3-in-1 Correlated Embedding via Adaptive Exploration of the Structure and Semantic Subspaces

Liang Yang[1,2,†], **Yuanfang Guo**[2,†], **Di Jin**[3,*], **Huazhu Fu**[4], **Xiaochun Cao**[2]
[1] School of Computer Science and Engineering, Hebei University of Technology
[2] State Key Laboratory of Information Security, Institute of Information Engineering, CAS
[3] School of Computer Science and Technology, Tianjin University
[4] Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore
{yangliang, guoyuanfang, caoxiaochun}@iie.ac.cn, jindi@tju.edu.cn, huazhufu@gmail.com

Abstract

Combinational network embedding, which learns the node representation by exploring both topological and non-topological information, becomes popular due to the fact that the two types of information are complementing each other. Most of the existing methods either consider the topological and non-topological information being aligned or possess predetermined preferences during the embedding process. Unfortunately, previous methods fail to either explicitly describe the correlations between topological and non-topological information or adaptively weight their impacts. To address the existing issues, three new assumptions are proposed to better describe the embedding space and its properties. With the proposed assumptions, nodes, communities and topics are mapped into one embedding space. A novel generative model is proposed to formulate the generation process of the network and content from the embeddings, with respect to the Bayesian framework. The proposed model automatically leans to the information which is more discriminative. The embedding result can be obtained by maximizing the posterior distribution by adopting the variational inference and reparameterization trick. Experimental results indicate that the proposed method gives superior performances compared to the state-of-the-art methods when a variety of real-world networks is analyzed.

1 Introduction

Network embedding, which learns the representation of every node in the network, challenges the end-to-end strategy for its crown in network analysis [Hamilton *et al.*, 2017; Cai *et al.*, 2017]. Tremendous efforts have been made to improve the performance of embedding in two directions. One is the preservation of structure and properties of the network [Wang *et al.*, 2017b; Ou *et al.*, 2016; Yang *et al.*, 2018], while the other is the integration of side information [Yang *et al.*, 2015; Tu *et al.*, 2017; 2016]. Recently, researchers started to jointly explore the two directions to further improve the embedding performance. To exploit the advantages of both the topological and non-topological information, which are actually complementing each other, combinational network embedding learns the node representation by exploring both of them and draws great attentions [Wang *et al.*, 2017a; Huang *et al.*, 2017b].

Most of the existing methods can be categorized into two classes. Class 1 methods, which assume the topological and non-topological information being aligned, usually force the entire embedding to represent the two kinds of information simultaneously [Chang and Blei, 2009; Wang *et al.*, 2016; Pan *et al.*, 2016; Liao *et al.*, 2017]. However, the alignment assumption doesn't usually hold in reality. Taking Twitter as an example, social relationships often indicate user groups

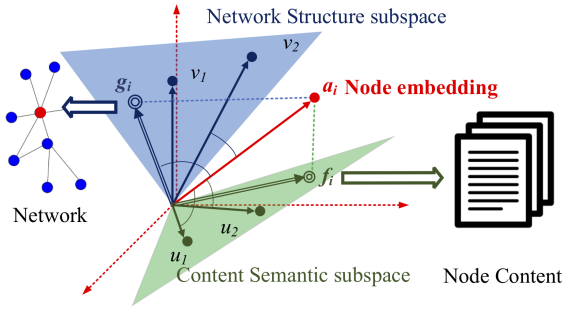


Figure 1: An example of the proposed embedding space. The embedding space (red coordinate system) is spanned by the content semantic subspace (green plane) and the network structure subspace (blue plane). The network structure subspace (blue plane) is spanned by the embeddings of communities v_1 and v_2, while the content semantic subspace (green plane) is spanned by the embeddings of topics u_1 and u_2. The projection of node embedding a_i to the network structure subspace forms the node's community weight g_i which represents the network topology, while the projection to the semantic subspace forms node's topic weight f_i which represents the content.

†These authors contributed equally to this work and should be considered co-first authors.
∗Corresponding author.

and users may post messages of diverse topics, i.e., the network structure and content semantic are not aligned at all.

Class 2 methods, which usually assume one kind of information to be more important than the other, tend to require the embeddings to perfectly represent one kind of information and then constrain the embeddings with the other kind of information [Yang *et al.*, 2015; Wang *et al.*, 2017a]. [Huang *et al.*, 2017b] regularizes the network embedding by employing the content embeddings. [Huang *et al.*, 2017a] constrains the embedding of the node attribute with the network topology information. From the literature review we can conclude that some methods in Class 2 are scheduled to prefer the topological information, while others are set to prefer the non-topological information. However, these predetermined preferences of the topological or non-topological information are not rigorous.

In fact, topological and non-topological information is correlated and they should be considered with no predetermined preferences during the embedding process. The ignorance of the correlation tends to degrade the integration of these information and the quality of embedding. The biased treatments of the two kind of information will reduce the versatility of the embedding methods.

In this paper, to alleviate these issues, we improve the embedding performance by considering the correlation in the topological and non-topological information. Three new assumptions about the embedding space and its properties have been proposed, as shown in Figure 1. With the assumptions, nodes, communities (mesoscopic properties of the topological information) and topics (mesoscopic properties of the non-topological information) are seamlessly transformed into one embedding space and a new generative model is developed to portray the generation process of the network and content. This model, which is constructed with respect to the Bayesian framework, can adaptively weight the impacts of the two types of information. i.e., it can automatically lean to the information which is more discriminative. Specifically, if the community structure of the network is more distinguishable than the topic semantics of node content, the model will assign more weight to the topological information, and vice versa. Then the embeddings can be achieved by maximizing the posterior distribution with the variational inference and reparameterization trick.

