Graph Relational Topic Model with Higher-order Graph Attention Auto-encoders

Qianqian Xie
Department of Computer Science,
University of Manchester
xieq@whu.edu.cn

Jimin Huang
School of Computer Science,
Wuhan University
huangjimin@whu.edu.cn

Pan Du
Department of Computer Science and Operations Research, University of Montreal
pandu@iro.umontreal.ca

Min Peng
School of Computer Science,
Wuhan University
pengm@whu.edu.cn

Abstract

Learning low-dimensional representations of networked documents is a crucial task for documents linked in network structures. Relational Topic Models (RTMs) have shown their strengths in modeling both document contents and relations to discover the latent topic semantic representations. However, higher-order correlation structure information among documents is largely ignored in these methods. Therefore, we propose a novel graph relational topic model (GRTM) for document network, to fully explore and mix neighborhood information of documents on each order, based on the Higher-order Graph Attention Network (HGAT) with the log-normal prior in the graph attention. The proposed method can address the aforementioned issue via the information propagation among document-document based on the HGAT probabilistic encoder, to learn efficient networked document representations in the latent topic space, which can fully reflect document contents, along with document connections. Experiments on several real-world document network datasets show that, through fully exploring information in documents and document networks, our model achieves better performance on unsupervised representation learning and outperforms existing competitive methods in various downstream tasks.

1 Introduction

Document networks, such as hyperlink networks of Web pages, citation networks of academic documents, and user profiles in social networks, have long been an intensively studied research subject due to their wide applications. Finding low-dimensional representations of networked documents to preserve document contents and connections among documents simultaneously is a crucial research task. Inspired by the wide application of topic models such as latent Dirichlet allocation (LDA) (Blei et al., 2003) on discovering the latent semantic structure of unconnected documents, a series of Relational Topic Models (RTMs) are proposed to explore the latent topic semantic structure of documents and links among them, based on probabilistic graphical models (Nallapati et al., 2008; Chang and Blei, 2009; Le and Lauw, 2014; Chen et al., 2014; Yang et al., 2016), deep generative models (Acharya et al., 2015; Wang et al., 2017; Bai et al., 2018), auto-encoders (AEs) (Zhang and Lauw, 2020) and graph auto-encoders (GAEs) (Wang et al., 2020a).

However, most RTMs consider only the pairwise correlation or the first-order neighbor correlation (Zhang and Lauw, 2020) among documents. Although the recently proposed deep relational topic model, GPFA (Wang et al., 2020a) based on graph neural networks (GNNs) can consider low-order indirect neighborhood information via stacked graph neural network (GNN) (Kipf and Welling, 2016b) layers, it still suffers from exploiting the deep interactions (higher-order) between indirectly connected documents due to the over-smoothing problem (Li et al., 2018). Such the higher-order correlation structure has been proved to be effective on various tasks (Abu-El-Haija et al., 2019) such as link prediction and recommendation (Zhang and McAuley, 2020).

To address the aforementioned issue, we propose the graph relational topic model (GRTM) for modeling the latent topic structure of document contents and links, based on the higher-order graph attention auto-encoders (HGTAEs), aiming to fully explore and fuse each order proximity (including the low-order and higher-order) of document network. Specifically, we propose to extract the higher-order document proximity network (HDPN) from the adjacent matrix of the document network via the calculation of the shortest path. The higher-order graph attention network (HGAT) is presented

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2604–2613

to efficiently model the neighborhood information propagation on HDPN via introducing the log-normal prior into the graph attention. We finally propose our GRTM with the higher-order graph attention auto-encoders (HGTAEs) based on HGAT and HDPN. The main contributions of our paper are as follows:

1. We propose a novel unsupervised deep relational topic model GRTM to fully explore multiple information: the higher-order document relations and latent topic semantic among document contents and networks.

2. We propose a novel graph attention network HGAT to efficiently explore each order correlations among networked documents.

3. Experimental results on document network datasets show that our model outperforms existing competitive methods on unsupervised representation learning, through fully exploring multi-granularity information in document networks.

2 Related Work

In this section, we briefly review existing Relational Topic Models (RTMs), Graph Auto-Encoders (GAEs), and Graph Topic Models.

RTMs generally extended LDA based topic models to further model the links between documents in networks. Chang and Blei (2009) first proposed to introduce additional binary conditional variables in the generation to model the document links. (Chen et al., 2014) (2014) proposed discriminative relational topic models (DRTMs) to learn discriminative latent representations of document networks. Le and Lauw (2014) proposed PLANE which can jointly extract topics and visualization coordinates. To apply the neural network based inference approach to RTMs, Bai et al. (2018) utilized Stacked Variational Auto-Encoder(SVAE) to derive more representative documents in topic distributions. However, these models only consider pair-wise document correlations, fail to model the full structural information (low-order and higher-order) embedded in the document network.

