Learning Graph Embedding with Adversarial Training Methods

Shirui Pan, Ruiqi Hu, Sai-fu Fung, Guodong Long, Jing Jiang, and Chengqi Zhang, *Senior Member, IEEE*

Abstract—Graph embedding aims to transfer a graph into vectors to facilitate subsequent graph analytics tasks like link prediction and graph clustering. Most approaches on graph embedding focus on preserving the graph structure or minimizing the reconstruction errors for graph data. They have mostly overlooked the embedding distribution of the latent codes, which unfortunately may lead to inferior representation in many cases. In this paper, we present a novel adversarially regularized framework for graph embedding. By employing the graph convolutional network as an encoder, our framework embeds the topological information and node content into a vector representation, from which a graph decoder is further built to reconstruct the input graph. The adversarial training principle is applied to enforce our latent codes to match a prior Gaussian or Uniform distribution. Based on this framework, we derive two variants of adversarial models, the adversarially regularized graph autoencoder (ARGA) and its variational version, adversarially regularized variational graph autoencoder (ARVGA), to learn the graph embedding effectively. We also exploit other potential variations of ARGA and ARVGA to get a deeper understanding on our designs. Experimental results compared among twelve algorithms for link prediction and twenty algorithms for graph clustering validate our solutions.

Index Terms—Graph Embedding, Graph Clustering, Link Prediction, Graph Convolutional Networks, Adversarial Regularization, Graph Autoencoder.

I. INTRODUCTION

GRAPHS are essential tools to capture and model complicated relationships among data. In a variety of graph applications, such as social networks, citation networks, protein-protein interaction networks, graph data analysis plays an important role in various data mining tasks including classification [1], clustering [2], recommendation [3], [4], [5], and graph classification [6], [7]. However, the high computational complexity, low parallelizability, and inapplicability of machine learning methods to graph data have made these graph analytic tasks profoundly challenging [8], [9]. *Graph embedding* has recently emerged as a general approach to these problems.

Graph embedding transfers graph data into a low dimensional, compact, and continuous feature space. The fundamental idea is to preserve the topological structure, vertex content,

S. Pan is with Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia (Email: shirui.pan@monash.edu).

R. Hu, G. Long, J. Jiang, C. Zhang are with Centre for Artificial Intelligence, FEIT, University of Technology Sydney, NSW 2007, Australia (E-mail: ruiqi.hu@uts.edu.au; jing.jiang@uts.edu.au; guodong.long@uts.edu.au; chengqi.zhang@uts.edu.au).

S.F. Fung is with Department of Applied Social Sciences, City University of Hong Kong, China.(E-mail: sffung@cityu.edu.hk).

Corresponding author: Ruiqi Hu.

Manuscript received April 19, 201x; revised August 26, 201x.

and other side information [10], [11]. This new learning paradigm has shifted the tasks of seeking complex models for classification, clustering, and link prediction [12] to learning a compact and informative representation for the graph data, so that many graph mining tasks can be easily performed by employing simple traditional models (e.g., a linear SVM for the classification task). This merit has motivated many studies in this area [4], [13].

Graph embedding algorithms can be classified into three categories: probabilistic models, matrix factorization-based algorithms, and deep learning-based algorithms. Probabilistic models like DeepWalk [14], node2vec [15] and LINE [16] attempt to learn graph embedding by extracting different patterns from the graph. The captured patterns or walks include global structural equivalence, local neighborhood connectivities, and other various order proximities. Compared with classical methods such as Spectral Clustering [17], these graph embedding algorithms perform more effectively and are scalable to large graphs.

Matrix factorization-based algorithms, such as GraRep [18], HOPE [19], M-NMF [20] pre-process the graph structure into an adjacency matrix and obtain the embedding by factorizing the adjacency matrix. It has been recently shown that many probabilistic algorithms including DeepWalk [14], LINE [16], node2vec [15], are equivalent to matrix factorization approaches [21], and Qiu et al. propose a unified matrix factorization approach NetMF [21] for graph embedding. Deep learning approaches, especially autoencoder-based methods, are also studied for graph embedding (a most up-to-date survey on graph neural networks can be found here [22]). SDNE [23] and DNGR [24] employ deep autoencoders to preserve the graph proximities and model the positive pointwise mutual information (PPMI). The MGAE algorithm utilizes a marginalized single layer autoencoder to learn representation for graph clustering [2]. The DNE-SBP model is proposed for signed network embedding with a stacked auto-encoder framework [25].

The approaches above are typically unregularized approaches which mainly focus on preserving the structure relationship (probabilistic approaches) or minimizing the reconstruction error (matrix factorization or deep learning methods). They have mostly ignored the latent data distribution of the representation. In practice, unregularized embedding approaches often learn a degenerate *identity* mapping where the latent code space is free of any structure [26], and can easily result in poor representation in dealing with real-world sparse and noisy graph data. One standard way to handle this problem is to introduce some regularization to the latent

codes and enforce them to follow some prior data distribution [26]. Recently generative adversarial based frameworks [27], [28], [29], [30] have also been developed for learning robust latent representation. However, none of these frameworks is specifically for graph data, where both topological structure and content information are required to be represented into a latent space.

In this paper, we propose a novel adversarially regularized algorithm with two variants, *adversarially regularized graph autoencoder* (ARGA) and its variational version, *adversarially regularized variational graph autoencoder* (ARVGA), for graph embedding. The theme of our framework is to not only minimize the reconstruction errors of the topological structure but also to enforce the learned latent embedding to match a prior distribution. By exploiting both graph structure and node content with a graph convolutional network, our algorithms encode the graph data in the latent space. With a decoder aiming at reconstructing the topological graph information, we further incorporate an adversarial training scheme to regularize the latent codes to learn a robust graph representation. The adversarial training module aims to discriminate if the latent codes are from a real prior distribution or the graph encoder. The graph encoder learning and adversarial regularization learning are jointly optimized in a unified framework so that each can be beneficial to the other and finally lead to a better graph embedding. To get further insight into the influence of prior distribution, we have varied it with the Gaussian distribution and Uniform distribution for all models and tasks. Moreover, we have examined the different ways to construct the graph decoders as well as the target of the reconstructions. By doing so, we have obtained a comprehensive view of the most influential factor of the adversarially regularized graph autoencoder models for different tasks. The experimental results on three benchmark graph datasets demonstrate the superb performance of our algorithms on two unsupervised graph analytic tasks, namely link prediction and node clustering. Our contributions can be summarized below:

- We propose a novel adversarially regularized framework for graph embedding, which represents topological structure and node content in a continuous vector space. Our framework learns the embedding to minimize the reconstruction error while enforcing the latent codes to match a prior distribution.
- We develop two variants of adversarial approaches, *adversarially regularized graph autoencoder* (ARGA) and *adversarially regularized variational graph autoencoder* (ARVGA) to learn the graph embedding.
- We have examined different prior distributions, the ways to construct decoders, and the targets of the reconstructions to point out the influence of the factors of the adversarially regularized graph autoencoder models on various tasks.
- Experiments on benchmark graph datasets demonstrate that our graph embedding approaches outperform the others on different unsupervised tasks.

The paper is structured as follows. Section II reviews the related work. Section III outlines the problem definition and our overall framework. Section IV presents the proposed algorithm and Section V describes the experimental results. We conclude the paper in Section VI.

II. RELATED WORK

A. Graph Embedding Models

Graph embedding, also known as network embedding [4] or network representation learning [10], transfers a graph into vectors. From the perspective of information exploration, graph embedding algorithms can be separated into two groups: topological network embedding approaches and content enhanced network embedding methods.

Topological network embedding approaches Topological network embedding approaches assume that there is only topological structure information available, and the learning objective is to preserve the topological information maximumly [31], [32]. Inspired by the word embedding approach [33], Perozzi et al. propose a DeepWalk model to learn the node embedding from a collection of random walks [14]. Since then, many probabilistic models have been developed. Specifically, Grover et al. propose a biased random walks approach, node2vec [15], which employs both breadth-first sampling (BFS) and Depth-first sampling (DFS) strategies to generate random walk sequences for network embedding. Tang et al. propose a LINE algorithm [16] to handle large-scale information networks while preserving both first-order and second-order proximity. Other random walk variants include hierarchical representation learning approach (HARP) [34], and discriminative deep random walk (DDRW) [35], and Walklets [36].

Because a graph can be mathematically represented as an adjacency matrix, many matrix factorization approaches are proposed to learn the latent representation for a graph. GraRep [18] integrates the global topological information of the graph into the learning process to represent each node into a low dimensional space; HOPE [19] preserves the asymmetric transitivity by approximating high-order proximity for a better performance on capturing topological information of graphs and reconstructing from partially observed graphs; DNE [37] aims to learn discrete embedding which reduces the storage and computational cost. Recently deep learning models have been exploited to learn the graph embedding. These algorithms preserve the first and second order of proximities [23], or reconstruct the positive pointwise mutual information (PPMI) [24] via different variants of autoencoders.

Content enhanced network embedding methods Content enhanced embedding methods assume node content information is available and exploit both topological information and content features simultaneously. TADW [38] proved that DeepWalk can be interpreted as a factorization approach and proposed an extension to DeepWalk to explore node features. TriDNR [39] captures structure, node content, and label information via a tri-party neural network architecture. UPP-SNE [40] employs an approximated kernel mapping scheme to exploit user profile features to enhance the embedding learning of users in social networks. SNE [41] learns a neural network model to capture both structural proximity

and attribute proximity for attributed social networks. DANE [42] deals with the dynamic environment with an incremental matrix factorization approach, and LANE [43] incorporates label information into the optimization process to learn a better embedding. Recently, BANE [44] is proposed to learn binarized embedding for an attributed graph which has the potential to increase the efficiency for latter graph analytic tasks.

Although these algorithms are well-designed for graph-structured data, they have largely ignored the embedding distribution, which may result in poor representation in real-graph data. In this paper, we explore adversarial training approaches to address this issue.