The main contributions are summarized as follows:

1. We propose three new assumptions to explicitly describe the correlations between the topological and non-topological information.
2. We propose a novel generative model to portray the generation process of the network and content from the latent embeddings of nodes, communities and topics, and adaptively weight their impacts.
3. We propose an efficient variation inference algorithm by adopting the reparameterization trick to maximize the posterior distribution and obtain the embeddings.

2 Correlated Embedding Model

A network with node content can be represented as an attributed graph $G = (O, E, W)$. $O = \{o_i | i = 1, ..., D\}$ is a set of D vertices, where o_i is associated with a bag of N_i words $\{w_{i,n} | n = 1, ..., N_i\}$. Each word $w_{i,n}$ is drawn from the vocabulary of L words. E is a set of edges, each of which connects two vertices in O. The adjacency matrix $Y = [y_{ij}] \in \mathbb{R}^{D \times D}$ is adopted to represent the network topology, where $y_{ij} = 1$ if an edge exists between the vertices o_i and o_j, and vice versa. Besides, we assume that the number of communities in the network, K, and the number of topics in the content, T, are known.

In this section, the nodes, communities and topics are designed to be represented in the same M-dimensional latent embedding space. We denote $A \in \mathbb{R}^{D \times M}$, $V \in \mathbb{R}^{K \times M}$ and $U \in \mathbb{R}^{T \times M}$ as the embedding matrices of the nodes, communities and topics, where $a_i \in \mathbb{R}^M$, $v_k \in \mathbb{R}^M$ and $u_t \in \mathbb{R}^M$ are the corresponding embedding of node i, community k and topic t, respectively.

2.1 Assumptions

To explore the correlations between the topological and non-topological information, we make three natural assumptions about the embedding space and its inner relationships.

Assumption 1: The 3-in-1 embedding space $\mathbb{R}^M$ is spanned by the network structure subspace $\mathcal{S} \subset \mathbb{R}^M$ and the content semantic subspace $\mathcal{C} \subset \mathbb{R}^M$. An example is shown in Figure 1, where the embedding space (red coordinate system) is spanned by the content semantic subspace (blue plane) and the network structure subspace (green plane).

Assumption 2: Subspace reflects the mesoscopic properties of the observations. As shown in previous work, mesoscopic properties, such as community structures in network embedding [Wang *et al.*, 2017b] and topics in document embedding [He *et al.*, 2017], should be preserved during embedding. Instead of forcing the entire embedding to represent the network structure and content semantic information simultaneously, we indirectly constrain the corresponding subspaces to be spanned by the embeddings of mesoscopic properties. In this paper, we define the network structure and content semantic subspaces as $\mathcal{S} = span(v_1, v_2, ..., v_K)$ and $\mathcal{C} = span(u_1, u_2, ..., u_T)$ respectively, where $v_k \in \mathbb{R}^M$ and $u_t \in \mathbb{R}^M$ are the embeddings of the communities and topics. Let $V = [v_1, v_2, ..., v_K] \in \mathbb{R}^{K \times M}$ and $U = [u_1, u_2, ..., u_T] \in \mathbb{R}^{T \times M}$. As the example shown in Figure 1, the network structure subspace (blue plane) is spanned by the embeddings of communities v_1 and v_2, while the content semantic subspace (green plane) is spanned by the embeddings of topics u_1 and u_2.

Assumption 3: The corresponding information is generated by the node's mesoscopic properties, which are the projections of the embedding to the subspaces. The projection of embedding a_i to the network structure subspace is $V'Va_i \in \mathbb{R}^M$, and the coordinate based on bases $\{v_1, v_2, ..., v_K\}$ is $g_i = Va_i \in \mathbb{R}^K$. g_i, which is the inner product of the embedding of node and communities, can be regarded as the community weight of node i. It can be transformed into the community distribution via $s_i = softmax(g_i)$ where $s_{ik} = softmax_k(g_i) = e^{g_{ik}} / \sum_j e^{g_{ij}}$. Similarly, the topic weight of node i is $f_i = Ua_i \in \mathbb{R}^T$ and the topic distribution is $r_i = softmax(f_i)$, where $r_{it} = softmax_t(f_i) =$

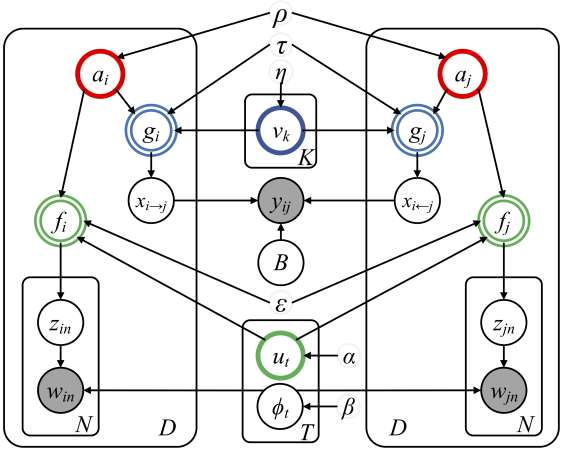


Figure 2: Graphical representation of the proposed model. The shaded circles denote the observed variables. The single rings with red, blue and green color denote the embeddings of nodes, communities and topics. The double rings with blue and green color denote the community and topic weights. The relationships among a_d, u_t, v_k, f_d and g_d are also shown in Figure 1.

Symbol	Description
D, R	number of nodes and edges
K, T	number of latent communities and topics
N_d	number of words in document d
L	vocabulary size
M	embedding dimension
y_{ij}	connection between nodes i and j
w_{dn}	the nth word in node d
a_d, u_t, v_k	embeddings of node d, topic t and community k
f_d	topic weight of node d
g_d	community weight of node d
z_{dn}	topic assignment of word w_{dn}
$x_{i\to j}$	community assignment of node pair (i, j) initiator
$x_{i\leftarrow j}$	community assignment of node pair (i, j) receiver
ϕ_t	word distribution of topic t
ρ, η, α	uncertainty degrees of three kinds of embeddings
τ, ε	uncertainty degrees of community and topic
β	prior of word distribution
B	probabilities of interactions between communities

Table 1: Notations.

