Unsupervised Attributed Multiplex Network Embedding

Chanyoung Park\textsuperscript{1}, Donghyun Kim\textsuperscript{2}, Jiawei Han\textsuperscript{1}, Hwanjo Yu\textsuperscript{3}

\textsuperscript{1}Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA
\textsuperscript{2}Yahoo! Research, CA, USA
\textsuperscript{3}Department of Computer Science and Engineering, Pohang University of Science and Technology, Korea
pcy1302@illinois.edu, donghyun.kim@verizonmedia.com, hanj@illinois.edu, hwanjoyu@postech.ac.kr

Abstract

Nodes in a multiplex network are connected by multiple types of relations. However, most existing network embedding methods assume that only a single type of relation exists between nodes. Even for those that consider the multiplexity of a network, they overlook node attributes, resort to node labels for training, and fail to model the global properties of a graph. We present a simple yet effective unsupervised network embedding method for attributed multiplex network called \textbf{DMGI}, inspired by Deep Graph Infomax (DGI) that maximizes the mutual information between local patches of a graph, and the global representation of the entire graph. We devise a systematic way to jointly integrate the node embeddings from multiple graphs by introducing 1) the consensus regularization framework that minimizes the disagreements among the relation-type specific node embeddings, and 2) the universal discriminator that discriminates true samples regardless of the relation types. We also show that the attention mechanism infers the importance of each relation type, and thus can be useful for filtering unnecessary relation types as a preprocessing step. Extensive experiments on various downstream tasks demonstrate that DMGI outperforms the state-of-the-art methods, even though DMGI is fully unsupervised.

1 Introduction

Analyzing and mining useful knowledge in graphs have been an actively researched topic for decades both in academia and industry. Among various graph mining techniques, network embedding, which learns low-dimensional vector representations for nodes in a graph, is shown to be especially effective for various network-based tasks (Tang et al. 2015; Wang et al. 2017; Meng et al. 2019).

However, most existing network embedding methods assume that only a single type of relation exists between nodes (Veličković et al. 2017; 2019; Kipf and Welling 2016), whereas in reality networks are multiplex (De Domenico et al. 2013) in nature, i.e., with multiple types of relations. Taking the publication network as an example, two papers can be connected due to various reasons, such as authors (two papers are authored by a common author), citation (one paper cites the other), or keywords (two papers share common keywords). As another example, in a movie database network, two movies can be connected via a common director, or a common actor.

Although different types of relations can independently form different graphs, \textit{these graphs are related}, and thus can mutually help each other for various downstream tasks. As a concrete example of the publication network, although it is hard to infer the topic of a paper only from its citations (citations can be diverse), also knowing other papers written by the same authors will help predict its topic, because authors usually work on a specific research topic. Furthermore, nodes in graphs may contain attribute information, which plays important roles in many applications (Zhang et al. 2018b). For example, if we are additionally given the abstract of the papers in the publication network, it will be much easier to infer their topics. As such, the main challenge is to learn a consensus representation of a node that not only considers its multiplexity, but also its \textit{attributes}.

Several recent studies have been conducted for multiplex network embedding, however, some issues remain that need further consideration. First, previous methods (Qu et al. 2017; Zhang et al. 2018a; Shi et al. 2018; Liu et al. 2017) focus on the integration of multiple graphs, but overlook \textbf{node attributes}. Second, even for those that consider node attributes (Schlichtkrull et al. 2018; Wang et al. 2019), they require \textbf{node labels} for training. However, as node labeling is often expensive and time-consuming, it would be the best if a method can show competitive performance even without any label. Third, most of these methods fail to model the \textbf{global properties} of a graph, because they are based on random walk-based skip-gram model or graph convolutional network (GCN) (Kipf and Welling 2016), both of which are known to be effective for capturing the local graph structure (Yadav et al. 2019). More precisely, nodes that are “close” (i.e., within the same context window or neighborhoods) in the graph are trained to have similar representations, whereas nodes that are far apart do not have similar representations, even though they are structurally similar (Ribeiro, Suverese, and Figueiredo 2017).

Keeping these limitations in mind, we propose a simple yet effective unsupervised method for embedding attributed multiplex networks. The core building block of our proposed method is Deep Graph Infomax (DGI) (Veličković et al. 2019) that aims to learn a node encoder that maximizes the mutual information between local patches of a graph,
and the global representation of the entire graph. DGI is the workhorse method for our task, because it 1) naturally integrates the node attributes by using a GCN, 2) is trained in a fully unsupervised manner, and 3) captures the global properties of the entire graph. However, it is challenging to apply DGI, which is designed for embedding a single network, to a multiplex network in which the interactions among multiple relation types, and the importance of each relation type should be considered.