To model the block correlation structure of the document network, Yang et al. (2016) incorporated weighted stochastic block model into relational topic models. Most recently, Zhang and Lauw (2020) proposed AdjEnc to reconstruct both documents and their neighborhoods in the network. However, it can only capture the first-order correlation structure with the adjacent-encoder. Wang et al. (2020a) proposed the deep relational topic model GPFA based GNNs to explore hierarchical relationships of interconnected documents. However, still, it can only capture the low-order hierarchical relationships of interconnected documents due to the well-known smoothing problem of GNNs, while long-range relations among documents are also critical for learning latent representations in document networks. To address this issue, we calculate the higher-order proximity network that allows considering the long-range topological information among documents, rather than merely pairwise or few-order relations.

Recently, GAEs has attracted a lot of attention, which incorporates GNNs into auto-encoder to unsupervised graph embedding learning, motivated by the successful applications of GNNs in modeling graph topological structure. The earliest attempt VGAE (Kipf and Welling, 2016a) extended variational auto-encoder (VAE) onto graph structure data for learning network embedding. Inspired by the advantage of GNNs, some works have explored VGAE for topic modeling, including the deep relational topic model GPFA mentioned before, and GraphBTM (Zhu et al., 2018) which improved the biterm topic model (Yan et al., 2013) with word co-occurrence graph encoded by GCNs. Except studies based on VGAE, there are also works combining topic models with graph neural networks in a different manner, such as the graph attention topic network (GATON) (Yang et al., 2020) proposed for unconnected documents, the dynamic hierarchical topic graph model DHTG (Wang et al., 2020b) used for unconnected document classification, the topic variational graph auto-encoder (TVGAE) (Xie et al., 2021b) for document classification and the graph topic neural network (GTNN) (Xie et al., 2021a) proposed for representation learning of both connected and unconnected documents. Different from them, we target connected documents. Moreover, to fully explore the deep topological structure of document networks, we propose the novel higher-order graph attention network, and then introduce it into the relational topic modelling based on variational graph auto-encoders.

3 Method

In this section, we present our graph relational topic model (GRTM) for the document network. We first introduce the construction of the higher-order proximity networks: HDPN from document adjacency matrices, then we present the novel graph attention network HGAT to fuse the information of HDPN. We end this section by introducing the variational graph auto-encoder structure for building GRTM.

3.1 Higher-order Proximity Network

Formally, we define a given document network as $\mathcal{G} = (D, A, X)$. $D = \{d_1, ..., d_n\}$ is the set of document nodes with n documents and a vocabulary V with m words. Relations between documents are represented as a 0-1 adjacency matrix $A \in \mathbb{R}^{n \times n}$, and $X \in \mathbb{R}^{n \times m}$ is the documen-word index matrix, in which X_{ij} represents the weight (e.g. TF-IDF) of word j in document i. For the given document sets D, Based on the given adjacency matrices A of document network, the key problem is to discover and preserve arbitrary-order neighborhood relations beyond first-order or few-order (including other higher-order). Intuitively, two nodes have a proximity correlation if and only if we can find at least one path between them (Liu et al., 2019). Thus, we can calculate the order of proximity correlation between two nodes according to the length of the shortest path between them based on the adjacency matrix, and directly preserve arbitrary-order information in the same matrix. Denoting the adjacency matrices of HDPN as $\hat{A} \in \mathbb{R}^{n \times n}$, the link of proximity correlation between two documents (d_i, d_j) is defined as:

$$\hat{A}_{i,j} = \begin{cases} k, & \textit{existing a k-length shortest path} \\ 1, & i = j \\ \infty, & \textit{no path} \end{cases} \tag{1}$$

According to the above definition, $\hat{A}$ can be calculated during the data pre-processing step in advance. The length of the shortest path of two nodes is calculated using classical search algorithms such as Dijkstra's algorithm or Bellman-Ford algorithm on the machine learning framework [1]. Compared with existing methods that calculate the higher-order proximity with the power of adjacency matrix or steps in a probabilistic transition process (Abu-El-Haija et al., 2019; Liu et al., 2019), our calculation is more suitable for explicitly calculating the length of the shortest path. Because the k-power of adjacency matrix has proximity information overlap on other power matrices before it, while the calculation of k-walk may lead to nodes return to their neighbors less than k-order rather than reach to their k order neighbor (Zhang and Xu, 2020).

[1] https://networkx.github.io/documentation/networkx-1.10/overview.html

3.2 Higher-order Graph Attention Network

In this section, we focus on how to better fuse information of neighbors at different orders on an HDPN to efficiently learn node representations. Intuitively, for a given node representation, the contributions of its neighbors vary according to their distances. However, directly utilizing GNNs such as graph convolutional networks (GCNs) and graph attention networks (GATs) on the HDPN will treat neighbors of nodes at different orders equally. Therefore, we present the higher-order graph attention network (HGAT) to solve the problem via introducing the log-normal prior into the graph attention.