B. Adversarial Models

Our method is motivated by the generative adversarial network (GAN) [45]. GAN plays an adversarial game with two linked models: the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$. The discriminator discriminates if an input sample comes from the prior data distribution or from the generator we built. Simultaneously, the generator is trained to generate the samples to convince the discriminator that the generated samples come from the prior data distribution. Typically, the training process is split into two steps: (1) Train the discriminator $\mathcal{D}$ for iterations to distinguish the samples from the expected data distribution from the samples generated via the generator. Then (2) train the generator to confuse the discriminator with its generated data. However, the original GAN does not fit the unsupervised data encoding, as the absence of the precise structure for inference. To implement the adversarial structure in learning data embedding, existing works like BiGAN[27], EBGAN[28] and ALI[29] arrive at extending the original adversarial framework with external structures for the inference, which have achieved non-negligible performance in applications, such as document retrieval[46] and image classification[27]. Other solutions manage to generate the embedding from the discriminator or generator for semi-supervised and supervised tasks via reconstructed layers. For example, DCGAN[30] bridges the gap between convolutional networks and generative adversarial networks with particular architectural constraints for unsupervised learning; and ANE[47] combines a structure-preserving component and an adversarial learning scheme to learn a robust embedding.

Makhzani et al. proposed an adversarial autoencoder (AAE) to learn the latent embedding by merging the adversarial mechanism into the autoencoder [26]. However, AAE is designed for general data rather than graph data. Recently there are some studies on applying the adversarial mechanism to graphs such as AIDW [47] and NetRA [48]. However, their approach can only exploit the topological information [47], [49], [48]. In contrast, our algorithm is more flexible and can handle both topological and content information for graph data. Furthermore, these models, such as NetRA, can only reconstruct the graph structure, while ARGA_AX reconstructs both topological structure and node characteristics, smoothly perservering the integrity of the given graph through entire encoding and decoding processing. Most recently, Ding et

al. proposed a GraphSGAN [50] for semi-supervised node classification with the GAN principle, and Hu et al. proposed the HeGAN [51] for heterogeneous information network embedding.

Though many adversarial models have achieved impressive success in computer vision, they cannot effectively and directly handle the graph-structured data. With some preliminary study in [52], we try to thoroughly exploit the graph convolutional models with different adversarial models to learn a robust graph embedding in this paper.

In particular, we have proposed four new algorithms to handle networks with limited labeled data. These algorithms aim to reconstruct different content in a network, including topological structure only or both the topological structure and node content, by using general graph encoder or variational graph encoder as a building block. We also conducted more extensive experiments to validate the proposed algorithms with a wide range of metrics including NMI, ACC, F1, Precision, ARI and Recall.

C. Graph Convolutional Nets based Models

Graph convolutional networks (GCN) [1] is a semi-supervised framework based on a variant of convolutional neural networks, which attempt to operate the graphs directly. Specifically, the GCN represents the graph structure and the interrelationship between node and feature with an adjacent matrix $\mathbf{A}$ and node-feature matrix $\mathbf{X}$. Hence, GCN can directly embed the graph structure with a spectral convolutional function $f(\mathbf{X}, \mathbf{A})$ for each layer and train the model on a supervised target for all labelled nodes. Because of the spectral function $f(\bullet)$ on the adjacent matrix $\mathbf{A}$ of the graph, the model can distribute the gradient from the supervised cost and learn the embedding of both the labelled and unlabelled nodes. Although GCN is powerful on graph-structured data sets for semi-supervised tasks like node classification, variational graph autoencoder VGAE [53] extends it into unsupervised scenarios. Specifically, VGAE integrates the GCN into the variational autoencoder framework [54] by framing the encoder with graph convolutional layers and remodeling the decoder with a link prediction layer. Taking advantage of GCN layers, VGAE can naturally leverage the information of node features, which expressively muscle the predictive performance. Recently GCN is used to learn the binary codes for improving the efficiency of information retrieval [55].

III. PROBLEM DEFINITION AND FRAMEWORK

A graph is represented as $\mathbf{G} = \{\mathbf{V}, \mathbf{E}, \mathbf{X}\}$, where $\mathbf{V} = \{\mathbf{v}_i\}_i = 1, \cdots, n$ is constitutive of a set of nodes in a graph and $\mathbf{e}_{i,j} = <\mathbf{v}_i, \mathbf{v}_j> \in \mathbf{E}$ represents a linkage coding the citation edge between the papers (nodes). The topological structure of graph $\mathbf{G}$ can be represented by an adjacency matrix $\mathbf{A}$, where $\mathbf{A}_{i,j} = 1$ if $\mathbf{e}_{i,j} \in \mathbf{E}$, otherwise $\mathbf{A}_{i,j} = 0$. $\mathbf{x}_i \in \mathbf{X}$ encodes the textual content features associated with each node $\mathbf{v}_i$.

Given a graph $\mathbf{G}$, our purpose is to map the nodes $\mathbf{v}_i \in \mathbf{V}$ to low-dimensional vectors $\mathbf{z}_i \in \mathbb{R}^d$ with the formal format as follows: $f : (\mathbf{A}, \mathbf{X}) \rightarrowtail \mathbf{Z}$, where $\mathbf{z}_i^\top$ is the i-th row of
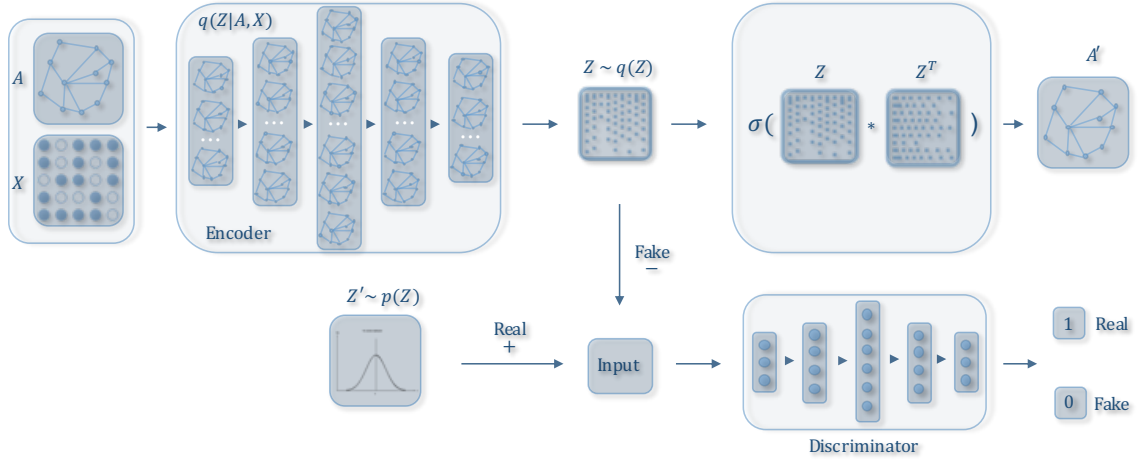
Fig. 1: The architecture of the adversarially regularized graph autoencoder (**ARGA**). The upper tier is a graph convolutional autoencoder that reconstructs a graph $\mathbf{A}$ from an embedding $\mathbf{Z}$ which is generated by the encoder which exploits graph structure $\mathbf{A}$ and the node content matrix $\mathbf{X}$. The lower tier is an adversarial network trained to discriminate if a sample is generated from the embedding or from a prior distribution. The adversarially regularized variational graph autoencoder (ARVGA) is similar to ARGA except that it employs a *variational* graph autoencoder in the upper tier (See Algorithm 1 for details).

the matrix $\mathbf{Z} \in \mathbb{R}^{n \times d}$. n is the number of nodes and d is the dimension of embedding. We take $\mathbf{Z}$ as the embedding matrix and the embeddings should well preserve the topological structure $\mathbf{A}$ as well as content information $\mathbf{X}$.

A. Overall Framework

The objective is to learn a robust embedding for a given graph $\mathbf{G} = \{\mathbf{V}, \mathbf{E}, \mathbf{X}\}$. To this end, we leverage an adversarial architecture with a graph autoencoder to directly process the entire graph and learn a robust embedding. Figure 1 demonstrates the workflow of ARGA which consists of two modules: the graph autoencoder and the adversarial network.

- **Graph convolutional autoencoder**. The autoencoder takes in the structure of graph $\mathbf{A}$ and the node content $\mathbf{X}$ as inputs to learn a latent representation $\mathbf{Z}$, and then reconstructs the graph structure $\mathbf{A}$ from $\mathbf{Z}$. We will further explore other variants of graph autoencoder in Section IV-D.
- **Adversarial regularization.** The adversarial network forces the latent codes to match a prior distribution by an adversarial training module, which discriminates whether the current latent code $\mathbf{z}_i \in \mathbf{Z}$ comes from the encoder or from the prior distribution.

IV. PROPOSED ALGORITHM

A. Graph Convolutional Autoencoder

Our graph convolutional autoencoder aims to embed a graph $\mathbf{G} = \{\mathbf{V}, \mathbf{E}, \mathbf{X}\}$ in a low-dimensional space. Two fundamental questions arise (1) how to simultaneously integrate graph structure $\mathbf{A}$ and content feature $\mathbf{X}$ in an encoder, and (2) what sort of information should be reconstructed via a decoder?

Graph Convolutional Encoder Model $\mathcal{G}(\mathbf{X}, \mathbf{A})$. To represent both graph structure $\mathbf{A}$ and node content $\mathbf{X}$ in a unified framework, we develop a variant of the graph

convolutional network (GCN) [1] as a graph encoder. GCN introduces the *convolutional* operation to graph-data from the spectral area, and leverages a spectral convolutional function $f(\mathbf{Z}^{(l)}, \mathbf{A}|\mathbf{W}^{(l)})$ to build a layer-wise transformation:

$$\mathbf{Z}^{(l+1)} = f(\mathbf{Z}^{(l)}, \mathbf{A}|\mathbf{W}^{(l)}) \tag{1}$$

Here, $\mathbf{Z}^l$ and $\mathbf{Z}^{(l+1)}$ are the input and output of the convolution respectively. We set $\mathbf{Z}^0 = \mathbf{X} \in \mathbb{R}^{n \times m}$ (n indicates the number of nodes and m indicates the number of features) for our problem. We need to learn a filter parameter matrix $\mathbf{W}^{(l)}$ in the neural network, and if the spectral convolution function is well defined, we can efficiently construct arbitrary deep convolutional neural networks.