$e^{f_{it}}/\sum_j e^{f_{ij}}$. g_i and f_i can be regarded as the structure and content embeddings of node i respectively. According to the mixed-membership model [Blei *et al.*, 2003; Airoldi *et al.*, 2008], the topic and community distributions of nodes are the latent variables which respectively generate the network topology and node content. An example is shown in Figure 1 and the detailed generation process will be discussed in Section 2.2.

2.2 Generative Model

The assumptions proposed above connect the embeddings (nodes, communities and topics) and mesoscopic properties (community and topic) . Here we continue to introduce the generative process from the mesoscopic properties of each node to the topological and non-topological observations. For topological observation, the stochastic block model is adopted. The community distribution of node generates a community assignment to the initiator and receiver for each pair of nodes. The probability of this pair of nodes being connected depends on the communities to which initiator and receiver belong. For non-topological observation, the topic model is adopted. The topic distribution of node generates a topic assignment to each word position. The probability of the word appearing in current position depends on the word distribution of this topic.

The generative process is described as follows. The proposed model, which is illustrated in Figure 2, is represented by three types of variables (observed variables, latent variables and parameter variables). Note that the notations are summarized in Table 1.

1. For each topic $t = 1, 2, ..., T$
   (a) Draw the topic word distribution $\phi_t \sim Dir(\beta)$
   (b) Draw the topic embedding $u_t \sim \mathcal{N}(0, \alpha^{-1}\mathbf{I})$
2. For each community $k = 1, 2, ..., K$
   (a) Draw the community embedding $v_k \sim \mathcal{N}(0, \eta^{-1}\mathbf{I})$
3. For each node $d = 1, 2, ..., D$
   (a) Draw node embedding $a_d \sim \mathcal{N}(0, \rho^{-1}\mathbf{I})$
   (b) Draw the node topic weight $f_d \sim \mathcal{N}(Ua_i, \varepsilon^{-1}\mathbf{I})$
   (c) Draw the node community weight $g_d \sim \mathcal{N}(Va_i, \tau^{-1}\mathbf{I})$
   (d) Derive the distribution over topics $r_d = softmax(f_d)$
   (e) Derive the distribution over communities $s_d = softmax(g_d)$
   (f) For each word $n = 1, 2, ..., N_d$
      i. Draw the topic assignment $z_{dn} \sim Mult(r_d)$
      ii. Draw the word $w_{dn} \sim Mult(\phi_{z_{dn}})$
4. For each pair of nodes i and j
   (a) Draw initiator indicator $x_{i\to j} \sim Mult(s_i)$
   (b) Draw receiver indicator $x_{i\leftarrow j} \sim Mult(s_j)$
   (c) Draw link $y_{ij} \sim Bernoulli(B_{x_{i\to j}, x_{i\leftarrow j}})$.

The joint distribution generated by the above model is

$$\begin{aligned} &P(Y, W, A, F, G, U, V, X, Z, B, \Phi|\alpha, \eta, \rho, \xi, \tau, \varepsilon, \beta) \\ &= \prod_i \prod_{i<j} P(x_{i\to j}|s_i)P(x_{i\leftarrow j}|s_j)P(y_{ij}|x_{i\to j}, x_{i\leftarrow j}, B) \\ &\prod_d \prod_n P(z_{dn}|r_d)P(w_{dn}|z_{dn}, \Phi) \prod_t P(u_t|\alpha)P(\phi_t|\beta) \\ &\prod_k P(v_k|\eta) \prod_d P(a_d|\rho)P(g_d|a_d, V, \tau)P(f_d|a_d, U, \varepsilon), \end{aligned}$$

where $\sum_l \phi_{tl} = 1$. To obtain the embeddings of nodes, communities and topics, A^*, F^*, G^*, U^*, V^*, X^*, Z^*, B^* and Φ^* that maximizing the posterior $P(A, F, G, U, V, X, Z, B, \Phi|Y, W, \alpha, \eta, \rho, \xi, \tau, \varepsilon, \beta)$ should be computed. However, since the likelihood is intractable, the posterior can be approximated via variational inference.

Remarks: One of the most remarkable advantages of the proposed method is that it automatically weights the impacts of two kinds of information and leans to the more discriminative one. For example, assume there exists an attributed network. The community structure in that network is more distinguishable than the topic semantics in that network, i.e., the community assignment of each node is more unequivocal (the entropy of each s_i is low); while the topic assignment of each

node is less unequivocal (the entropy of each r_i is very high). Since both $s_i = softmax(Va_i)$ and $r_i = softmax(Ua_i)$ are derived from the projections of a_i to the network structure subspaces $\mathcal{S}$ and the content semantic subspace $\mathcal{C}$ respectively, the correlation between a_i and $\mathcal{S}$ is high, while the correlation between a_i and $\mathcal{C}$ is low, i.e., the model leans to the topological information.

3 Variational Inference

In this section, we solve the proposed formulation by adopting the variational EM algorithm. By grouping all the variables into the observed variables as $J = \{Y, W\}$ and the latent variables as $H = \{A, F, G, U, V, X, Z, \Phi\}$, the log marginal probability can be decomposed as

$$\begin{aligned} \ln P(J) &= \mathcal{L}(q(H)) + \mathrm{KL}(q(H)||P(H|J)), \\ \mathcal{L}(q(H)) &= \int q(H) \ln \Big\{ \frac{P(H,J)}{q(H)} \Big\} dH, \qquad (1) \\ \mathrm{KL}(q(H)||P(H|J)) &= -\int q(H) \ln \Big\{ \frac{P(H|J)}{q(H)} \Big\} dH. \end{aligned}$$