In this paper, we present a systematic way to jointly integrate the embeddings from multiple types of relations between nodes, so as to facilitate them to mutually help each other learn high-quality embeddings useful for various downstream tasks. More precisely, we introduce the consensus regularization framework that minimizes the disagreements among the relation-type specific node embeddings, and the universal discriminator that discriminates true samples, i.e., ground truth “(graph-level summary, local patch)” pairs, regardless of the relation types. Moreover, we demonstrate that through the attention mechanism, we can infer the importance of each relation type in generating the consensus node embeddings, which can be used for filtering unnecessary relation types as a preprocessing step. Our extensive experiments demonstrate that our proposed method, Deep Multiplex Graph Infomax (DMGI), outperforms the state-of-the-art attributed multiplex network embedding methods in terms of node clustering, similarity search, and especially, node classification even though DMGI is fully unsupervised.

2 Problem Statement

Definition 1. (Attributed Multiplex Network) An attributed multiplex network is a network \( G = \{ G^1, G^2, ..., G^{|R|} \} = \{ V, E, X \} \), where \( G^r = \{ V, E^r, X \} \) is a graph of the relation type \( r \in R, V \) is the set of \( n \) nodes, \( E = \bigcup_{r \in R} E^r \subseteq V \times V \) is the set of all edges with relation type \( r \in R \), and \( X \in \mathbb{R}^{n \times f} \) is a matrix that encodes node attributes information for \( n \) nodes. Note that \( |R| > 1 \) for multiplex networks, and \( |R| = 1 \) for a single network. Given the network \( G, A = \{ A^{(1)}, ..., A^{(|R|)} \} \) is a set of adjacency matrices, where \( A^{(r)} \in \{0, 1\}^{V \times V} \) is an adjacency matrix of the network \( G^r \).

Task: Unsupervised Attributed Multiplex Network Embedding. Given an attributed multiplex network \( G = \{ V, E, X \} \), and the set of adjacency matrices \( A \), the task of unsupervised attributed multiplex network embedding is to learn a \( d \)-dimensional vector representation \( z_i \in Z \in \mathbb{R}^{n \times d} \) for each node \( v_i \in V \) without using any labels.

3 Unsupervised Attributed Multiplex Network Embedding

We begin by introducing Deep Graph Infomax (DGI) (Veličković et al. 2019), then we discuss about its limitations, and present our proposed method.

Deep Graph Infomax (DGI). Veličković et al. (2019) proposed an unsupervised method for learning node representations, called DGI, that relies on the infomax principle (Linsker 1988). More precisely, DGI aims to learn a low-dimensional vector representation for each node \( v_i \), i.e., \( h_i \in \mathbb{R}^d \), such that the average mutual information (MI) between the graph-level (global) summary representation \( s \in \mathbb{R}^d \) and the representations of the local patches \( \{h_1, h_2, ..., h_n\} \) is maximized. To this end, DGI introduces a discriminator \( D \) that discriminates the true samples, i.e., \( (h_i, s) \), from its negative counterparts, i.e., \( (\hat{h}_i, s) \):

\[
L = \sum_{v_i \in V} \log D(h_i, s) + \sum_{j=1}^n \log (1 - D(\hat{h}_j, s))
\]

where \( h_i = \sigma \left( \sum_{j \in N(i)} \frac{1}{c_{ij}} x_j W \right), N(i) \) is the set of neighboring nodes of \( v_i \) including \( v_i \) itself, \( W \in \mathbb{R}^{f \times d} \), and \( c_{ij} \) is a normalizing constant for edge \( (v_i, v_j) \), \( s = \sigma \left( \frac{1}{n} \sum_{i=1}^n h_i \right) \), and \( \sigma \) is the sigmoid nonlinearity. Negative patch representation \( \hat{h}_j \) is obtained by row-wise shuffling the original attribute matrix \( X \). Veličković et al. (2019) theoretically proved that the binary cross entropy loss shown in Eqn. 1 amounts to maximizing the mutual information (MI) between \( h_i \) and \( s \), based on the Jensen-Shannon divergence (Veličković et al. 2019). Refer to Section 3.3 of (Veličković et al. 2019) for the detailed proof. As the local patch representations \( \{h_1, h_2, ..., h_n\} \) are learned to preserve the MI with the graph-level representation \( s \), each \( h_i \) is expected to capture the global properties of the entire graph.

Limitation. Despite its effectiveness, DGI is designed for a single attributed network, and thus it is not straightforward to apply it to a multiplex network. As a naive extension of DGI to a multiplex attributed network, we can independently apply DGI to each graph formed by each relation type, and then compute the average of the embeddings obtained from each graph to get the final node representations. However, we argue that this fails to model the multiplexity of the network, because the interactions among the node embeddings from different relation types is not captured. Thus, we need a more systematic way to integrate multiple independent models to obtain the final consensus embedding that every model can agree on.