Each layer of our graph convolutional network can be expressed with the the spectral convolution function $f(\mathbf{Z}^{(l)}, \mathbf{A}|\mathbf{W}^{(l)})$ as follows:

$$f(\mathbf{Z}^{(l)}, \mathbf{A}|\mathbf{W}^{(l)}) = \phi(\widetilde{\mathbf{D}}^{-\frac{1}{2}} \widetilde{\mathbf{A}} \widetilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{Z}^{(l)} \mathbf{W}^{(l)}), \tag{2}$$

where $\widetilde{\mathbf{D}}_{ii} = \sum_j \widetilde{\mathbf{A}}_{ij}$ and $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$. $\mathbf{I}$ is the identity matrix of $\mathbf{A}$ and ϕ is an activation function such as sigmoid$(t) = \frac{1}{1+e^t}$ or Relu$(t) = \max(0, t)$. Overall, the graph encoder $\mathcal{G}(\mathbf{X}, \mathbf{A})$ is constructed with a two-layer GCN. In our paper, we develop two variants of the encoder, e.g., Graph Encoder and Variational Graph Encoder.

The *Graph Encoder* is constructed as follows:

$$\mathbf{Z}^{(1)} = f_{\text{Relu}}(\mathbf{X}, \mathbf{A}|\mathbf{W}^{(0)}); \tag{3}$$
$$\mathbf{Z}^{(2)} = f_{\text{linear}}(\mathbf{Z}^{(1)}, \mathbf{A}|\mathbf{W}^{(1)}). \tag{4}$$

Relu$(\cdot)$ and linear activation functions are used for the first and second layers. Our graph convolutional encoder $\mathcal{G}(\mathbf{Z}, \mathbf{A}) = q(\mathbf{Z}|\mathbf{X}, \mathbf{A})$ encodes both graph structure and node content into a representation $\mathbf{Z} = q(\mathbf{Z}|\mathbf{X}, \mathbf{A}) = \mathbf{Z}^{(2)}$.

A *Variational Graph Encoder* is defined by an inference model:

$$q(\mathbf{Z}|\mathbf{X}, \mathbf{A}) = \prod_{i=1}^{n} q(\mathbf{z_i}|\mathbf{X}, \mathbf{A}), \tag{5}$$

$$q(\mathbf{z_i}|\mathbf{X}, \mathbf{A}) = \mathcal{N}(\mathbf{z_i}|\boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}^2)) \tag{6}$$

Here, $\boldsymbol{\mu} = \mathbf{Z}^{(2)}$ is the matrix of mean vectors $\mathbf{z}_i$; similarly $\log\boldsymbol{\sigma} = f_{\text{linear}}(\mathbf{Z}^{(1)}, \mathbf{A}|\mathbf{W}'^{(1)})$ which shares the weights $\mathbf{W}^{(0)}$ with $\boldsymbol{\mu}$ in the first layer in Eq. (3).

Decoder Model. Our decoder model is used to reconstruct the graph data. We can reconstruct either the graph structure $\mathbf{A}$, content information $\mathbf{X}$, or both. In the basic version of our model (ARGA), we propose to reconstruct graph structure $\mathbf{A}$, which provides more flexibility in the sense that our algorithm will still function properly even if there is no content information $\mathbf{X}$ available (e.g., $\mathbf{X} = \mathbf{I}$). We will provide several variants of decoder model in Section IV-D. Here the ARGA decoder $p(\hat{\mathbf{A}}|\mathbf{Z})$ predicts whether there is a link between two nodes. More specifically, we train a link prediction layer based on the graph embedding:

$$p(\hat{\mathbf{A}}|\mathbf{Z}) = \prod_{i=1}^{n} \prod_{j=1}^{n} p(\hat{\mathbf{A}}_{ij}|\mathbf{z}_i, \mathbf{z}_j); \tag{7}$$

$$p(\hat{\mathbf{A}}_{ij} = 1|\mathbf{z}_i, \mathbf{z}_j) = \text{sigmoid}(\mathbf{z}_i^\top, \mathbf{z}_j), \tag{8}$$

here the prediction $\hat{\mathbf{A}}$ should be close to the ground truth $\mathbf{A}$.

Graph Autoencoder Model. The embedding $\mathbf{Z}$ and the reconstructed graph $\hat{\mathbf{A}}$ can be presented as follows:

$$\hat{\mathbf{A}} = \text{sigmoid}(\mathbf{Z}\mathbf{Z}^\top), \text{ here } \mathbf{Z} = q(\mathbf{Z}|\mathbf{X}, \mathbf{A}) \tag{9}$$

Optimization. For the graph encoder, we minimize the reconstruction error of the graph data by:

$$\mathcal{L}_0 = \mathbb{E}_{q(\mathbf{Z}|(\mathbf{X}, \mathbf{A}))}[\log p(\mathbf{A}|\mathbf{Z})] \tag{10}$$

For the variational graph encoder, we optimize the variational lower bound as follows:

$$\mathcal{L}_1 = \mathbb{E}_{q(\mathbf{Z}|(\mathbf{X}, \mathbf{A}))}[\log p(\mathbf{A}|\mathbf{Z})] - \mathbf{KL}[q(\mathbf{Z}|\mathbf{X}, \mathbf{A}) \parallel p(\mathbf{Z})] \tag{11}$$

where $\mathbf{KL}[q(\bullet)||p(\bullet)]$ is the Kullback-Leibler divergence between $q(\bullet)$ and $p(\bullet)$. $p(\bullet)$ is a prior distribution which can be a uniform distribution or a Gaussian distribution $p(\mathbf{Z}) = \prod_i p(\mathbf{z}_i) = \prod_i \mathcal{N}(\mathbf{z}_i|0, \mathbf{I})$ in practice.

B. Adversarial Model $\mathcal{D}(\mathbf{Z})$

The fundamental idea of our model is to enforce latent representation $\mathbf{Z}$ to match a prior distribution, which is achieved by an adversarial training model. The adversarial model is built on a standard multi-layer perceptron (MLP) where the output layer only has one dimension with a sigmoid function. The adversarial model acts as a discriminator to distinguish whether a latent code is from the prior p_z (positive) or graph encoder $\mathcal{G}(\mathbf{X}, \mathbf{A})$ (negative). By minimizing the cross-entropy cost for training the binary classifier, the embedding will finally be regularized and improved during the training process. The cost can be computed as follows:

$$-\frac{1}{2}\mathbb{E}_{\mathbf{z} \sim p_z}\log\mathcal{D}(\mathbf{Z}) - \frac{1}{2}\mathbb{E}_{\mathbf{X}}\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{X}, \mathbf{A}))), \tag{12}$$

Algorithm 1 Adversarially Regularized Graph Embedding

Require:
  $\mathbf{G} = \{\mathbf{V}, \mathbf{E}, \mathbf{X}\}$: a Graph with links and features;
  T: the number of iterations;
  K: the number of steps for iterating discriminator;
  d: the dimension of the latent variable
Ensure: $\mathbf{Z} \in \mathbb{R}^{n \times d}$
1: **for** iterator = 1,2,3, $\cdots\cdots$, T **do**
2:     Generate latent variables matrix $\mathbf{Z}$ through Eq.(4);
3:     **for** k = 1,2, $\cdots\cdots$, K **do**
4:         Sample m entities $\{\mathbf{z}^{(1)}, ..., \mathbf{z}^{(m)}\}$ from latent matrix $\mathbf{Z}$
5:         Sample m entities $\{\mathbf{a}^{(1)}, ..., \mathbf{a}^{(m)}\}$ from the prior distribution p_z
6:         Update the discriminator with its stochastic gradient:

$$\bigtriangledown \frac{1}{m} \sum_{i=1}^{m} [\log \mathcal{D}(\mathbf{a}^i) + \log (1 - \mathcal{D}(\mathbf{z}^{(i)}))]$$

7:     **end for**
8:     Update the graph autoencoder with its stochastic gradient by Eq. (10) for ARGA or Eq. (11) for ARVGA;
9: **end for**
10: **return** $\mathbf{Z} \in \mathbb{R}^{n \times d}$

In our paper, we have examined both Gaussian distribution and Uniform distribution as p_z for all models and tasks.

Adversarial Graph Autoencoder Model. The equation for training the encoder model with Discriminator $\mathcal{D}(\mathbf{Z})$ can be written as follows:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}_{\mathbf{z} \sim p_z}[\log\mathcal{D}(\mathbf{Z})] + \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})}[\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{X}, \mathbf{A})))] \tag{13}$$

where $\mathcal{G}(\mathbf{X}, \mathbf{A})$ and $\mathcal{D}(\mathbf{Z})$ indicate the generator and discriminator explained above.

C. Algorithm Explanation

Algorithm 1 is our proposed framework. Given a graph $\mathbf{G}$, step 2 gets the latent variables matrix $\mathbf{Z}$ from the graph convolutional encoder. Then we take the same number of samples from the generated $\mathbf{Z}$ and the real data distribution p_z in step 4 and 5 respectively, to update the discriminator with the cross-entropy cost computed in step 6. After K runs of training the discriminator, the graph encoder will try to confuse the trained discriminator and update itself with the generated gradient in step 8. We can update Eq. (10) to train the **adversarially regularized graph autoencoder (ARGA),** or Eq. (11) to train the **adversarially regularized variational graph autoencoder (ARVGA)**, respectively. Finally, we will return the graph embedding $\mathbf{Z} \in \mathbb{R}^{n \times d}$ in step 9.

D. Decoder Variations

In ARGA and ARVGA models, the decoder is merely a link prediction layer which performs as a dot product of the embedding $\mathbf{Z}$. In practice, the decoder can also be a graph convolutional layer or a combination of link prediction layer and graph convolutional decoder layer.