Maximizing the lower bound $\mathcal{L}(q(H))$ w.r.t. $q(H)$ is equivalent to minimizing the KL divergence between $q(H)$ and $P(H|J)$. Therefore, the posterior $P(H|J)$ can be approximated by $q(H)$. Here, $q(H)$ is restricted to mean-field family of variational distributions

$$\begin{aligned} & q(H) = q(A, F, G, U, V, X, Z, \Phi) \\ &= \prod_k q(v_k) \prod_t q(u_t) q(\phi_t) \prod_d q(a_d) q(f_d) q(g_d) \\ & \prod_d \prod_n q(z_{dn}) \prod_i \prod_{i<j} q(x_{i\to j}) q(x_{i\leftarrow j}), \end{aligned}$$

where the factors have the following parametric forms,

$$\begin{aligned} q(u_t) &= \mathcal{N}(u_t|\mu_t, \Sigma_t^{(u)}), \qquad q(v_k) = \mathcal{N}(v_k|\nu_k, \Sigma_k^{(v)}), \\ q(a_d) &= \mathcal{N}(a_d|\gamma_d, \Sigma_d^{(a)}), \qquad q(f_d) = \mathcal{N}(f_d|\zeta_d, \Sigma_d^{(f)}), \\ q(g_d) &= \mathcal{N}(g_d|\psi_d, \Sigma_d^{(g)}), \qquad q(z_{dn}) = Mult(z_{dn}|\kappa_{dn}), \\ q(\phi_t) &= Dir(\phi_t|\lambda_t), \qquad q(x_{i\to j}) = Mult(x_{i\to j}|\chi_{i\to j}), \\ & \qquad\qquad q(x_{i\leftarrow j}) = Mult(x_{i\leftarrow j}|\chi_{i\leftarrow j}). \end{aligned}$$

For simplicity, the covariance matrices for $q(f_d)$ and $q(g_d)$ are both assumed to be diagonal, i.e.,

$$\Sigma_d^{(f)} = \mathrm{diag}(\sigma_{d1}^{(f)}, ..., \sigma_{dT}^{(f)}), \quad \Sigma_d^{(g)} = \mathrm{diag}(\sigma_{d1}^{(g)}, ..., \sigma_{dK}^{(g)}).$$

Then, the objective function can be obtained as

$$\begin{aligned} & \mathcal{L}(q(H)) = \mathcal{L}(q(A, F, G, U, V, X, Z, \Phi)) \\ &= \sum_k \mathbb{E}_q\Big[\log \frac{P(v_k)}{q(v_k)}\Big] + \sum_t \mathbb{E}_q\Big[\log \frac{P(u_t)P(\phi_t)}{q(u_t)q(\phi_t)}\Big] \\ &+ \sum_d \mathbb{E}_q\Big[\log \frac{P(a_d)P(g_d)P(f_d)}{q(a_d)q(g_d)q(f_d)}\Big] \\ &+ \sum_d \sum_n \mathbb{E}_q\Big[\log \frac{P(z_{dn})P(w_{dn})}{q(z_{dn})}\Big] \\ &+ \sum_i \sum_{i<j} \mathbb{E}_q\Big[\log \frac{P(x_{i\to j})P(x_{i\leftarrow j})P(y_{ij})}{q(x_{i\to j})q(x_{i\leftarrow j})}\Big]. \end{aligned}$$

$\mathcal{L}(q(H))$ is iteratively minimized by varying each $q(.)$ and fixing the others. For each topic t, we update $q(u_t) = \mathcal{N}(u_t|\mu_t, \Sigma_t^{(u)})$ by minimizing

$$\begin{aligned} \mathcal{L}(q(u_t)) &= \mathbb{E}_q[\log P(u_t|\alpha)] + \sum_d \mathbb{E}_q[\log P(f_d|a_d, U, \varepsilon)] \\ &- \mathbb{E}_q[\log q(u_t)]. \end{aligned}$$

By rearranging the items which do not contain u_t, we obtain

$$\ln q^*(u_t) = \mathbb{E}_{-u_t}[\log P(u_t|\alpha)] + \sum_d \mathbb{E}_{-u_t}[\log P(f_d|a_d)],$$

where

$$\begin{aligned} & \mathbb{E}_{-u_t}[\log\{(2\pi)^{-\frac{M}{2}} \alpha^{\frac{M}{2}} \exp(-\tfrac{\alpha}{2} u_t' u_t)\}] \propto -\tfrac{\alpha}{2} u_t' u_t, \\ & \mathbb{E}_{-u_t}[\log\{(2\pi)^{-\frac{M}{2}} \varepsilon^{\frac{M}{2}} \exp(-\tfrac{\varepsilon}{2}(f_d - Ua_d)'(f_d - Ua_d))\}] \\ & \propto -\tfrac{\varepsilon}{2} u_t' \Big[\textstyle\sum_d (\Sigma_d^{(a)} + \gamma_d \gamma_d')\Big] u_t + \varepsilon \textstyle\sum_d \zeta_{dt} \gamma_d' u_t. \end{aligned}$$

Then, we can conclude that $q^*(u_t)$ is proportional to

$$\exp\{-\frac{1}{2} u_t' [\alpha I + \varepsilon \textstyle\sum_d (\Sigma_d^{(a)} + \gamma_d \gamma_d')] u_t' + \textstyle\sum_d \zeta_{dt} \gamma_d' u_t\}$$