Instead of utilizing uniform prior in GATs, we exploit the log-normal distribution to model the importance decaying of neighbors of the current node on different orders. For simplicity sake, we use the log-normal distribution with zero-mean value, and calculate the attention coefficient between two nodes i, j:

$$\begin{aligned} \tilde{h_i}^l, \tilde{h_j}^l &= W^l h_i^l + b^l, W^l h_j^l + b^l \\ e_{ij} &= \rho(\lambda \phi(p_{ij})(\tilde{h_i}^l \cdot (\tilde{h_j}^l)^\mathsf{T}) \\ \alpha_{i,j} &= softmax(e_{i,j}) = \frac{\exp(e_{ij})}{\sum_{o \in \mathcal{N}(i)} \exp(e_{io})} \end{aligned} \tag{2}$$

where $\phi(p) = \frac{\exp^{-(\ln p)^2/\sigma^2}}{p\sigma\sqrt{2\pi}}$ is the probability density function of the log-normal distribution, σ is the variance of the log-normal prior, p_{ij} is the length of the shortest path between nodes i, j in a HPN, h_i^l, h_j^l are representations of node i, j in the l-th layer, ρ is the activation function, λ is the parameter to control the influence of the log-normal prior, $\mathcal{N}(i)$ is neighbors of node i in a HPN, W^l, b^l are the weight matrix and bias of the l-th layer. The attention mechanism based on the log-normal prior allow nodes to select their neighbors at arbitrary-order with different importance. Calculated attention coefficients are further exploited to propagate information of neighbors of each node at arbitrary-

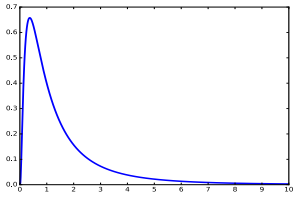


Figure 1: The example of the log-normal distribution with $\mu = 0, \sigma = 1$.

order:

$$h_i^{l+1} = \rho(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} h_j^l) \tag{3}$$

Although there are other distributions, such as the normal distribution used in the Gaussian transformer (Guo et al., 2019), the log-normal is more suitable for model the importance calculation in our case. This is because the density function $\phi(p)$ of the log-normal prior with zero-mean value has always a real value larger than 0 rather than that of the normal distribution with the negative value or Poisson distribution with an integer value. Moreover, $\phi(d)$ always decreases monotonically after $\exp^{-\sigma^2} < 1$, while the path length in our HDPN is always greater or equal to 1. Therefore, the log-normal prior can naturally model the decay of importance weight by increasing the path length between two nodes, as shown in Figure 1. When conducting the HGAT on the vanilla adjacency matrix (only with first-order proximity), the HGAT will be degenerated into the GAT method due to there are only 1 or ∞-length of the shortest path in the vanilla adjacency matrix.

3.3 Graph Relational Topic Modelling

To fully explore the long-range document relations to model the latent topic semantic among document contents and networks, we present the GRTM model with the higher-order graph attention auto-encoders (HGTAEs). Let's assume K is the topic number, θ is the document topic proportion, and β is the topics, namely the topic word proportion.

Firstly, we present the generative process of GRTM as in Algorithm 1. Similar to previous RTMs, we assume the document topic proportion is generated from the Dirichlet prior. However, the Dirichlet prior makes it difficult to make the neural variational inference for GRTM, due to the challenge of reparameterizing the Dirichlet prior. Thus, to simplify the inference process, we approximate the Dirichlet distribution with its Laplace approximation: the logistic normal distribution following many previous works (Srivastava and Sutton, 2017).

Algorithm 1: Generative Process of GRTM

for *each document* $d \in D$ **do**
- Generate the mean vector: $\mu_d^0 \sim \text{HGAT}_\mu(X_d, \hat{A}_d)$.
- Generate the diagonal covariance: $\sigma_d^0 \sim diag(\text{HGAT}_\sigma(X_d, \hat{A}_d))$.
- Draw the noise variable $\epsilon_d \sim \mathcal{N}(0, I)$.
- Draw the document topic proportion: $\theta_d \sim LN(\mu_d^0, \sigma_d^0) = softmax(\mu_d + (\sigma_d^0)^{\frac{1}{2}}\epsilon_d)$.
- **for** *each word* $w_{dv} \in V$ **do**
  - Draw the word $w_{dv}|\theta_d, \beta \sim \text{Mult}(\theta_d \beta)$.

for *each pair of document* $d, d' \in D$ **do**
- Draw the observed link $A_{d,d'}|\theta_d, \theta_{d'} \sim \text{Bernoulli}(f_y(\theta_d, \theta_{d'}))$

To incorporate the higher-order relations among documents, we generate the document topic proportion with logistic normal distribution parameterized by HGAT probabilistic encoder. Specifically, for each document d, we draw the mean and covariance of a multinomial distribution variable and then transform it with the softmax function:

$$\begin{aligned} \mu_d^0 &= \text{HGAT}_\mu(X_d, \hat{A}_d) \\ \sigma_d^0 &= diag(\text{HGAT}_\sigma(X_d, \hat{A}_d)) \\ \theta_d &= softmax(\mu_d^0 + (\sigma_d^0)^{\frac{1}{2}}\epsilon_d) \end{aligned} \tag{4}$$

where HGAT is the message passing process as in Equation 3, ϵ_d is the noise variable. For HGAT encoders of mean and covariance, the input feature h_d^0 is set to the normalized document-word index feature X_d following previous methods (Kipf and Welling, 2016a). The message passing based on HGAT makes latent topic proportions of each document influenced by its neighbors at different orders with different importance.

