0.404	0.287	0.287	0.281	0.414	0.411	0.401
GibbsLDA	0.786	0.722	0.715	0.399	0.400	0.396	0.346	0.405	0.346
Doc2Vec	0.523	0.496	0.448	0.454	0.438	0.431	0.385	0.394	0.350
LFTM	0.739	0.653	0.634	0.619	0.575	0.569	0.419	0.418	0.405
RTM	0.560	0.563	0.549	0.565	0.550	0.553	0.407	0.411	0.393
DCSBM	0.461	0.430	0.393	0.451	0.487	0.439	0.359	0.358	0.325
TopicVec	0.729	0.689	0.681	0.624	0.600	0.599	0.405	0.417	0.393
CeTe	**0.790**	**0.736**	**0.731**	**0.671**	**0.650**	**0.654**	**0.437**	**0.433**	**0.417**

It partly shows that the topic embeddings derived by our method have a high quality. We also noticed that, there is a word 'software' in the lower-middle set, which should originally belong to the 'software engineering' field, but actually belongs to the area of 'high-performance computing' here. That is, the keywords such as 'distributed', 'power' and 'energy' in this topic are all from 'high-performance computing', while there is another keyword 'software' which seems to be a mistake. But after a deep investigation we find that, the 'high-performance computing' topic also contains some papers using the idea of software engineering. At the same time, these papers are linked with other papers in 'high-performance computing' and are classified in the same community and thus corrected by our method. This just corresponds to the real situation where these papers (with semantic fuzzy words) belong to the field of 'high-performance computing'. This further validates that the problem of semantic fuzziness of topic embeddings brought by the existence of semantic fuzzy words has been addressed by using links with communities. This visualization can further help discover the meaningful semantic structures to understand topics in real applications.

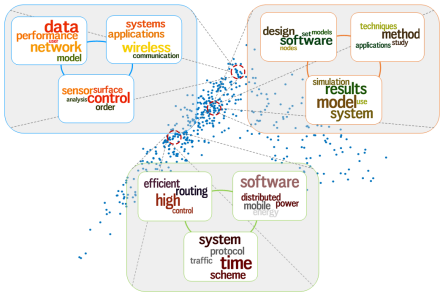

Figure 2: An example of visualizing the 500 topic embeddings derived by CeTe on the DBLP dataset. Each point in the cloud represents a latent topic. The distance between points indicates the correlation between topics, and smaller distance means stronger correlation. We mainly show three sets of points (topics) with small distance in the embedding space. Each topic is characterized by keywords according to the word distribution of this topic (by calculating the cosine similarity between topic embedding and word embeddings). The word sizes are proportional to the probability that this word belongs to the topic. Edges represent the strength of the correlations between topics.

4.4 A case study

Here is to illustrate how our new model CeTe alleviates the topical fuzziness problem of topic embedding methods by introducing the high-order community structure. We also test it on the DBLP dataset with three categories (i.e. theoretical computer science, high-performance computing and computer network), and compare it with TopicVec which is one of the state-of-the-art topic embedding models [19]. We randomly select 10% of the documents as test set, and use the topic distributions of documents derived from CeTe (and TopicVec) as document representations. There are 78 of 375 documents that are classified into wrong categories by TopicVec, while our method corrects about half of them. The results are shown in Figure 3. Due to space limit, we only discuss three of the papers in details since the rests are similar.

The first sample paper is 'Microcode compaction with timing constraints' which is clearly in the field of 'high-performance computing'. However, some words such as 'timing constraints', 'optimization' and 'heuristic' often appear in its maintext. The embedding of these fuzzy topical words is the reason why this paper is assigned to the 'theoretical computer science' category by TopicVec. However, this paper and other seven papers in 'high-performance computing' are divided into the same community by our CeTe, and thus have similar topic distributions. In other words, the topical fuzziness of this paper has been corrected.