Thus, $q^*(u_t)$ satisfies our assumption that $q(u_t) = \mathcal{N}(u_t|\mu_t, \Sigma_t^{(u)})$ where

$$\begin{aligned} \Sigma^{(u)} = \Sigma_t^{(u)} &= \Big[\alpha I + \varepsilon \textstyle\sum_d (\Sigma_d^{(a)} + \gamma_d \gamma_d')\Big]^{-1}, \\ \mu_t &= \varepsilon \Sigma^{(u)} \textstyle\sum_d \zeta_{dt} \gamma_d, \qquad (2) \end{aligned}$$

where $\Sigma_t^{(u)}$ is independent with t and all topics share the same covraiance matrix $\Sigma^{(u)}$. Similarly, $q^*(v_k)$ meets our assumption that $q(v_k) = \mathcal{N}(v_k|\nu_k, \Sigma_k^{(v)})$ where

$$\begin{aligned} \Sigma^{(v)} = \Sigma_k^{(v)} &= \Big[\eta I + \tau \textstyle\sum_d (\Sigma_d^{(a)} + \gamma_d \gamma_d')\Big]^{-1}, \\ \nu_k &= \tau \Sigma^{(v)} \textstyle\sum_d \psi_{dk} \gamma_d. \qquad (3) \end{aligned}$$

Similarly, $q(a_d) = \mathcal{N}(a_d|\gamma_d, \Sigma^{(a)})$, where $\Sigma^{(a)}$ and γ_d are

$$\begin{aligned} & \Big[\rho I + \tau \sum_k (\Sigma^{(v)} + \nu_k \nu_k') + \varepsilon \sum_t (\Sigma^{(u)} + \mu_t \mu_t')\Big]^{-1}, \\ & \Sigma^{(a)} \Big(\tau \textstyle\sum_k \psi_{dk} \nu_k + \varepsilon \textstyle\sum_t \zeta_{dt} \mu_t\Big). \qquad (4) \end{aligned}$$

After approximate distributions for three embeddings, $q(a_d)$, $q(u_t)$ and $q(v_k)$ are obtained. Then we optimize $\mathcal{L}(q(H))$ according to the latent variables ϕ_t, z_{dn} and f_d in the topic model branch. Similar to LDA, we can obtain

$$q^*(\phi_t) \propto \textstyle\prod_l \phi_{tl}^{\beta - 1 + \sum_{d,n} \mathbb{1}(w_{dn}=l)\mathbb{1}(z_{dn}=t)},$$

which satisfies the assumption $q(\phi_t) = Dir(\phi_t|\lambda_t)$ where

$$\lambda_{tl} = \beta + \textstyle\sum_{d,n} \mathbb{1}(w_{dn} = l)\mathbb{1}(z_{dn} = t). \qquad (5)$$

The approximate distribution $q(z_{dn} = t)$ is proportional to

$$\begin{aligned} & \exp\big\{\mathbb{E}_{-z_{dn}}[\log P(z_{dn} = t)] + \mathbb{E}_{-z_{dn}}[\log P(w_{dn})]\big\} \\ &\propto \exp\Big\{\mathbb{E}_{-z_{dn}}[\log soft_t(f_d) + \log \textstyle\prod_l \phi_{tl}^{\mathbb{1}(w_{dn}=l)}]\Big\} \\ &\propto \exp\Big\{\zeta_{dt} + \sum_l \mathbb{1}(w_{dn} = l)\Big(\Psi(\lambda_{tl}) - \Psi(\sum_{l'} \lambda_{tl'})\Big)\Big\}, \end{aligned} \qquad (6)$$

where $\Psi(.)$ is the diagamma function, i.e., the first derivative of the log Gamma function. The last latent variable in the topic model branch to be approximated is f_d. By isolating the terms containing f_d, the objective function is

$$\begin{aligned}\mathcal{L}(q(f_d)) &= \sum\nolimits_n \mathbb{E}_q[\log P(z_{dn}|f_d)] \\ &\quad +\mathbb{E}_q[\log P(f_d|a_d,U,\varepsilon)] - \mathbb{E}_q[\log q(f_d)], \qquad (7)\\ \mathbb{E}_q[\log P(f_d|a_d,U)] &= -\frac{\varepsilon}{2}\sum\nolimits_t(\zeta_{dt}^2+\sigma_{dt}^{(f)2}) + \varepsilon\zeta_d'\mu\gamma_d,\\ \mathbb{E}_q[\log q(f_d)] &= -\sum\nolimits_t \log\sigma_{dt}^{(f)},\\ \mathbb{E}_q[\log P(z_{dn}|f_d)] &= \sum\nolimits_t \mathbb{1}(z_{dn}=k)\mathbb{E}_q[\log soft_t(f_d)].\end{aligned}$$

Here, $\mu = [\mu_1, ..., \mu_T] \in \mathbb{R}^{M\times T}$ is the collection of the means of $q(u_t) = \mathcal{N}(u_t|\mu_t, \Sigma_t^{(u)})$. Due to the normalization term in the softmax function, $\mathbb{E}_q[\log P(z_{dn}|f_d)$ does not have a close-form solution. Thus, reparameterization trick and Monto Carlo sampling [Kingma and Welling, 2013] are adopted to approximate this expectation. Since we assume $q(f_d) = \mathcal{N}(f_d|\zeta_d, \text{diag}(\sigma_d^{(f)}))$, f_d can be reparameterized as

$$f_d = \zeta_d + \sigma_d^{(f)} \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

where $\odot$ denotes the element-wise multiplication. Then, S samples can be obtained from $\epsilon^s \sim \mathcal{N}(0, I)$ and $\mathbb{E}_q[\log P(z_{dn}|f_d)$ can be rewritten as

$$(1/S)\sum\nolimits_s\sum\nolimits_t \mathbb{1}(z_{dn}=t)softmax(\zeta_d + \sigma_d^{(f)} \odot \epsilon^s).$$

The derivations of the three terms in $\mathcal{L}(q(f_d))$ (Eq. (7)) w.r.t. the variational parameter ζ_d are

$$\begin{aligned}&\nabla_{\zeta_d}\mathbb{E}_q[\log P(f_d|a_d,U,\varepsilon)] = \varepsilon(\mu\gamma_d - \zeta_d),\\ &\nabla_{\zeta_d}\mathbb{E}_q[\log q(f_d)] = 0,\\ &\nabla_{\zeta_d}\mathbb{E}_q[\log P(z_{dn}|f_d)]\\ =\ &(1/S)\sum\nolimits_s\sum\nolimits_t \mathbb{1}(z_{dn}=t)\big[e_t - softmax(f_d^s)\big]\\ =\ &\sum\nolimits_t \mathbb{1}(z_{dn}=t)e_t - (1/S)\sum\nolimits_s softmax(f_d^s).\end{aligned}$$