GCN Decoder for Graph Structure Reconstruction (ARGA_GD) We have modified the encoder by adding two graph convolutional layers to reconstruct the graph structure.
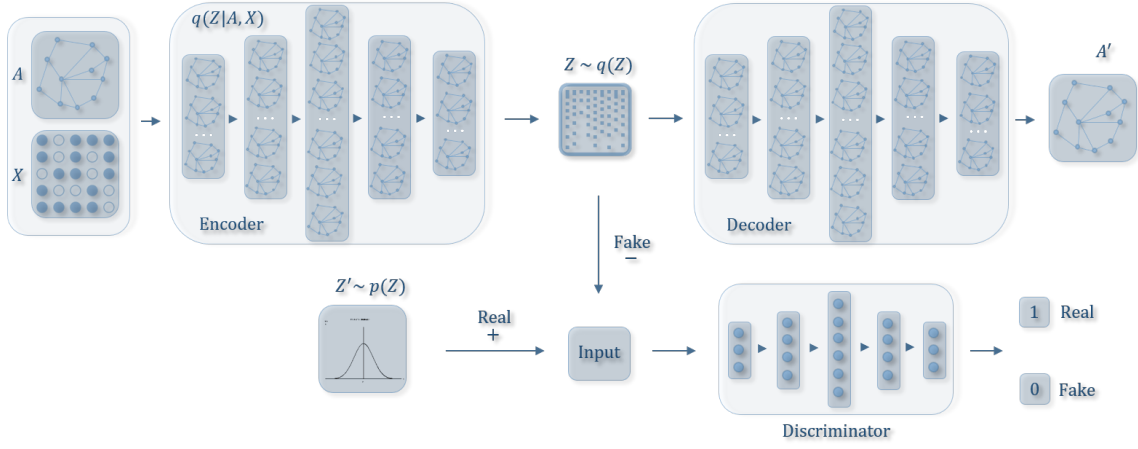
Fig. 2: The architecture of adversarially regularized graph autoencoder with a graph convolutional decoder (**ARGA_GD**) to reconstruct the topological structure **A**.The upper tier is a standard graph convolutional autoencoder. The decoder employs the graph convolutional networks. The lower tier keeps the same with both Gaussian distribution and Uniform distribution. **ARVGA_GD** is similar to **ARGA_GD** except that it employs a *variational* graph autoencoder in the upper tier.
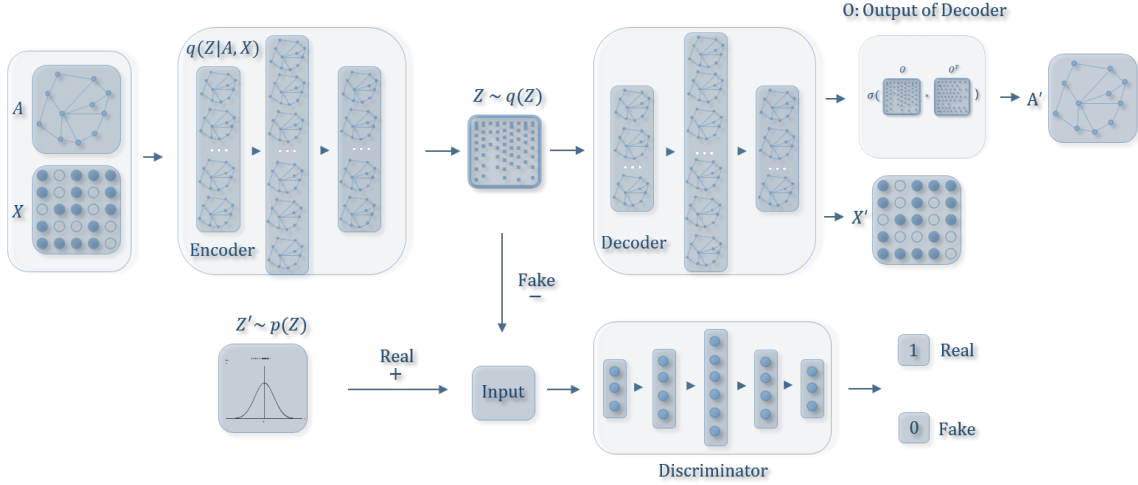

Fig. 3: The architecture of the **ARGA_AX** which simultaneously reconstructs the graph topological structure **A** and the node content matrix **X**. The lower tier keeps the same, and we also exploit the variational version of the **ARVGA_AX**.

This variant of approach is named ARGA_GD. Fig. 2 demonstrates the architecture of ARGA_GD. In this approach, the input of the decoder will be the embedding from the encoder, and the graph convolutional decoder is constructed as follows:

$$\mathbf{Z}_D = f_{\text{linear}}(\mathbf{Z}, \mathbf{A}|\mathbf{W}_D^{(1)}). \tag{14}$$

$$\mathbf{O} = f_{\text{linear}}(\mathbf{Z}_D, \mathbf{A}|\mathbf{W}_D^{(2)}). \tag{15}$$

where $\mathbf{Z}$ is the embedding learned from the graph encoder while $\mathbf{Z}_D$ and $\mathbf{O}$ are the outputs from the first and second layer of the graph decoder. The number of the horizontal dimension of $\mathbf{O}$ is equal to the number of nodes. Then we calculate the reconstruction error as follows:

$$\mathcal{L}_{ARGA\_GD} = \mathbb{E}_{q(\mathbf{O}|(\mathbf{X},\mathbf{A}))}[\log p(\mathbf{A}|\mathbf{O})] \tag{16}$$

GCN Decoder for both Graph Structure and Content Information Reconstruction (ARGA_AX) We have further modified our graph convolutional decoder to reconstruct both the graph structure **A** and content information **X**. The architecture is illustrated in Fig 3. We fixed the dimension of second graph convolutional layer with the same number of the features associated with every node, thus the output from the second layer $\mathbf{O} \in \mathbb{R}^{n \times f} \ni \mathbf{X}$. In this case, the reconstruction loss is composed of two errors. First, the reconstruction error of graph structure can be minimized as follows:

$$\mathcal{L}_A = \mathbb{E}_{q(\mathbf{O}|(\mathbf{X},\mathbf{A}))}[\log p(\mathbf{A}|\mathbf{O})], \tag{17}$$

Then the reconstruction error of node content can be minimized with a similar formula:

$$\mathcal{L}_X = \mathbb{E}_{q(\mathbf{O}|(\mathbf{X},\mathbf{A}))}[\log p(\mathbf{X}|\mathbf{O})]. \tag{18}$$

The final reconstruction error is the sum of the reconstruction error of graph structure and node content:

$$\mathcal{L}_0 = \mathcal{L}_A + \mathcal{L}_X. \tag{19}$$

V. Experiments

We report our results on both link prediction and node clustering tasks. The benchmark graph datasets used in the paper, Cora [56], Citeseer [57] and Pubmed [58], are summarized in table 1. Each dataset consists of scientific publications as nodes and citation relationships as edges. The features are unique words in each document.

TABLE I: Real-world Graph Datasets Used in the Paper

Data Set	# Nodes	# Links	# Content Words	# Features
Cora	2,708	5,429	3,880,564	1,433
Citeseer	3,327	4,732	12,274,336	3,703
PubMed	19,717	44,338	9,858,500	500

A. Link Prediction

Baselines.　**Twelve** algorithms in total are compared for the link prediction task:

- **DeepWalk** [14] is a network representation approach which encodes social relations into a continuous vector space.
- **Spectral Clustering** [17] is an effective approach to learn social embedding.
- **GAE** [53] is the most recent autoencoder-based unsupervised framework for graph data, which naturally leverages both topological structure $\mathbf{A}$ and content information $\mathbf{X}$. **GAE*** is the version of GAE which only considers the topological information $\mathbf{A}$, i.e., $\mathbf{X} = \mathbf{I}$.
- **VGAE** [53] is the variational graph autoencoder for graph embedding with both topological and content information. Likewise, **VGAE*** is a simplified version of VGAE which only leverages the topological information.
- **ARGA** is our proposed adversarially regularized autoencoder algorithm which uses graph autoencoder to learn the embedding.
- **ARVGA** is our proposed algorithm, which uses a *variational* graph autoencoder to learn the embedding.
- **ARGA_DG** is a variant of our proposed ARGA which takes graph convolutional layers as its decoder to reconstruct graph structure. **ARVGA_DG** is the variational version of **ARGA_DG**.
- **ARGA_AX** is a variant of our proposed ARGA which takes graph convolutional layers as its decoder to simultaneously reconstruct graph structure and node content. **ARVGA_AX** is the variational version of **ARGA_AX**.

Metric.　We report the results concerning AUC score (the area under a receiver operating characteristic curve) and average precision (AP) [53] score which can be computed as follow:

$$\mathrm{AUC} = \frac{\sum_i^1 \sum_j^1 \mathtt{pred}(x_i) > \mathtt{pred}(y_j)}{N * M}$$

where $\mathtt{pred}(\bullet)$ is the outputs from the predictor and N and M are the number of positive samples $x_i \in X$ and the number of negative samples $y_j \in Y$ respectively. We also report the Average Precision (AP) which indicates the area under the precision-recall curve:

$$\mathrm{Precision} = \frac{\mathtt{true\_positive}}{\mathtt{true\_positive} + \mathtt{false\_positive}}$$

$$\mathrm{AP} = \frac{\sum_k \mathtt{Precision}(k)}{\#\{\mathtt{positive\_sample}\}}$$

where k is an index for the class k.

We conduct each experiment 10 times and report the mean values with the standard errors as the final scores. Each dataset is separated into a training, testing set, and a validation set. The validation set contains 5% citation edges for hyperparameter optimization, the test set holds 10% citation edges to verify the performance, and the rest are used for training.