In the decoding process, the word is generated from the multinomial distribution based on the topic proportion of the document it belongs to and its topic proportion: $p(w_d|\theta_d, \beta) = \text{Mult}([\theta_d \beta])$. The links between two documents are modeled as Bernoulli binary variables, which are conditionally generated based on the latent topic proportions of these documents: $p(A_{d,d'}|\theta_d, \theta_{d'}) = sigmoid(f_y(\theta_d, \theta_{d'}))$, where f_y is the multi-layer perception.

Following the auto-encoding variational Bayes inference method (Kingma and Welling, 2014), we can yield the evidence lower bound (ELBO) to the marginal log-likelihood according to the above

generative process:

$$\begin{aligned}\mathcal{L}(\Theta) = &- D_{KL}[q(\theta|w,\hat{A})||p(\theta|\alpha)] \\ &+ \mathbb{E}_{q(\theta|w,\hat{A})} logp(w|\theta,\beta) \\ &+ \mathbb{E}_{q(\theta|w,\hat{A})} logp(A|\theta)\end{aligned} \tag{5}$$

where Θ is the parameter set of the whole process, $q(\theta|w,\hat{A})$ is the approximate Dirichlet variational posterior as parameterized in Equation 3, $p(\theta|\alpha)$ is assumed the true Dirichlet posterior and α is the prior parameter. We still approximate it with its Laplacian approximation: the softmax variable on the multivariate normal with mean and covariance matrix as follows:

$$\begin{aligned}\mu_d^1 &= log\alpha - \frac{1}{k}\sum_i \alpha_i \\ \sigma_d^1 &= \frac{1}{\alpha}(1-\frac{2}{K}) + \frac{1}{k^2}\sum_i(\frac{1}{\alpha_i})\end{aligned} \tag{6}$$

We seek to minimize the KL divergence between the variational posterior and the true posterior in the first term. The second and third terms aim to reconstruct the document contents and links.

Based on the gradient variational Bayes (SGVB) estimator (Kingma and Welling, 2014), we can further yield the detailed formulation of each term:

$$\begin{aligned}D_{KL} = &\frac{1}{2}\{tr(\sigma^0(\sigma^1)^{-1}) \\ &+ (\mu^1-\mu^0)^T(\sigma^1)^{-1}(\mu^1-\mu^0) \\ &- k + log(\frac{|\sigma^1|}{|\sigma^0|})\}\end{aligned} \tag{7}$$

$$\begin{aligned}logp(w|\theta,\beta) = &\sum_{d=1}^{n} -X_d log(\tilde{X}_d) \\ &- (1-X_d)log(1-\tilde{X}_d) \\ logp(A|\theta) = &\sum_{d=1}^{n} -A_d log(\tilde{A}_d) \\ &- (1-A_d)log(1-\tilde{A}_d)\end{aligned} \tag{8}$$

where $\tilde{X} = (\theta\beta)$ is the reconstructed document contents, $\tilde{A} = sigmoid(f_y(\theta))$ is the reconstructed document links. Based on these, we can optimize the ELBO with stochastic gradient descent to infer the whole model end to end.

Table 1: Statistics of the document network datasets (Zhang and Lauw, 2020)

Datasets	Classes	Documents	Edges	Vocabulary
DS	9	570	1336	3,085
HA	6	223	515	2,073
ML	7	1,980	5,748	4,431
PL	9	1,553	4,851	4,105

4 Experiments

We conduct experiments in several real-world document network datasets. The statistics are reported in Table 1. Four datasets are subsets extracted from Cora: Data Structure (DS), Hardware and Architecture (HA), Machine Learning (ML), and Programming Language (PL) as in (Zhang and Lauw, 2020), in which Cora is the scientific article citation dataset collected from scholar websites. To evaluate the unsupervised representation learning capability of our method, we infer the latent topic portions of documents θ with our model, and then use it for three types of downstream tasks, namely document classification, document clustering, link prediction. We compare our method against baselines from the following three categories:

- **Auto-Encoders**: including variants of auto-encoders such as AE, DAE (Vincent et al., 2010), CAE (Rifai et al., 2011), VAE (Kingma and Welling, 2014), KSAE (Makhzani and Frey, 2014), KATE (Chen and Zaki, 2017) use Auto-Encoder to encode texts;
- **Relational Topic Models**: including generative models based on relational topic model such as RTM (Chang and Blei, 2009), PLANE (Le and Lauw, 2014), NRTM (Bai et al., 2018), ProdLDA (Srivastava and Sutton, 2017), and also a relational topic model based on the auto-encoder method: ADE (Zhang and Lauw, 2020);
- **Graph Embedding**: including graph embedding method based on GCN - VGAE (Kipf and Welling, 2016a).