Similarly, the derivations of these terms w.r.t. $\sigma_d^{(f)}$ are

$$\begin{aligned}&\nabla_{\sigma_d^{(f)}}\mathbb{E}_q[\log P(f_d|a_d,U,\varepsilon)] = -\varepsilon\sigma_d^{(f)},\\ &\nabla_{\sigma_d^{(f)}}\mathbb{E}_q[\log q(f_d)] = -1/\sigma_d^{(f)},\\ &\nabla_{\sigma_d^{(f)}}\mathbb{E}_q[\log P(z_{dn}|f_d)] = 0,\end{aligned}$$

where e_t is a vector whose elements are all 0 except the t^{th} element being 1. Then the derivatives of $\mathcal{L}(q(f_d))$ (Eq. (7)) w.r.t. the variational parameter ζ_d and $\sigma_d^{(f)}$ are

$$\begin{aligned}\nabla_{\sigma_d^{(f)}}\mathcal{L}(q(f_d)) &= -\varepsilon\sigma_d^{(f)} + 1/\sigma_d^{(f)},\\ \nabla_{\zeta_d}\mathcal{L}(q(f_d)) &= \varepsilon(\mu\gamma_d - \zeta_d) + \sum\nolimits_n\sum\nolimits_t \mathbb{1}(z_{dn}=t)e_t.\\ &\quad -(N_d/S)\sum\nolimits_s softmax(f_d^s). \qquad (8)\end{aligned}$$

Thus, $\sigma_d^{(f)} = \sqrt{\varepsilon}$ and ζ_d can be updated with Adagrad.

$\mathcal{L}(q(H))$ is optimized w.r.t. the latent variables θ_k, $x_{i\to j}$, $x_{i\leftarrow j}$ and g_d in the stochastic block model branch. Similar to the derivation of $q(f_d)$, the variational parameters of $q(g_d) = \mathcal{N}(g_d|\psi_d, \text{diag}(\sigma_d^{(g)}))$ can be updated as follows. $\sigma_d^{(g)} = \sqrt{\tau}$, and ψ_d is updated with Adagrad according to

$$\begin{aligned}\nabla_{\psi_d}\mathcal{L}(q(g_d)) &= \tau(\nu\gamma_d - \psi_d) + \sum\nolimits_j\sum\nolimits_k \mathbb{1}(x_{d\to j}=k)e_k\\ &\quad -(D/S)\sum\nolimits_s softmax(g_d^s), \qquad (9)\end{aligned}$$

where $\nu = [\nu_1, ..., \nu_K] \in \mathbb{R}^{M\times K}$ is the collection of the means of $q(v_k) = \mathcal{N}(v_k|\nu_k, \Sigma_k^{(v)})$ and $g_d = \psi_d + \sigma_d^{(g)} \odot \delta, \delta \sim \mathcal{N}(0, I)$.

Next, we optimize $\mathcal{L}(q(H))$ w.r.t. $q(x_{i\to j})$ and obtain

$$\begin{aligned}&\chi_{i\to j}^k = q(x_{i\to j}=k)\\ &\propto \exp\big\{\mathbb{E}_{-x_{i\to j}}[\log softmax_k(g_d)]\\ &\quad +\sum\nolimits_h \mathbb{E}_{-x_{i\to j}}[P(x_{i\leftarrow j}=h)\log B_{kh}^{y_{ij}}(1-B_{kh})^{1-y_{ij}}]\big\}\\ &\propto \exp\{\phi_{ik}\}\prod\nolimits_h (B_{kh}^{y_{ij}}(1-B_{kh})^{1-y_{ij}})^{\chi_{i\leftarrow j}^h}. \qquad (10)\end{aligned}$$

Similarly, $\chi_{i\leftarrow j}^h = q(x_{i\leftarrow j}=h)$ is proportional to

$$\exp\{\phi_{jh}\}\prod\nolimits_k (B_{kh}^{y_{ij}}(1-B_{kh})^{1-y_{ij}})^{\chi_{i\to j}^k}. \qquad (11)$$

Until then, we have finished the E-step in the variational EM algorithm, which computes all the approximate distributions of the latent variables. In the M-step. the hyper-parameters $\alpha, \eta, \rho, \xi, \tau, \varepsilon, \beta$ are fixed and only the model parameter B is updated by maximizing ELBO according to the updated variational parameters. By isolating terms containing B and maximizing $\mathcal{L}(B)$, we obtain

$$B_{kh} = \big(\sum\nolimits_{i,j} y_{ij}\chi_{i\to j}^k\chi_{i\leftarrow j}^h\big)/\big(\sum\nolimits_{i,j}\chi_{i\to j}^k\chi_{i\leftarrow j}^h\big). \qquad (12)$$

By performing the E-step (variational parameters update) and M-step (model parameters update) alternatively, the posterior distribution can be effectively approximated. At last, γ_d, μ_t and ν_k are the embeddings of nodes, topics and communities.