Parameter Settings.　For the Cora and Citeseer data sets, we train all autoencoder-related models for 200 iterations and optimize them with the Adam algorithm. Both the learning rate and discriminator learning rate are set as 0.001. As the PubMed dataset is relatively large (around 20,000 nodes), we iterate 2,000 times for adequate training with a 0.008 discriminator learning rate and 0.001 learning rate. We construct encoders with a 32-neuron hidden layer and a 16-neuron embedding layer for all the experiments and all the discriminators are built with two hidden layers(16-neuron, 64-neuron respectively). For the rest of the baselines, we retain the settings described in the corresponding papers.

Experimental Results.　The details of the experimental results on the link prediction are shown in Table 2. The results show that by incorporating an effective adversarial training module into our graph convolutional autoencoder, ARGA and ARVGA achieve outstanding performance: all AP and AUC scores are as higher as 92% on all three data sets. Compared with all the baselines, ARGA increased the AP score from around 2.5% compared with VGAE incorporating with node features, 11% compared with VGAE without node features; 15.5% and 10.6% compared with DeepWalk and Spectral Clustering respectively on the large PubMed data set.

The approaches which use both node content and topological information are always straightforward to get better performance compared to those only consider graph structure. The gap between ARGA and GAE models demonstrates that regularization on the latent codes has its advantage to learn a robust embedding. The impact of various distributions, architectures of the decoder as well as the reconstructions will be discussed in Section V-C: ARGA Architectures Comparison.
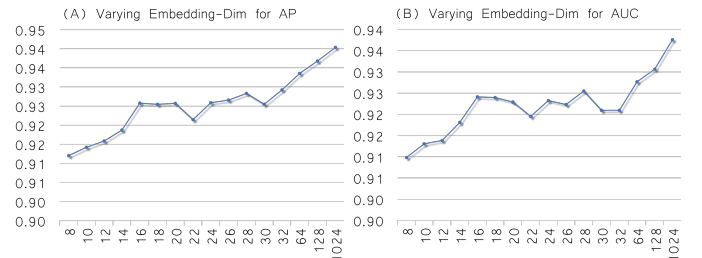

Fig. 4: Average performance on different dimensions of the embedding. (A) Average Precision score; (B) AUC score.

Parameter Study.　We conducted experiments on Cora dataset by varying the dimension of embedding from 8 neurons to 1024 and report the results in Fig 4.

The results from both Fig 4 (A) and (B) reveal similar trends: when adding the dimension of embedding from 8-

TABLE II: Results for Link Prediction. GAE* and VGAE* are variants of GAE and VGAE, which only explore topological structure, i.e., $\mathbf{X} = \mathbf{I}$.

Approaches	Cora		Citeseer		PubMed	
	AUC	AP	AUC	AP	AUC	AP
SC	84.6 ± 0.01	88.5 ± 0.00	80.5 ± 0.01	85.0 ± 0.01	84.2 ± 0.02	87.8 ± 0.01
DW	83.1 ± 0.01	85.0 ± 0.00	80.5 ± 0.02	83.6 ± 0.01	84.4 ± 0.00	84.1 ± 0.00
GAE*	84.3 ± 0.02	88.1 ± 0.01	78.7 ± 0.02	84.1 ± 0.02	82.2 ± 0.01	87.4 ± 0.00
VGAE*	84.0 ± 0.02	87.7 ± 0.01	78.9 ± 0.03	84.1 ± 0.02	82.7 ± 0.01	87.5 ± 0.01
GAE	91.0 ± 0.02	92.0 ± 0.03	89.5 ± 0.04	89.9 ± 0.05	96.4 ± 0.00	96.5 ± 0.00
VGAE	91.4 ± 0.01	92.6 ± 0.01	90.8 ± 0.02	92.0 ± 0.02	94.4 ± 0.02	94.7 ± 0.02
ARGA	92.4 ± 0.003	**93.2 ± 0.003**	91.9 ± 0.003	93.0± 0.003	**96.8 ± 0.001**	**97.1 ± 0.001**
ARVGA	**92.4 ± 0.004**	92.6 ± 0.004	**92.4 ± 0.003**	**93.0 ± 0.003**	96.5± 0.001	96.8± 0.001
ARGA_DG	77.9 ± 0.003	78.9 ± 0.003	74.4 ± 0.003	76.2± 0.003	95.1 ± 0.001	95.2 ± 0.001
ARVGA_DG	88.0 ± 0.004	87.9 ± 0.004	89.7 ± 0.003	90.5 ± 0.003	93.2± 0.001	93.6 ± 0.001
ARGA_AX	91.3 ± 0.003	91.3 ± 0.003	91.9 ± 0.003	93.4± 0.003	96.6 ± 0.001	96.7 ± 0.001
ARVGA_AX	90.2 ± 0.004	89.2 ± 0.004	89.8 ± 0.003	90.4 ± 0.003	96.7± 0.001	97.1 ± 0.001

TABLE III: Algorithm Comparison

	K-means	Spectral	BigClam	GraphEncoder	DeepWalk	DNGR	Circles	RTM	RMSC	TADW	GAE*	VGAE*	GAE	**ARGA**	*ARGA_DG*	*ARGA_AX*
Content	★						★	★	★	★			★	★	★	★
Structure		★	★	★	★	★	★	★	★	★	★	★	★	★	★	★
Adversarial														★	★	★
GCN encoder											★	★	★	★	★	★
GCN dncoder															★	★
Recover A				★	★	★					★	★	★	★	★	★
Recover X																★

neuron to 16-neuron, the performance of embedding on link prediction steadily rises; when we further increase the number of the neurons at the embedding layer to 32-neuron, the performance fluctuates, however, the results for both the AP score and the AUC score remain good.

It is worth mentioning that if we continue to set more neurons, for examples, 64-neuron, 128-neuron and 1024-neuron, the performance rises dramatically.

B. Node Clustering

For the node clustering task, we first learn the graph embedding, and after that, we perform the K-means clustering method based on the embedding.

Baselines We compare both embedding based approaches as well as approaches directly for graph clustering. Except for the baselines we compared for link prediction, we also include baselines which are designed for clustering. **Twenty** approaches in total are compared in the experiments. For a comprehensive validation, we take the algorithms which only consider one perspective of the information source, say, network structure or node content, as well as algorithms considering both factors.

Node Content or Graph Structure Only: The baselines which leverage only one source of information are listed as follows.

1) **K-means** is a classical method and also the foundation of many clustering algorithms.
2) **Big-Clam** [17] is a community detection algorithm based on NMF.

3) **Graph Encoder** [59] learns graph embedding for spectral graph clustering.
4) **DNGR** [24] trains a stacked denoising autoencoder for graph embedding.

Both Content and Structure The baselines which simultaneously consider both topological structure and node characteristics are listed as follows.

5) **Circles** [60] is an overlapping graph clustering algorithm which treats each node as ego and builds the ego graph with the linkages between the ego's friends.
6) **RTM** [61] learns the topic distributions of each document from both text and citation.
7) **RMSC** [62] is a multi-view clustering algorithm which recovers the shared low-rank transition probability matrix from each view for clustering. In this paper, we treat node content and topological structure as two different views.
8) **TADW** [38] applies matrix factorization for network representation learning.

Table III gives the detailed comparison of most of the baselines. For space saving, we did not list the variational versions of our models. Recovering $\mathbf{A}$ and $\mathbf{X}$ in the table demonstrates whether the model reconstructs the graph structure ($\mathbf{A}$) and node content ($\mathbf{X}$). Please note that we do not report the clustering results from Circle on PubMed dataset as the single experiment have been running more than three days without any outcome and error. We think this is because of the large size of the PubMed dataset (around 20,000 nodes). Note that the Circle algorithm works well on the other two datasets.
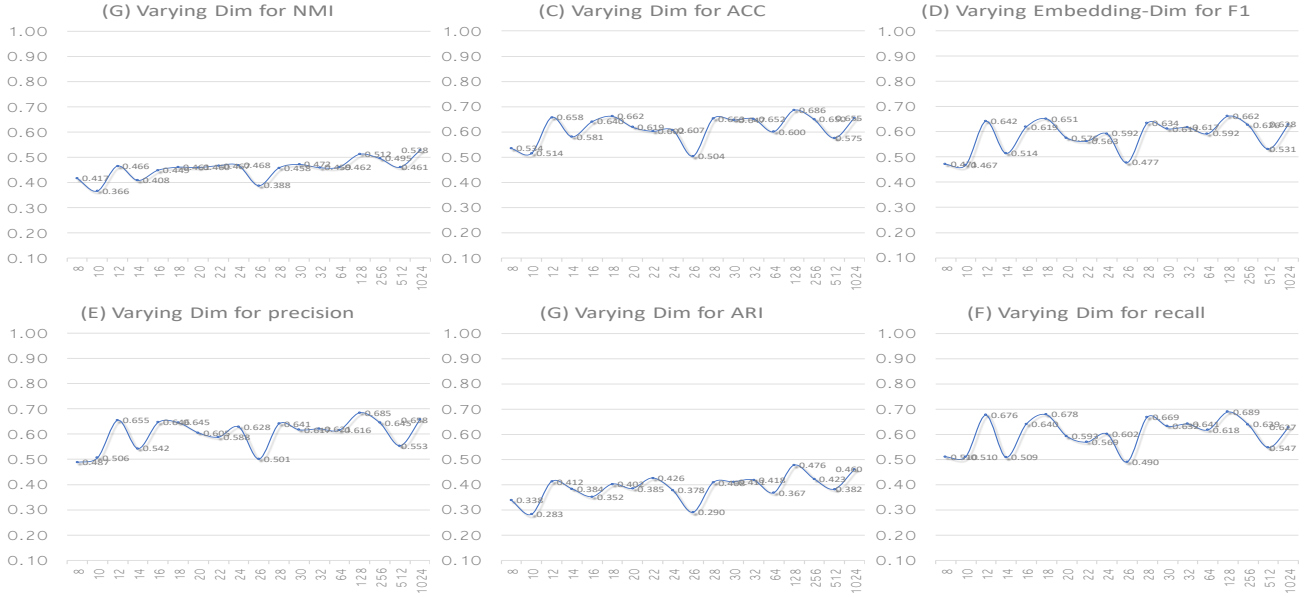
Fig. 5: Average node clustering performance on different dimensions of the embedding.