We follow the settings for all baselines as in (Zhang and Lauw, 2020) and also compare methods in both transductive and inductive learning settings. For inductive learning, we randomly select a subset of 70% documents as the training set, a subset of 10% of documents as the validation set, and use the remaining 20% documents as the testing set. For transductive learning, all documents are involved

in the training process. All experimental results are averaged over the results of 10 independent runs. Following (Zhang and Lauw, 2020), we set the topic number K as 64, the layer number L of message passing in HGAT as 1. The hidden size of weight matrices in HGAT is equal to the topic number of 64. For the log-normal prior in HGAT, we set the the parameter $\lambda = \sqrt{\pi}$, the variance $\sigma = \frac{1}{\sqrt{2}}$. We use the tanh activation function in HGAT. We use the Dirichlet distribution with parameter $\alpha = \frac{1}{K}$ for the logistic normal approximation. The learning rate on all datasets is 0.065, the maximum training epochs with Adam is 40000, the early stop epoch is 500. The parameter setting in all baseline models is the same as in (Zhang and Lauw, 2020).

We infer document topic representations with trained GRTM model for both train and test documents, which are then used in three downstream tasks to evaluate the effectiveness: 1) **Document classification**: we adopt K-Nearest Neighbors to predict each document's label based on the Euclidean distance of generated representations. We use the classification accuracy as the metric. 2) **Document clustering**: We also compare our method with baselines in clustering documents via K-means, to investigate whether our method can generate similar representations for documents in the same category. In this case, the ground truth labels are only utilized to calculate normalized mutual information (NMI) in evaluation. 3) **Link Prediction**: The generated representations are used to predict the links between documents in this experiment. We use Mean Average Precision (MAP) as the evaluation metric following the previous method (Zhang and Lauw, 2020). To better understand the semantic information our method captured in generated representations, we also conduct experiments to present a detailed analysis of our generated topics in inductive learning: 1) **Topic Coherence**: As in previous work (Zhang and Lauw, 2020), we adopt PMI - $PMI(w_i, w_j) = log\frac{p(w_i,w_j)}{p(w_i)p(w_j)}$ to evaluate topic coherence. We calculate the average pairwise PMI of the top 10 words in each topic. Better topics should produce higher PMI.

2) **Visualization**: We apply t-SNE to project text representations generated by different models into a 2-dimensional space.

4.1 Overall Results

As shown in Table 2 and 3, our method achieves the best performance in all tasks on the four datasets on both inductive and transductive settings. Compared with auto-encoder based methods (AE, DAE, CAE, VAE, KSAE, and KATE), which only consider document contents without document networks. ADE, relational topic model methods (ProdLDA, RTM, PLANE, NRTM), and our method achieve better performance due to considering the links among documents, which benefits the downstream tasks of document classification/clustering and link prediction. Compared with relational topic model methods (ProdLDA, RTM, PLANE, NRTM), we found that the graph embedding method VGAE and the adjacent auto-encoder method ADE perform better than other baselines, which demonstrates the advantage of using high-order proximity information. But they are still inferior to our proposed GRTM, which proves the benefits of fully exploring information of various orders in document networks. A similar performance among these methods can also be observed in the results of topic coherence in Table 4.

There is no mean standard deviation evaluation by the previous methods (Zhang and Lauw, 2020; Chen and Zaki, 2017), so we only report the results of our method in Table 5 to illustrate the statistical effectiveness of our model. These results are obtained in both transductive and inductive settings through repeating each 10 times.

4.2 Effect of Topic Number

To investigate the sensitivity of our method to topic numbers, we present the classification accuracy of our model on different topic numbers in the inductive setting. As shown in Figure 2, the test accuracy on four datasets generally improves with the increase of the number of topics and reaches the peak when the topic number is around 64. From these curves, we can find that the performance of our model is not too sensitive to the topic number, and also the topic number does not seem to be so related to the ground truth number of classes of datasets. In figure 3, we further show transductive test classification accuracy of different models under different topic numbers. We can see that our model consistently outperforms all baselines under different topic numbers on four datasets.

Table 2: Transductive learning results on document classification, document clustering, and link prediction.