Complexity Analysis

The complexity of the proposed variational algorithm is analyzed as follows. In Eq. (2), updating the means of variational topic embedding $\{\mu_t\}_{t=1}^T$ requires $\mathcal{O}(T(DM+M^2))$ operations. Updating the covariance of variational embedding $\Sigma^{(u)}$, which is independent of topic t and only need to be computed once, requires $\mathcal{O}(DM^2+M^3)$ operations. Similarly, the complexities of Eqs. (3) and (4) are $\mathcal{O}(K(DM+M^2)+DM^2+M^3)$ and $\mathcal{O}(D(TM+KM+M^2)+(T+K)M^2+M^3)$, respectively. In Eq (5), only a one-time traversal over all the words in every node is required, whose complexity is $\mathcal{O}(DN)$ where N is the average number of words in each node. The computations in Eq. (6) consists of two parts, computing $\Phi(.)$ for each pair of the topic and word and computing κ_{dn}^t. Their complexities are $\mathcal{O}(TL)$ and $\mathcal{O}(DN)$ respectively, because computing κ_{dn}^t only demands a single traversal over all the words. The complexities of Eqs. (8) and (9) are $\mathcal{O}(D(TM+N))$ and $\mathcal{O}(DKM+R)$.

Dataset	D	R	N_d	C
Texas	187	328	1,703	5
Cornell	195	304	1,703	5
Washington (WA)	230	446	1,703	5
Wisconsin (WI)	265	530	1,703	5
Citeseer	3,312	4,732	3,703	6
Cora	2,708	5,429	1,433	7
Wiki	3,363	45,006	4,972	19
Pubmed	19,729	44,338	500	3

Table 2: Datasets.

Due to the sparsity of network structre [Airoldi *et al.*, 2008], Eqs. (10) and (11) can be performed with only one traversal over all the edges, and thus their time complexities are $\mathcal{O}(R)$. Similarly, B can be updated via $\mathcal{O}(R)$ operations. In general, the total complexity of the proposed algorithm is $\mathcal{O}(M^3+(D+T+K)M^2+D(T+K)M+DN+TL+R)$ in each iteration, which is linear with the network scale.

4 Evaluations

4.1 Experimental Setups

Datasets. In the experiments, eight public real networks, as shown in Table 2, are employed. Four of them, i.e., Texas, Cornell, Washington and Wisconsin, are the sub-networks of WebKB network. Each of them is the collection of webpages from an American university. Similarly, nodes in Wiki network are webpages from Wikipedia. Citeseer, Cora and Pubmed are three citation networks whose nodes and edges are scientific publications with binary-valued word attributes and citation relationships, respectively.

Baseline Methods. To demonstrate the superiority of our proposed method, nine state-of-the-art embedding methods are employed for comparisons. These methods can be classified into two categories. The methods in the first category, such as DeepWalk [Perozzi *et al.*, 2014], node2vec [Grover and Leskovec, 2016], LINE [Tang *et al.*, 2015], GraRep [Cao *et al.*, 2015] and M-NMF [Cao *et al.*, 2015], only exploit the network topology information. The methods in the second category, such as TADW [Yang *et al.*, 2015], AANE [Huang *et al.*, 2017a], TriDNR [Pan *et al.*, 2016] and ASNE [Liao *et al.*, 2017] jointly exploit both the topological and non-topological information. Each baseline method employs the default settings in original paper. For fair comparison, the embedding dimension M is set to 64 for every method.

Parameter Settings. In the experiments, the hyper-parameters are set to $\beta = 1/T, \alpha = \rho = \eta = 0.1$ and $\tau = \varepsilon = 1$. For the number of communities and topics, we simply fix $T = K = 40$, though they are larger than the true number of groups C and $T + K > M$, for the following reasons: 1) We are interested in computing the embedding instead of directly detecting the communities and topics; 2) Setting $T + K > M$ reveals the correlations between topological and non-topological information more obviously.

4.2 Node Classification and Clustering

For node classification, the SVM implemented by Liblinear is adopted as the classifier. For each network, 10% of the nodes are randomly selected for training, while the rest of them is employed for testing. This process is repeated 10 times, and the average accuracy are reported in Table 3. The results indicate that the proposed method gives a 5% performance improvement (on average) compared to the state-of-the-art methods, which not only demonstrates the effectiveness and efficiency of the proposed model, but also shows the success of correlation exploration and information adaptation.

Methods	Cornell	Texas	WA	WI	Citeseer	Cora	Wiki	Pubmed
DeepWalk	38.46	48.09	53.92	49.62	48.42	75.52	60.61	78.36
Node2Vec	37.95	50.27	45.62	46.95	52.44	70.31	60.38	81.08
LINE	44.10	63.39	56.22	54.96	40.56	72.25	53.93	74.92
GraRep	53.33	69.40	51.15	60.31	53.61	76.02	63.74	81.37
MNMF	34.87	57.38	58.53	51.15	46.42	68.19	54.06	70.41
TADW	61.03	67.76	64.98	67.56	72.53	81.64	68.50	86.80
AANE	41.54	53.01	61.75	38.93	22.24	70.74	43.04	77.99
TriDNR	34.87	42.08	43.32	41.60	52.91	66.53	57.94	78.40
ASNE	45.64	59.02	55.76	59.92	44.35	71.49	29.87	77.20
Ours	**69.33**	**70.19**	**77.14**	**81.21**	**73.96**	**83.37**	**70.73**	**89.93**

Table 3: Node classification results (Accuracy).

Methods	Cornell	Texas	WA	WI	Citeseer	Cora	Wiki	Pubmed
DeepWalk	7.06	6.16	5.66	7.65	10.58	30.21	34.28	26.55
Node2Vec	6.65	4.49	2.94	7.86	12.99	32.10	33.87	25.02
LINE	9.27	23.16	24.95	9.39	5.62	28.95	29.01	7.17
GraRep	8.80	12.43	5.18	8.02	9.61	30.73	33.64	17.76
MNMF	11.63	17.20	22.15	10.10	8.95	18.00	19.23	1.41
TADW	11.13	10.90	11.63	17.52	31.60	39.90	37.92	20.11
AANE	9.55	3.52	13.19	2.86	1.19	17.57	15.49	0.01
TriDNR	7.20	4.32	8.10	6.60	9.59	35.99	32.37	19.28
ASNE	11.11	12.63	17.43	23.94	7.31	33.26	26.59	26.61
Ours	**23.28**	**28.63**	**31.57**	**30.02**	**35.55**	**43.72**	**40.61**	**29.33**

Table 4: Node clustering results (NMI).