TABLE IV: Clustering Results on Cora

Cora	Acc	NMI	F1	Precision	ARI
K-means	0.492	0.321	0.368	0.369	0.230
Spectral	0.367	0.127	0.318	0.193	0.031
BigClam	0.272	0.007	0.281	0.180	0.001
GraphEncoder	0.325	0.109	0.298	0.182	0.006
DeepWalk	0.484	0.327	0.392	0.361	0.243
DNGR	0.419	0.318	0.340	0.266	0.142
Circles	0.607	0.404	0.469	0.501	0.362
RTM	0.440	0.230	0.307	0.332	0.169
RMSC	0.407	0.255	0.331	0.227	0.090
TADW	0.560	0.441	0.481	0.396	0.332
GAE*	0.439	0.291	0.417	0.453	0.209
VGAE*	0.443	0.239	0.425	0.430	0.175
GAE	0.596	0.429	0.595	0.596	0.347
VGAE	0.609	0.436	0.609	0.609	0.346
ARGA	0.640	0.449	0.619	0.646	0.352
ARVGA	0.638	0.450	0.627	0.624	0.374
ARGA_DG	0.604	0.425	0.594	0.600	0.373
ARVGA_DG	0.463	0.387	0.455	0.524	0.265
ARGA_AX	0.597	0.455	0.579	0.593	0.366
ARVGA_AX	**0.711**	**0.526**	**0.693**	**0.710**	**0.495**

TABLE V: Clustering Results on Citeseer

| Citeseer	Acc	NMI	F1	Precision	ARI
K-means	0.540	0.305	0.409	0.405	0.279
Spectral	0.239	0.056	0.299	0.179	0.010
BigClam	0.250	0.036	0.288	0.182	0.007
GraphEncoder	0.225	0.033	0.301	0.179	0.010
DeepWalk	0.337	0.088	0.270	0.248	0.092
DNGR	0.326	0.180	0.300	0.200	0.044
Circles	0.572	0.301	0.424	0.409	0.293
RTM	0.451	0.239	0.342	0.349	0.203
RMSC	0.295	0.139	0.320	0.204	0.049
TADW	0.455	0.291	0.414	0.312	0.228
GAE*	0.281	0.066	0.277	0.315	0.038
VGAE*	0.304	0.086	0.292	0.331	0.053
GAE	0.408	0.176	0.372	0.418	0.124
VGAE	0.344	0.156	0.308	0.349	0.093
ARGA	0.573	0.350	0.546	0.573	0.341
ARVGA	0.544	0.261	0.529	0.549	0.245
ARGA_DG	0.479	0.231	0.446	0.456	0.203
ARVGA_DG	0.448	0.256	0.410	0.496	0.149
ARGA_AX	0.547	0.263	0.527	0.549	0.243
ARVGA_AX	**0.581**	**0.338**	**0.525**	**0.537**	**0.301**
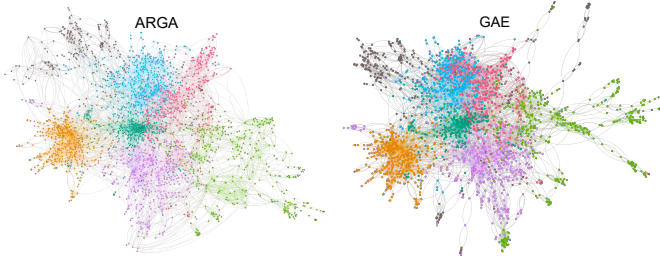
Fig. 6: Visualization with edges of the latent space of unsupervised ARGA and GAE trained on Cora data set. Colors indicate different clusters, and edges are represented with the links between nodes. Best view for both models

Metrics: Following [62], we employ five metrics to validate the clustering results: accuracy (Acc), F-one score (F1), normalized mutual information (NMI), precision and average rand index (ARI).

Experimental Results. The clustering results on the Cora, Citeseer and Pubmed data sets are given in table IV, table V and table VI. The results show that ARGA and ARVGA have achieved a dramatic improvement on all five metrics compared with all the other baselines. For instance, on Citeseer, ARGA has increased the accuracy from 6.1% compared with K-means to 154.7% compared with GraphEncoder; increased the F1
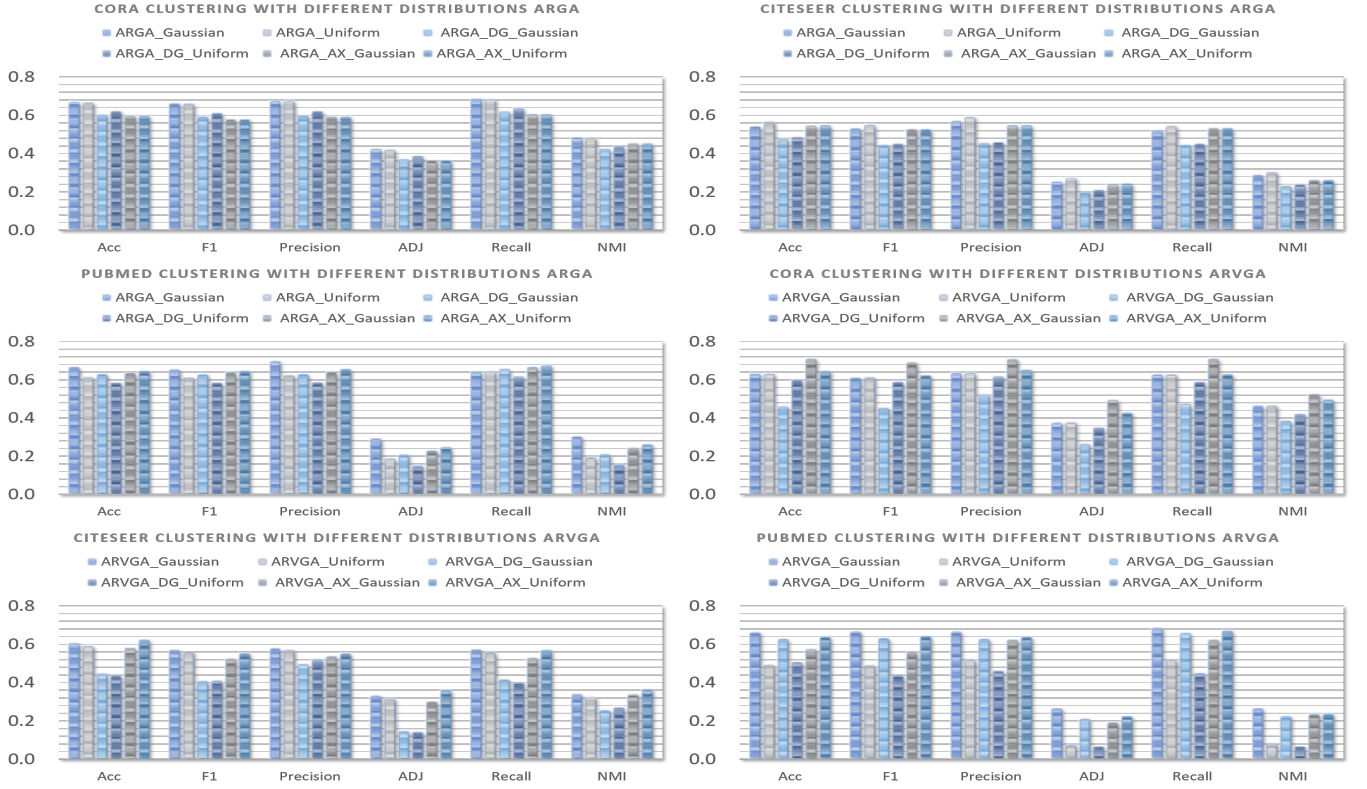
Fig. 7: The ARGA related models comparison on the **clustering** task with different prior distributions.

TABLE VI: Clustering Results on Pubmed

Pubmed	Acc	NMI	F1	Precision	ARI
K-means	0.398	0.001	0.195	0.579	0.002
Spectral	0.403	0.042	0.271	0.498	0.002
BigClam	0.394	0.006	0.223	0.361	0.003
GraphEncoder	0.531	0.209	0.506	0.456	0.184
DeepWalk	0.684	0.279	0.670	0.686	0.299
DNGR	0.458	0.155	0.467	0.629	0.054
RTM	0.574	0.194	0.444	0.455	0.148
RMSC	0.576	0.255	0.521	0.482	0.222
TADW	0.354	0.001	0.335	0.336	0.001
GAE*	0.581	0.196	0.569	0.636	0.162
VGAE*	0.504	0.162	0.504	0.631	0.088
GAE	0.672	0.277	0.660	0.684	0.279
VGAE	0.630	0.229	0.634	0.630	0.213
ARGA	0.668	**0.305**	0.656	**0.699**	0.295
ARVGA	**0.690**	0.290	**0.678**	0.694	**0.306**
ARGA_DG	0.630	0.212	0.629	0.631	0.209
ARVGA_DG	0.630	0.226	0.632	0.629	0.212
ARGA_AX	0.637	0.245	0.639	0.642	0.231
ARVGA_AX	0.640	0.239	0.644	0.639	0.226

score from 31.9% compared with TADW to 102.2% compared with DeepWalk; and increased NMI from 14.8% compared with K-means to 124.4% compared with VGAE.

Furthermore, as we can see from the three tables, the clustering results from approaches BigClam and DeepWalk, which only consider one perspective information of the graph, are inferior to the results from those which consider both topological information and node content of the graph. However, both purely GCNs-based approaches or the methods considering multi-view information still only obtain sub-optimal results compared to the adversarially regularized graph convolutional models.

The wide margin in the results between ARGA and GAE (and the others) has further demonstrated the superiority of our adversarially regularized graph autoencoder.

Parameter Study. We conducted experiments on Cora dataset with varying the dimension of embedding from 8 neurons to 1024 and report the results in Fig 5. All metrics demonstrated a similar fluctuation as the dimension of the embedding is increased. We cannot extract apparent trends to represent the relations between the embedding dimensions and the score of each clustering metric. This observation indicates that the unsupervised clustering task is more sensitive to the parameters compared to the supervised learning tasks (e.g., link prediction in Section V-A).