The second is the paper 'MILP approach to pattern generation in logical analysis of data' which belongs to the 'theoretical computer science' category. But there are also some high-frequency fuzzy topical words such as 'linear', 'efficient', 'optimal' and 'hundreds of' in this paper which lead to that it is misplaced by TopicVec. However, through links, this paper and three other papers in 'theoretical computer science' are classified into the same community, and hence its topic is corrected again.

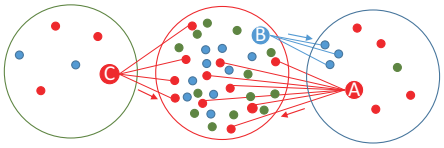

Figure 3: An example that uses network communities to correct inaccurate document representations suffered from the semantic fuzziness of topics. The blue, red and green circles are used to represent the three categories, i.e. theoretical computer science, high-performance computing and computer network, respectively. The blue (red or green) nodes denotes that it belongs to the blue (red or green) category. The nodes with connections in this figure denotes that they are divided into the same community by the topological component of our model. Here only some typical examples are shown in each category. Most of them are papers indeed influenced by semantic fuzziness of word embeddings, which are wrongly classified by TopicVec while corrected by our CeTe. Especially, the titles of papers A, B and C are 'Microcode compaction with timing constraints', 'MILP approach to pattern generation in logical analysis of data' and 'Fast Path-Based Neural Branch Prediction', respectively.

The third paper is 'Fast Path-Based Neural Branch Prediction' which belongs to 'high-performance computing'. But as shown in Figure 3, it is incorrectly assigned to 'computer network' by TopicVec. After a deeper investigation of its maintext, we found that

there are often some keywords such as 'latency', 'path', 'branch' and 'instructions' frequently appear in this paper which may mislead its topic distribution. But by utilizing links between papers, this paper and other three papers in 'high-performance computing' are assigned to the same community, which helps to correct its topic distributions by our CeTe.

To sum up, this case study further validates that, with the help of network community, the inaccurate topic distributions of documents brought by embedding semantic fuzzy words can be well alleviated by our CeTe approach.

5 CONCLUSION AND DISCUSSIONS

We proposed a novel generative topic embedding model CeTe using documents with topics and network structure with communities together, and developed an efficient variational expectation-maximization algorithm to learn the model. The new approach utilizes the high-order network information, i.e., community structure, to solve the problem of topical fuzziness suffered by topic embedding models which is from introducting the embeddings of semantic fuzzy words. The experiment results on two tasks (i.e. document classification and topic visualization) validate its superiority compared with the state-of-the-art methods.

There have been some topic embedding methods presented recently [14, 19, 21]. However, it is the first time, to the best of our knowledge, to use network topology, specifically the community structure, to improve their performance. In addition, there are also some topic models [4, 8, 15, 31] incorporating network information. However, they have essential differences. First, the introduction of links in topic models is to capture the additional correlations between documents, while our purpose is to use links to alleviate the topical fuzziness problem suffered by topic embedding methods (due to the introduction of word embeddings of the semantic fuzzy words). Second, rather than using the links directly, we use the high-order network information, i.e., community structure, to overcome the shortcomings brought by the noise and sparseness of networks. Third, we further use probability transition to describe the relationship between topics and communities, in order to make the method robust even when the topics and communities do not match very well.

ACKNOWLEDGEMENTS

This work was partly supported by the National Key R&D Program of China (No.2017YFC0820106) and the National Natural Science Foundation of China (No.61772361, 61876128, 61503281, 61502334).