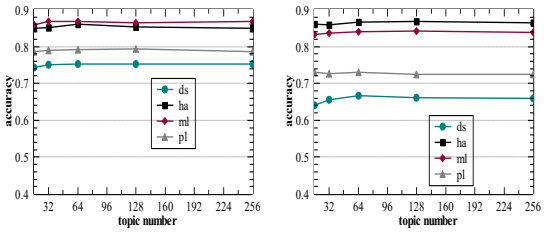
Model	Document Classification				Document Clustering				Link Prediction			
	DS	HA	ML	PL	DS	HA	ML	PL	DS	HA	ML	PL
AE	0.558	0.688	0.739	0.616	0.250	0.315	0.368	0.230	0.144	0.195	0.107	0.102
DAE	0.656	0.799	0.790	0.694	0.372	0.409	0.441	0.278	0.204	0.296	0.121	0.147
CAE	0.558	0.685	0.741	0.620	0.261	0.309	0.371	0.228	0.145	0.188	0.108	0.103
VAE	0.652	0.789	0.796	0.679	0.356	0.394	0.447	0.286	0.193	0.283	0.122	0.135
KSAE	0.537	0.672	0.710	0.581	0.245	0.295	0.345	0.222	0.136	0.182	0.092	0.088
KATE	0.628	0.808	0.762	0.651	0.325	0.378	0.342	0.267	0.174	0.267	0.095	0.114
ProdLDA	0.637	0.780	0.764	0.631	0.374	0.460	0.423	0.289	0.162	0.324	0.080	0.095
RTM	0.543	0.637	0.663	0.574	0.082	0.094	0.126	0.127	0.117	0.194	0.072	0.075
PLANE	0.690	0.799	0.750	0.648	0.417	0.406	0.439	0.288	0.284	0.226	0.107	0.160
NRTM	0.591	0.816	0.549	0.503	0.313	0.404	0.137	0.190	0.149	0.221	0.036	0.049
VGAE	0.671	0.827	0.807	0.718	0.335	0.362	0.495	0.308	0.285	0.265	0.132	0.171
ADE	0.744	0.846	0.857	0.780	0.445	0.548	0.571	0.392	0.374	0.326	0.251	0.271
DGTAE	**0.753**	**0.860**	**0.869**	**0.792**	**0.501**	**0.562**	**0.592**	**0.416**	**0.402**	**0.340**	**0.270**	**0.294**

Table 3: Inductive learning results on document classification, document clustering, and link prediction.

Model	Document Classification				Document Clustering				Link Prediction			
	DS	HA	ML	PL	DS	HA	ML	PL	DS	HA	ML	PL
AE	0.405	0.580	0.632	0.509	0.213	0.337	0.340	0.248	0.185	0.233	0.181	0.129
DAE	0.516	0.749	0.732	0.595	0.375	0.436	0.415	0.299	0.347	0.286	0.259	0.198
CAE	0.400	0.573	0.644	0.519	0.212	0.279	0.362	0.253	0.192	0.232	0.185	0.132
VAE	0.491	0.785	0.738	0.594	0.373	0.361	0.404	0.300	0.391	0.346	0.243	0.192
KSAE	0.390	0.569	0.614	0.491	0.269	0.319	0.334	0.232	0.188	0.238	0.148	0.111
KATE	0.484	0.800	0.712	0.573	0.321	0.440	0.354	0.290	0.277	0.336	0.205	0.178
ProdLDA	0.202	0.401	0.184	0.158	0.302	0.292	0.399	0.306	0.220	0.297	0.192	0.140
RTM	0.327	0.498	0.652	0.564	0.000	0.046	0.091	0.048	0.260	0.276	0.210	0.149
PLANE	0.282	0.544	0.275	0.218	0.162	0.192	0.000	0.000	0.306	0.345	0.176	0.134
NRTM	0.456	0.811	0.482	0.408	0.339	0.398	0.167	0.207	0.076	0.097	0.020	0.049
VGAE	0.509	0.748	0.736	0.607	0.280	0.185	0.442	0.291	0.315	0.309	0.237	0.274
ADE	0.640	0.845	0.836	0.724	0.416	0.489	0.522	0.363	0.400	0.427	0.363	0.322
GRTM	**0.687**	**0.867**	**0.841**	**0.731**	**0.466**	**0.564**	**0.535**	**0.394**	**0.432**	**0.449**	**0.538**	**0.358**

Table 4: Topic Coherence

Model	PMI			
	DS	HA	ML	PL
AE	0.294	0.446	0.665	0.969
DAE	1.170	1.125	1.203	1.553
CAE	0.348	0.558	0.526	0.684
VAE	0.685	0.793	1.831	1.132
KSAE	0.547	0.285	0.770	0.759
KATE	1.312	1.755	1.619	2.003
ProdLDA	1.638	1.315	1.837	2.088
RTM	1.279	1.678	1.199	1.615
PLANE	1.585	1.847	1.756	2.099
NRTM	1.533	2.041	1.328	1.632
ADE	1.872	1.887	2.337	2.321
GRTM	**2.073**	**2.361**	**2.512**	**2.610**



(a) Transductive (b) Inductive

Figure 2: The test classification accuracy of our model vary different topic numbers

4.3 Effect of Log-normal Prior

We vary the variance σ of the log-normal prior to explore the impact of the log-normal prior in HGAT. We present results of our model under the value of $\sigma \in \{\frac{1}{2}, \frac{1}{\sqrt{2}}, 1\}$, and the results are shown in Table 7. The larger value of σ makes the slower decay of the importance on higher-order proximity information otherwise the faster decay. From the table, we can see that our model generally achieves the best performance at $\sigma = \frac{1}{\sqrt{2}}$. We suspect that too much noisy higher-order information is introduced when σ is set too large, while insufficient higher-order information can be used when too small. Hence it yields poor performance in both cases.

4.4 Different layer numbers

Although one layer message passing process of HGAT is able to capture arbitrary-order proximity information among document networks, we can still report the results of our model under the different layers of HGAT in Table 6. We can see that our model with one layer HGAT achieves the best performance under both settings. When the layer number of HGAT is set to 0, the HGAT encoder of our model degenerates to the feed-forward neural network. With two-layer HGAT, document representations may be disturbed by the information of their noisy neighbors.