Graph Visualization with Linkages.

Inspired by [53], we visualized the well-learned latent space with the linkages of both GAE and our proposed ARGA trained on Cora data set. As shown in Fig. 6, many nodes in the latent space of GAE (Right side) which belong to the GREEN cluster have been located nearer to the PINK cluster. Similar circumstance happened in the bond between the RED cluster and the BLUE cluster, where some of nodes of RED mixed in the BLUE cluster. This could be caused by the unregularized embedding space, which is free for any structure. Adversarially
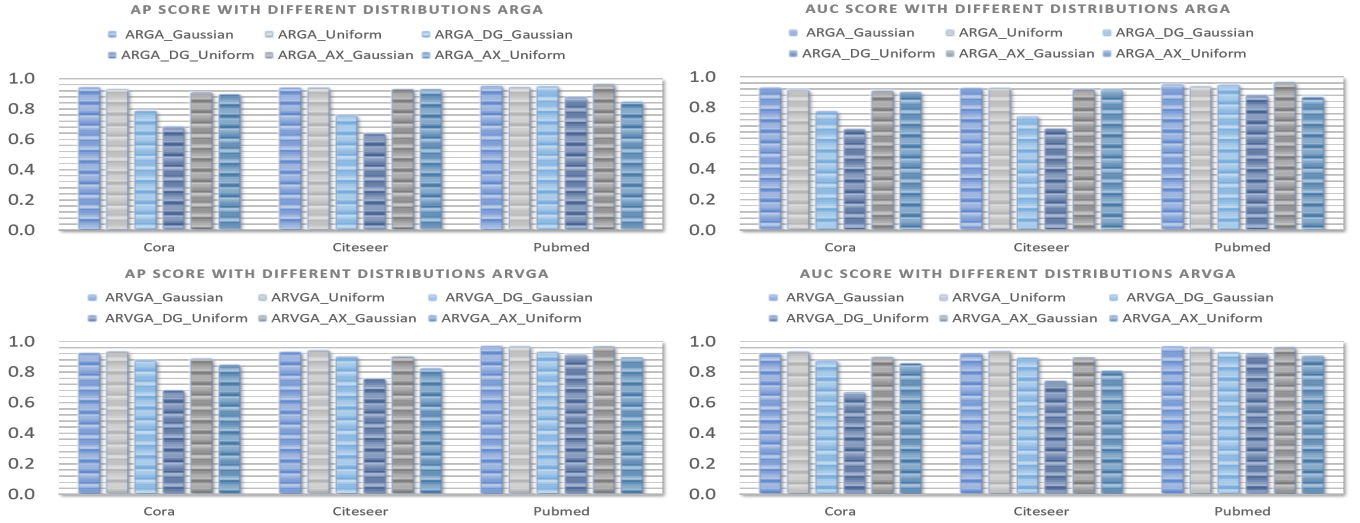
Fig. 8: The ARGA related models comparison on the **link prediction** task with different prior distributions.

regularized embedding shows better visualization with clear boundary line between two clusters. Considering the only difference between ARGA and the GAE is the adversarial training regularization scheme, it is reasonable to claim that adversarial regularization is helpful to enhance the quality of graph embedding.

C. ARGA Architectures Comparison

In this section, we construct six versions of the model: *adversarially regularized graph autoencoder* (ARGA), *adversarially regularized graph autoencoder with graph convolutional decoder* (ARGA_DG) and *adversarially regularized graph autoencoder for reconstructing both graph structure and node content* (ARGA_AX) and their variational versions. Meanwhile, we conduct all experiments with a prior Gaussian distribution and a prior Uniform Distribution respectively for every model. We analyze the comparison experiments and try to figure out the reasons behind the results. The experimental results are illustrated in Fig, 7 and 8.

Gaussian Distribution vs Uniform Distribution. The performance of the proposed models is not very sensitive to the prior distributions, especially for the node clustering task. As shown in Fig. 7, if we compare the results of two distributions with the same metric, the results from one same model, in most cases, are very similar.

As for the link prediction (Fig. 8), the Uniform distribution dramatically lowers the performance of ARGA_DG on all datasets and metrics, compared to the results with Gaussian distribution. ARGA and its variational version are not as sensitive to the different distributions as ARGA_DG models. The standard version of ARGA with Gaussian distribution slightly outperforms the ones with Uniform distribution. The situation reversed with the variational ARGA models.

Decoders and Reconstructions. As shown in Fig 8, the ARGA with the Gaussian distribution and inner product decoder for reconstructing graph structure has a significant advantage in link prediction since $p(\hat{\mathbf{A}}_{ij} = 1|\mathbf{z}_i, \mathbf{z}_j)$ is

designed to predict whether there is a link between two nodes. Simply replacing the decoder with graph convolutional layers to reconstruct adjacency matrix $\hat{\mathbf{A}}$ (ARGA_DG) has a sub-optimal performance in link prediction compared to ARGA. According to the statistic in Fig. 7, although the performance of ARGA_DG on clustering is comparable with original ARGA, there is still a gap between these two variations. Two graph convolutional layers in the decoder cannot effectively decode the topological information of the graph, which leads to the sub-optimal results. The model with graph convolutional decoder for reconstructing both topological information $\mathbf{A}$ and node content $\mathbf{X}$ (ARGA_AX) may prove this hypothesis. As can be seen in Fig. 7 and 8, ARGA_AX has dramatically improved the performance on both link prediction and clustering compared to ARGA_DG which purely reconstructs the topological structure. ARGA and ARGA_AX have very similar performances on both link prediction and clustering. The variational version of ARGA_AX (ARVGA_AX) has outstanding performance on clustering which has achieved 12.2% improvement on clustering accuracy on Cora dataset and 5.4% improvement on Citeseer dataset compared to ARVGA.

D. Time Complexity on Convollution

Our graph encoder requires the computation $\mathbf{Z}' = \phi(\widetilde{\mathbf{D}}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{Z}^{(l)}\mathbf{W}^{(l)})$, which can be computed efficiently using sparse matrix computation. Specifically, let $\mathbf{P} = \widetilde{\mathbf{D}}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-\frac{1}{2}}$, which is the Laplacian matrix. As $\widetilde{\mathbf{D}}$ is a diagonal matrix, the inverse of $\widetilde{\mathbf{D}}$ is the inverse of its diagonal values with time complexity $O(|V|)$. Let $\mathbf{W}^{(l)} \in \mathbb{R}^{m \times d}$, and $\mathbf{Z}^{(l)} \in \mathbb{R}^{n \times m}$. The complexity of our convolution operation is $O(|\mathbf{E}|md)$, as $\widetilde{\mathbf{A}}\mathbf{Z}^{(l)}$ can be efficiently implemented as a product of a sparse matrix with a dense matrix (See [53] for details).

We conducted experiments with six ARGA models and two GAE models for the training time comparison. We conducted 200 training epochs for link prediction task of each model on Cora data set and report the average time for the

comparison. The results are shown in Fig. 9. The results show that ARGA models take more time than original GAE models due to the additional regularization module in the architecture. ARGA_AX model requires more computation for simultaneously reconstructing both topological structure (A) and node characteristics (X).
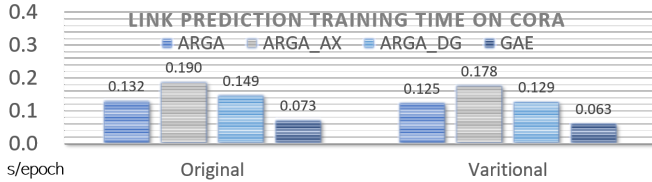

Fig. 9: Average training time per epoch of link prediction. (Left) Original Architectures; (Right) Varitional Architectures.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel adversarial graph embedding framework for graph data. We argue that most existing graph embedding algorithms are unregularized methods that ignore the data distributions of the latent representation and suffer from inferior embedding in real-world graph data. We proposed an adversarial training scheme to *regularize* the latent codes and enforce the latent codes to match a prior distribution. The adversarial module is jointly learned with a graph convolutional autoencoder to produce a robust representation. We also exploited some interesting variations of ARGA like ARGA_DG and ARGA_AX to discuss the impact of graph convolutional decoder for reconstructing both graph structure and node content. Experiment results demonstrated that our algorithms ARGA and ARVGA outperform baselines in link prediction and node clustering tasks.

There are several directions for the adversarially regularized graph autoencoders (ARGA). We will investigate how to use the ARGA model to generate some realistic graphs [63], which may help discover new drugs in biological domains. We will also study how to incorporate label information into ARGA to learn robust graph embedding.

REFERENCES

[1] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[2] C. Wang, S. Pan, G. Long, X. Zhu, and J. Jiang, "Mgae: Marginalized graph autoencoder for graph clustering," in *CIKM*. ACM, 2017, pp. 889–898.

[3] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, no. 99, pp. 1–13, 2018.

[4] H. Cai, V. W. Zheng, and K. C.-C. Chang, "A comprehensive survey of graph embedding: Problems, techniques and applications," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[5] C. Shi, B. Hu, W. X. Zhao, and S. Y. Philip, "Heterogeneous information network embedding for recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 2, pp. 357–370, 2018.

[6] S. Pan, J. Wu, and X. Zhu, "Cogboost: Boosting for fast cost-sensitive graph classification," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 2933–2946, 2015.

[7] S. Pan, J. Wu, X. Zhu, C. Zhang, and P. S. Yu, "Joint structure feature exploration and regularization for multi-task graph classification," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 715–728, 2016.

[8] P. Cui, X. Wang, J. Pei, and W. Zhu, "A survey on network embedding," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[9] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip, "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17–37, 2017.

[10] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "Network representation learning: A survey," *IEEE Transactions on Big Data*, 2018.

[11] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *The World Wide Web Conference*, 2019, pp. 2022–2032.