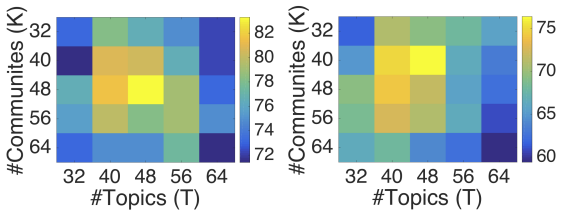
For node clustering, the k-means algorithm is applied to the learned embeddings. NMI is adopted to measure the clustering performance. This process is repeated 10 times for each network, and the average results are shown in Table 4. The proposed methods significantly outperform the existing state-of-the-art methods by preserving the mesoscopic properties in network structure and node content.

4.3 Correlation Analysis

To reveal the effectiveness of the correlations between the topological and non-topological information, the dimensions of the network structure subspace K and content subspace T vary from 32 to 64 when the dimension of embedding space is fixed as 64. The results are presented in Figure 3 and demonstrate several observations that: 1) The best results are achieved when both K and T are between 40 and 48, i.e., some (not all) dimensions of the two subspaces are aligned. 2) When $K = T = 32$, i.e., the two subspaces are independent, the performance is poor. 3) When either K or T is 64, i.e., one subspace contains the other one, the performance also degrades. 4) When $K = T = 64$, i.e., two subspaces are perfectly aligned, the performance is extremely poor. These observations satisfy the proposed assumptions of this paper.

5 Conclusions

Based on the three new assumptions about the embedding space and its properties, in this paper, nodes, communities



(a) Wisconsin (b) Washington

Figure 3: The effects of T and K in node classification task.

and topics are seamlessly mapped into one embedding space, and a novel correlated embedding approach is proposed to better utilize the correlations between the topological and non-topological information and adaptively weight the impacts of them. Extensive results demonstrate the superiority of the proposed method compared to the current state-of-the-art methods.

Acknowledgments

Supported by National Key R&D Program of China (No.2017YFC0820106, 2016YFC0801004), National Natural Science Foundation of China (No.61503281, U1636214, U1605252, 61332012, 61772361), Key Program of the Chinese Academy of Sciences (No. QYZDB-SSW-JSC003) and Fundamental Theory and Cutting Edge Technology Research Program of Institute of Information Engineering, CAS (No. Y7Z0381102).

References

[Airoldi *et al.*, 2008] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *JMLR*, 9(9):1981–2014, 2008.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *JMLR*, 3(1):993–1022, 2003.

[Cai *et al.*, 2017] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques and applications. *arXiv preprint arXiv:1709.07604*, 2017.

[Cao *et al.*, 2015] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Grarep: Learning graph representations with global structural information. In *CIKM*, pages 891–900. ACM, 2015.

[Chang and Blei, 2009] Jonathan Chang and David Blei. Relational topic models for document networks. In *AISTATS*, pages 81–88, 2009.

[Grover and Leskovec, 2016] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *SIGKDD*, pages 855–864. ACM, 2016.

[Hamilton *et al.*, 2017] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

[He *et al.*, 2017] Junxian He, Zhiting Hu, Taylor Berg-Kirkpatrick, Ying Huang, and Eric P Xing. Efficient correlated topic modeling with topic embedding. In *SIGKDD*, pages 225–233. ACM, 2017.

[Huang *et al.*, 2017a] Xiao Huang, Jundong Li, and Xia Hu. Accelerated attributed network embedding. In *ICDM*, pages 633–641. SIAM, 2017.

[Huang *et al.*, 2017b] Xiao Huang, Jundong Li, and Xia Hu. Label informed attributed network embedding. In *WSDM*, pages 731–739. ACM, 2017.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Liao *et al.*, 2017] Lizi Liao, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. Attributed social network embedding. *arXiv preprint arXiv:1705.04969*, 2017.

[Ou *et al.*, 2016] Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. Asymmetric transitivity preserving graph embedding. In *SIGKDD*, pages 1105–1114. ACM, 2016.

[Pan *et al.*, 2016] Shirui Pan, Jia Wu, Xingquan Zhu, Chengqi Zhang, and Yang Wang. Tri-party deep network representation. *Network*, 11(9):12, 2016.

[Perozzi *et al.*, 2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *SIGKDD*, pages 701–710. ACM, 2014.

[Tang *et al.*, 2015] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *WWW*, pages 1067–1077, 2015.

[Tu *et al.*, 2016] Cunchao Tu, Weicheng Zhang, Zhiyuan Liu, and Maosong Sun. Max-margin deepwalk: Discriminative learning of network representation. In *IJCAI*, pages 3889–3895, 2016.

[Tu *et al.*, 2017] Cunchao Tu, Han Liu, Zhiyuan Liu, and Maosong Sun. Cane: Context-aware network embedding for relation modeling. In *ACL*, pages 1722–1731, 2017.

[Wang *et al.*, 2016] Suhang Wang, Jiliang Tang, Charu Aggarwal, and Huan Liu. Linked document embedding for classification. In *CIKM*. ACM, 2016.

[Wang *et al.*, 2017a] Suhang Wang, Charu Aggarwal, Jiliang Tang, and Huan Liu. Attributed signed network embedding. In *CIKM*, pages 137–146. ACM, 2017.

[Wang *et al.*, 2017b] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *AAAI*, pages 203–209, 2017.

[Yang *et al.*, 2015] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. Network representation learning with rich text information. In *IJCAI*, pages 2111–2117, 2015.

[Yang *et al.*, 2018] Liang Yang, Yuanfang Guo, and Xiaochun Cao. Multi-facet network embedding: Beyond the general solution of detection and representation. In *AAAI*, pages 499–506, 2018.