Table 5: Mean ± standard deviation results of our model on document classification, document clustering.

Model	Document Classification				Document Clustering			
	DS	HA	ML	PL	DS	HA	ML	PL
Trans	0.753 ± 0.021	0.86 ± 0.032	0.869 ± 0.016	0.792 ± 0.030	0.501±0.009	0.562 ±0.011	0.592 ±0.0014	0.416 ±0.008
Induc	0.687 ± 0.018	0.867 ± 0.024	0.841 ± 0.020	0.731 ± 0.021	0.466 ± 0.011	0.564 ± 0.012	0.535 ± 0.015	0.394 ± 0.009

Table 6: Transductive results of our model on document classification, document clustering, and link prediction under different layer numbers of HGAT.

Layer Number	Document Classification				Document Clustering				Link Prediction			
	DS	HA	ML	PL	DS	HA	ML	PL	DS	HA	ML	PL
0	0.731	0.852	0.850	0.784	0.486	0.553	0.572	0.401	0.382	0.324	0.247	0.269
1	**0.753**	**0.860**	**0.869**	**0.792**	**0.501**	**0.562**	**0.592**	**0.416**	**0.402**	**0.340**	**0.270**	**0.294**
2	0.742	0.849	0.855	0.780	0.488	0.546	0.568	0.406	0.390	0.327	0.255	0.271

4.5 Ablation Study

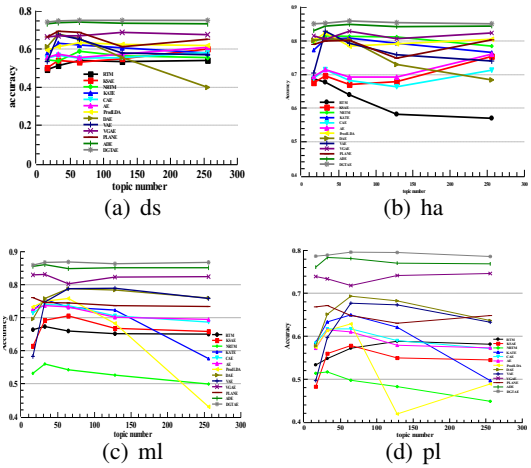
We also perform an ablation study on our method to verify the effectiveness of each module in the inductive setting. We compare our model with its variants by removing one of the components HDPN and HGAT respectively, as shown in Table 8. From which we can see that each component makes a certain contribution to the overall performance. In the case of removing the HDPN (W/HDPN) module, our model directly takes the document adjacency matrix as input, in which it degenerates into the relational topic model based on graph attention auto encoder without considering the higher-order proximity. As in the case of removing the HGAT (W/HGAT), although our model takes the higher-order information into consideration, it doesn't make the important selection for different order correlation information. We can also see that missing the higher-order proximity has a more significant negative influence than missing the HGAT based encoder module, illustrating the relative effectiveness of the higher-order information in improving the discrimination of latent representations.

4.6 Visualization

Finally, to intuitively demonstrate the effectiveness of our model, we visualize the learned representations of the test documents on the ML dataset in Figure 4. It shows that documents are better grouped by our model than ADE (with first-order correlations) and VGAE (with few-order correlations), due to the incorporation of the in-direct correlation information among documents.

5 Conclusion

In this paper, we propose a novel graph relational topic model GTM for document networks to fully explore each order of relations among document networks, which is efficiently fused by a proposed novel graph attention network HGAT equipped with log-normal attention prior. Experimental results show that full consideration of each order proximity information on the document-document graph is beneficial for improving the learned document representations. In future work, we would like to explore the better-suited method and more elegant prior distributions for discovering and fus-



(a) ds (b) ha (c) ml (d) pl

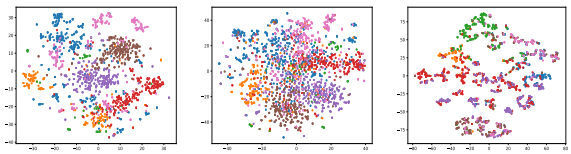
Figure 3: The test classification accuracy of different models vary different topic numbers

Table 7: Results of our model based on vary values of σ on document classification, document clustering, and link prediction.

σ	Document Classification		Document Clustering		Link Prediction	
	DS	HA	DS	HA	DS	HA
1	0.739	0.861	**0.52**	0.556	0.389	0.332
$\frac{1}{\sqrt{2}}$	**0.753**	**0.861**	0.501	0.562	**0.402**	**0.340**
$\frac{1}{2}$	0.736	0.85	0.498	**0.599**	0.396	0.338

Table 8: Ablation Study

Module	document classification				document clustering			
	DS	HA	ML	PL	DS	HA	ML	PL
All	**0.687**	**0.867**	**0.841**	**0.731**	**0.466**	**0.564**	**0.535**	**0.394**
W/HDPN	0.631	0.833	0.814	0.705	0.427	0.533	0.510	0.366
W/HGAT	0.651	0.852	0.827	0.710	0.450	0.557	0.524	0.390



(a) DGTAE (b) ADE (c) VGAE

Figure 4: The t-SNE visualization of document representation learned by different models in ML under the inductive setting. (Each color denotes one categorical label of documents)

ing higher-order proximity in document networks.