[12] X. Cao, Y. Zheng, C. Shi, J. Li, and B. Wu, "Link prediction in schema-rich heterogeneous information network," in *PAKDD*. Springer, 2016, pp. 449–460.

[13] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," *Knowledge-Based Systems*, vol. 151, pp. 78–94, 2018.

[14] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *SIGKDD*. ACM, 2014, pp. 701–710.

[15] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *SIGKDD*. ACM, 2016, pp. 855–864.

[16] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: large-scale information network embedding," in *WWW*, 2015, pp. 1067–1077.

[17] L. Tang and H. Liu, "Leveraging social media networks for classification," *DMKD*, vol. 23, no. 3, pp. 447–478, 2011.

[18] S. Cao, W. Lu, and Q. Xu, "Grarep: Learning graph representations with global structural information," in *CIKM*. ACM, 2015, pp. 891–900.

[19] M. Ou, P. Cui, J. Pei, and et.al, "Asymmetric transitivity preserving graph embedding." in *KDD*, 2016, pp. 1105–1114.

[20] X. Wang, P. Cui, J. Wang, and et.al, "Community preserving network embedding." in *AAAI*, 2017, pp. 203–209.

[21] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang, "Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 459–467.

[22] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *arXiv preprint arXiv:1901.00596*, 2019.

[23] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *SIGKDD*. ACM, 2016, pp. 1225–1234.

[24] S. Cao, W. Lu, and Q. Xu, "Deep neural networks for learning graph representations." in *AAAI*, 2016, pp. 1145–1152.

[25] X. Shen and F. Chung, "Deep network embedding for graph representation learning in signed networks," *IEEE Transactions on Cybernetics*, pp. 1–8, 2018.

[26] A. Makhzani, J. Shlens, N. Jaitly, and et.al, "Adversarial autoencoders," *ICLR*, 2016.

[27] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *ICLR*, 2017.

[28] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial networks," *ICLR*, 2017.

[29] V. Dumoulin, I. Belghazi, B. Poole, and et.al, "Adversarially learned inference," *ICLR*, 2017.

[30] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *ICLR*, 2016.

[31] D. Zhu, P. Cui, Z. Zhang, J. Pei, and W. Zhu, "High-order proximity preserved embedding for dynamic networks," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[32] H. Gui, J. Liu, F. Tao, M. Jiang, B. Norick, L. Kaplan, and J. Han, "Embedding learning with events in heterogeneous information networks,"

IEEE transactions on knowledge and data engineering, vol. 29, no. 11, pp. 2428–2441, 2017.

[33] T. Mikolov, K. Chen, G. Corrado, and et.al, "Efficient estimation of word representations in vector space," *ICLR*, 2013.

[34] H. Chen, B. Perozzi, Y. Hu, and S. Skiena, "HARP: hierarchical representation learning for networks," in *AAAI*, 2018.

[35] J. Li, J. Zhu, and B. Zhang, "Discriminative deep random walk for network classification," in *ACL*, vol. 1, 2016, pp. 1004–1013.

[36] B. Perozzi, V. Kulkarni, and S. Skiena, "Walklets: Multiscale graph embeddings for interpretable network classification," *Computing Research Repository*, 2016.

[37] X. Shen, S. Pan, W. Liu, Y. Ong, and Q. Sun, "Discrete network embedding," in *IJCAI*, 2018, pp. 3549–3555.

[38] C. Yang, Z. Liu, D. Zhao, and et.al, "Network representation learning with rich text information." in *IJCAI*, 2015, pp. 2111–2117.

[39] S. Pan, J. Wu, X. Zhu, C. Zhang, and Y. Wang, "Tri-party deep network representation," in *IJCAI*, 2016, pp. 1895–1901.

[40] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "User profile preserving social network embedding," in *IJCAI*. AAAI Press, 2017, pp. 3378–3384.

[41] L. Liao, X. He, H. Zhang, and T.-S. Chua, "Attributed social network embedding," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2018.

[42] J. Li, H. Dani, X. Hu, J. Tang, Y. Chang, and H. Liu, "Attributed network embedding for learning in a dynamic environment," in *CIKM*, 2017, pp. 387–396.

[43] X. Huang, J. Li, and X. Hu, "Label informed attributed network embedding," in *WSDM*. ACM, 2017, pp. 731–739.

[44] H. Yang, S. Pan, P. Zhang, L. Chen, D. Lian, and C. Zhang, "Binarized attributed network embedding," in *ICDM*. IEEE, 2018, pp. 1476–1481.

[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, and et.al, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[46] J. Glover, "Modeling documents with generative adversarial networks," *NIPS 2016 Workshop on Adversarial Training*, 2016.

[47] Q. Dai, Q. Li, J. Tang, and D. Wang, "Adversarial network embedding," in *AAAI*, 2018.

[48] W. Yu, C. Zheng, W. Cheng, C. C. Aggarwal, D. Song, B. Zong, H. Chen, and W. Wang, "Learning deep network representations with adversarially regularized autoencoders," in *SIGKDD*. ACM, 2018, pp. 2663–2671.

[49] H. Wang, J. Wang, J. Wang, M. Zhao, W. Zhang, F. Zhang, X. Xie, and M. Guo, "Graphgan: Graph representation learning with generative adversarial nets," *AAAI*, 2018.

[50] M. Ding, J. Tang, and J. Zhang, "Semi-supervised learning on graphs with generative adversarial nets," in *CIKM*. ACM, 2018, pp. 913–922.

[51] Y. F. Binbin Hu and C. Shi, "Adversarial learning on heterogeneous information network," in *KDD*. ACM, 2019.

[52] S. Pan, R. Hu, G. Long, J. Jiang, L. Yao, and C. Zhang, "Adversarially regularized graph autoencoder for graph embedding." in *IJCAI*, 2018, pp. 2609–2615.

[53] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *NIPS*, 2016.

[54] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *ICLR*, 2014.

[55] X. Zhou, F. Shen, L. Liu, W. Liu, L. Nie, Y. Yang, and H. T. Shen, "Graph convolutional network hashing," *IEEE Transactions on Cybernetics*, pp. 1–13, 2018.

[56] Q. Lu and L. Getoor, "Link-based classification," in *ICML*, 2003, pp. 496–503.

[57] P. Sen, G. Namata, M. Bilgic, and et.al, "Collective classification in network data," *AI magazine*, vol. 29, no. 3, p. 93, 2008.

[58] G. Namata, B. London, L. Getoor, and et.al, "Query-driven active surveying for collective classification," in *MLG*, 2012.

[59] F. Tian, B. Gao, Q. Cui, and et.al, "Learning deep representations for graph clustering." in *AAAI*, 2014, pp. 1293–1299.

[60] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *NIPS*, 2012, pp. 539–547.

[61] J. Chang and D. Blei, "Relational topic models for document networks," in *Artificial Intelligence and Statistics*, 2009, pp. 81–88.

[62] R. Xia, Y. Pan, L. Du, and et.al, "Robust multi-view spectral clustering via low-rank and sparse decomposition." in *AAAI*, 2014, pp. 2149–2155.

[63] J. You, R. Ying, X. Ren, W. Hamilton, and J. Leskovec, "Graphrnn: Generating realistic graphs with deep auto-regressive models," in *ICML*, 2018, pp. 5694–5703.

Shirui Pan Shirui Pan received the Ph.D. degree in computer science from the University of Technology Sydney (UTS), Ultimo, NSW, Australia. He is a Lecturer with the Faculty of Information Technology, Monash University, Australia. His research interests include data mining and machine learning. To date, Dr Pan has published over 60 research papers in top-tier journals and conferences, including the IEEE Transactions on Neural Networks and Learning Systems (TNNLS), IEEE Transactions on Knowledge and Data Engineering (TKDE), and IEEE Transactions on Cybernetics (TCYB).

Ruiqi Hu received the Ph.D. degree in computer science from the University of Technology Sydney (UTS), Ultimo, NSW, Australia. He is currently a Research Associate in the Centre for Artificial Intelligence (CAI), UTS. His research concentrates on data mining and graph embedding. His recent works have been accepted by top-tier conferences, such as International Joint Conference on Artificial Intelligence (IJCAI) and Conference on Information and Knowledge Management (CIKM).

Sai-fu Fung received the Ph.D. degree in in politics and international studies from the University of Warwick, Coventry, U.K. He is currently with the Department of Applied Social Sciences, City University of Hong Kong, Hong Kong. His current research interests include comparative methods and democratisation and politics of East and Southeast Asia.
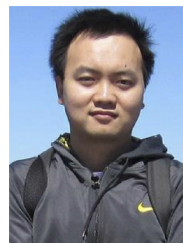
Guodong Long received his Ph.D. degree in computer science from University of Technology, Sydney (UTS), Australia, in 2014. He is a Senior Lecturer and core member in the Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology, Sydney (UTS). His research focuses on machine learning, data mining and cloud computing. To date, Dr Long has published over 50 research papers on highly ranked conference and jounals. He also served as a reviewer for a few jounals including IEEE TPAMI, TNNLS, TCYB, TKDE, and TSE.

Jing Jiang is currently a Lecturer at the Centre for Artificial Intelligence, Faculty of Engineering and Information Technology at the University of Technology Sydney (UTS), Australia. She has a PhD degree in Information Technology. Her research interest lies in data mining and machine learning applications with the focuses on deep reinforcement learning and sequential decision-making.

Chengqi Zhang (SM'95) received his Ph.D. degree in computer science from the University of Queensland, Brisbane, Australia, in 1991 and the D.Sc. degree from Deakin University, Geelong, Australia, in 2002. Since December 2001, he has been a Distinguished Professor of Information Technology with the University of Technology Sydney (UTS), Sydney, Australia, and he has been the Director of the UTS Priority Investment Research Centre for Quantum Computation and Intelligent Systems since April 2008. His research interests mainly focus on data mining and its applications. He is General Co-Chair of KDD 2015 in Sydney, the Local Arrangements Chair of IJCAI-2017 in Melbourne, a Fellow of the Australian Computer Society and a Senior Member of the IEEE.