REFERENCES

[1] David Blei and John Lafferty. 2006. Correlated topic models. *Advances in neural information processing systems* 18 (2006), 147.
[2] David M Blei and John D Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics* 1, 1 (2007), 17–35.
[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[4] Jonathan Chang and David Blei. 2009. Relational topic models for document networks. In *Artificial Intelligence and Statistics*. 81–88.
[5] Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang. 2013. Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems*. 2445–2453.
[6] Pin-Yu Chen and Alfred O Hero. 2015. Deep community detection. *IEEE Transactions on Signal Processing* 63, 21 (2015), 5706–5719.
[7] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101, suppl 1 (2004), 5228–5235.
[8] Dongxiao He, Zhiyong Feng, Di Jin, Xiaobao Wang, and Weixiong Zhang. 2017. Joint identification of network communities and semantics via integrative modeling of network topologies and node contents. In *Thirty-First AAAI Conference on Artificial Intelligence*. AAAI Press, 116–124.
[9] Junxian He, Zhiting Hu, Taylor Berg-Kirkpatrick, Ying Huang, and Eric P Xing. 2017. Efficient correlated topic modeling with topic embedding. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 225–233.
[10] Yuan He, Cheng Wang, and Changjun Jiang. 2018. Discovering canonical correlations between topical and topological information in document networks. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 30, 3 (2018), 460–473.
[11] Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 289–296.
[12] Xin Huang and Laks VS Lakshmanan. 2017. Attribute-driven community search. *Proceedings of the VLDB Endowment* 10, 9 (2017), 949–960.
[13] Seungil Huh and Stephen E Fienberg. 2012. Discriminative topic modeling based on manifold learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 5, 4 (2012), 20.
[14] Di Jiang, Lei Shi, Rongzhong Lian, and Hua Wu. 2016. Latent topic embedding. In *Proceedings of the 26th International Conference on Computational Linguistics*. 2689–2698.
[15] Di Jin, Xiaobao Wang, Ruifang He, Dongxiao He, Jianwu Dang, and Weixiong Zhang. 2018. Robust detection of link communities in large social networks by exploiting link semantics. In *Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press, 314–321.
[16] Indika Kahanda and Jennifer Neville. 2009. Using Transactional Information to Predict Link Strength in Online Social Networks. *The International AAAI Conference on Web and Social Media* 9 (2009), 74–81.
[17] Brian Karrer and Mark EJ Newman. 2011. Stochastic blockmodels and community structure in networks. *Physical Review E* 83, 1 (2011), 016107.
[18] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*. 1188–1196.
[19] Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. 2016. Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 666–675.
[20] Shaohua Li, Jun Zhu, and Chunyan Miao. 2015. A Generative Word Embedding Model and its Low Rank Positive Semidefinite Solution. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1599–1609.
[21] Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical Word Embeddings.. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2418–2424.
[22] Sharad Nandanwar and M Narasimha Murty. 2016. Structural neighborhood based classification of nodes in a network. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1085–1094.
[23] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics* 3 (2015), 299–313.
[24] Guo-Jun Qi, Charu Aggarwal, Qi Tian, Heng Ji, and Thomas Huang. 2012. Exploring context and content links in social media: A latent space method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 5 (2012), 850–862.
[25] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 990–998.
[26] Mirwaes Wahabzada, Zhao Xu, and Kristian Kersting. 2010. Topic models conditioned on relations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 402–417.
[27] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Manifold-Ranking Based Topic-Focused Multi-Document Summarization.. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, Vol. 7. 2903–2908.
[28] Suhang Wang, Jiliang Tang, Charu Aggarwal, and Huan Liu. 2016. Linked document embedding for classification. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. ACM, 115–124.
[29] Joyce Jiyoung Whang, David F Gleich, and Inderjit S Dhillon. 2016. Overlapping community detection using neighborhood-inflated seed expansion. *IEEE Transactions on Knowledge and Data Engineering* 28, 5 (2016), 1272–1284.
[30] Wenhui Wu, Sam Kwong, Yu Zhou, Yuheng Jia, and Wei Gao. 2018. Nonnegative matrix factorization with mixed hypergraph regularization for community detection. *Information Sciences* 435 (2018), 263–281.
[31] Ge Zhang, Di Jin, Jian Gao, Pengfei Jiao, Francoise Fogelman-Soulié, and Xin Huang. 2018. Finding communities with hierarchical semantics by distinguishing general and specialized topics. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 3648–3654.