Acknowledgments

We thank Benyou Wang for his valuable comments on this manuscript. We also would like to thank the anonymous reviewers for their constructive comments. We thank MindSpore for the partial support of this work, which is a new deep learning computing framework[2]. This research is supported by the CSC Scholarship offered by China Scholarship Council and the joint laboratory on Credit Technology.

References

Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, H. Harutyunyan, Nazanin Alipourfard, Kristina Lerman, G. V. Steeg, and A. Galstyan. 2019. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *ICML*.

Ayan Acharya, Dean Teffer, Jette Henderson, Marcus Tyler, Mingyuan Zhou, and Joydeep Ghosh. 2015. Gamma process poisson factorization for joint modeling of network and documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 283–299. Springer.

Haoli Bai, Zhuangbin Chen, Michael R. Lyu, Irwin King, and Zenglin Xu. 2018. Neural relational topic models for scientific article analysis. *CIKM*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Jonathan Chang and David M. Blei. 2009. Relational topic models for document networks. In *AISTATS*.

Ning Chen, Jun Zhu, Fei Xia, and Bo Zhang. 2014. Discriminative relational topic models. *IEEE transactions on pattern analysis and machine intelligence*, 37(5):973–986.

Yu Chen and Mohammed J. Zaki. 2017. Kate: K-competitive autoencoder for text. *SIGKDD*.

[2]https://www.mindspore.cn/

Maosheng Guo, Yu Zhang, and Ting Liu. 2019. Gaussian transformer: a lightweight approach for natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6489–6496.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. *ICLR*.

Thomas Kipf and Max Welling. 2016a. Variational graph auto-encoders. *NIPS workshop*, abs/1611.07308.

Thomas N Kipf and Max Welling. 2016b. Semi-supervised classification with graph convolutional networks. *ICLR*.

Tuan M. V. Le and Hady Wirawan Lauw. 2014. Probabilistic latent document network embedding. *2014 ICDM*, pages 270–279.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. *arXiv preprint arXiv:1801.07606*.

S. Liu, Lingwei Chen, Hanze Dong, Zihao Wang, D. Wu, and Zengfeng Huang. 2019. Higher-order weighted graph convolutional networks. *ArXiv*, abs/1911.04129.

Alireza Makhzani and Brendan J. Frey. 2014. k-sparse autoencoders. *CoRR*, abs/1312.5663.

Ramesh M Nallapati, Amr Ahmed, Eric P Xing, and William W Cohen. 2008. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. 2011. Contractive auto-encoders: Explicit invariance during feature extraction. In *ICML*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *ICLR*.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408.

Chaojie Wang, Hao Zhang, Bo Chen, Dongsheng Wang, Zhengjue Wang, and Mingyuan Zhou. 2020a. Deep relational topic modeling via graph poisson gamma belief network. *Advances in Neural Information Processing Systems*, 33.

Hao Wang, Xingjian Shi, and Dit-Yan Yeung. 2017. Relational deep learning: A deep latent variable model for link prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Zhengjue Wang, Chaojie Wang, Hao Zhang, Zhibin Duan, Mingyuan Zhou, and Bo Chen. 2020b. Learning dynamic hierarchical topic graph with graph convolutional network for document classification. In *International Conference on Artificial Intelligence and Statistics*, pages 3959–3969. PMLR.

Qianqian Xie, Jimin Huang, Pan Du, Min Peng, and Jian-Yun Nie. 2021a. Graph topic neural network for document representation. In *Proceedings of The Web Conference 2021*.

Qianqian Xie, Jimin Huang, Pan Du, Min Peng, and Jian-Yun Nie. 2021b. Inductive topic variational graph auto-encoder for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4218–4227, Online. Association for Computational Linguistics.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

Liang Yang, Fan Wu, Junhua Gu, Chuan Wang, Xiaochun Cao, Di Jin, and Yuanfang Guo. 2020. Graph attention topic modeling network. In *Proceedings of The Web Conference 2020*, pages 144–154.

Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. 2016. A discriminative topic model using document network structure. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 686–696, Berlin, Germany. Association for Computational Linguistics.

Ce Zhang and Hady W. Lauw. 2020. Topic modeling on document networks with adjacent-encoder. In *AAAI*.

Hengrui Zhang and Julian McAuley. 2020. Stacked mixed-order graph convolutional networks for collaborative filtering. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 73–81. SIAM.

Luoyi Zhang and Ming Xu. 2020. Epine: Enhanced proximity information network embedding. *arXiv preprint arXiv:2003.02689*.

Qile Zhu, Zheng Feng, and Xiaolin Li. 2018. Graphbtm: Graph enhanced autoencoded variational inference for biterm topic model. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4663–4672.