3.1 Deep Multiplex Graph Infomax: DMGI

We present our unsupervised method for embedding an attributed multiplex network. We first describe how to independently model each graph pertaining to each relation type, then explain how to jointly integrate them to finally obtain the consensus node embedding matrix.

Relation-type specific Node Embedding. For each relation type \( r \in R \), we introduce a relation-type specific node encoder \( g_r : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^{f \times d} \) to generate the relation-type specific node embedding matrix \( H^{(r)} \) of nodes in \( G^{(r)} \). The encoder is a single–layered GCN:

\[
H^{(r)} = g_r(X, A^{(r)} W^{(r)}) = \sigma \left( \tilde{D}^{-\frac{1}{2}} \tilde{A}^{(r)} \tilde{D}^{-\frac{1}{2}} X W^{(r)} \right)
\]

where \( \tilde{A}^{(r)} = A^{(r)} + w I_n, \tilde{D}_{ii} = \sum_j \tilde{A}_{ij} W^{(r)} \in \mathbb{R}^{f \times d} \) is a trainable weight matrix of the relation-type specific decoder \( g_r \), and \( \sigma \) is the ReLU nonlinearity. Unlike conventional GCNs (Kipf and Welling 2016), we control the weight...
of the self-connections by introducing a weight $w \in \mathbb{R}$. Larger $w$ indicates that the node itself plays a more important role in generating its embedding, which in turn diminishes the importance of its neighboring nodes. Then, we compute the graph-level summary representation $\mathbf{s}^{(r)}$ that summarizes the global content of the graph $\mathcal{G}^{(r)}$. We employ a readout function $\text{Readout} : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^d$:

$$\mathbf{s}^{(r)} = \text{Readout}(\mathbf{H}^{(r)}) = \sigma \left( \frac{1}{n} \sum_{i=1}^{n} \mathbf{h}_i^{(r)} \right)$$  \hspace{1cm} (3)$$

where $\sigma$ is the logistic sigmoid nonlinearity, and $\mathbf{h}_i^{(r)}$ denotes the $i$-th row vector of the matrix $\mathbf{H}^{(r)}$. We also note that various pooling methods such as maxpool, and SAGPool (Lee, Lee, and Kang 2019) can be used as $\text{Readout}^{(r)}$.

Next, given the relation-type specific node embedding matrix $\mathbf{H}^{(r)}$, and its graph-level summary representation $\mathbf{s}^{(r)}$, we compute the relation-type specific cross entropy:

$$\mathcal{L}^{(r)} = \sum_{v \in \mathcal{V}} \log \mathcal{D} \left( \mathbf{h}_v^{(r)}, \mathbf{s}^{(r)} \right) + \sum_{j=1}^{d} \log \left( 1 - \mathcal{D} \left( \mathbf{h}_j^{(r)}, \mathbf{s}^{(r)} \right) \right)$$  \hspace{1cm} (4)$$

where $\mathcal{D} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a discriminator that scores patch-summary representation pairs, i.e., $(\mathbf{h}_v^{(r)}, \mathbf{s}^{(r)})$. In this paper, we apply a simple bilinear scoring function as it empirically performs the best in our experiments:

$$\mathcal{D} \left( \mathbf{h}_v^{(r)}, \mathbf{s}^{(r)} \right) = \sigma (\mathbf{h}_v^{(r)T} \mathbf{M}^{(r)} \mathbf{s}^{(r)})$$  \hspace{1cm} (5)$$

where $\sigma$ is the logistic sigmoid nonlinearity, and $\mathbf{M}^{(r)} \in \mathbb{R}^{d \times d}$ is a trainable scoring matrix. To generate the negative node embedding $\tilde{\mathbf{h}}_v^{(r)}$, we corrupt the original attribute matrix by shuffling it in the row-wise manner (Veličković et al. 2019), i.e., $\tilde{\mathbf{X}} \leftarrow \mathbf{X}$, and reuse the encoder in Eqn. 2. i.e., $\tilde{\mathbf{H}}^{(r)} = g_r(\tilde{\mathbf{X}}, \mathbf{A}^{(r)}|\mathbf{W}^{(r)})$.

**Joint Modeling and Consensus Regularization.** Heretofore, by independently maximizing the average MI between the local patches $\{\mathbf{h}^{(r)}, \mathbf{h}_2^{(r)}, ..., \mathbf{h}_n^{(r)}\}$ and the graph-level summary $\mathbf{s}^{(r)}$ pertaining to each graph $\mathcal{G}^{(r)}(\forall r \in \mathcal{R})$, we obtained relation-type specific node embedding matrix $\mathbf{H}^{(r)}$ that captures the global information in $\mathcal{G}^{(r)}$. However, as each $\mathbf{H}^{(r)}$ is trained independently for each $r \in \mathcal{R}$, these embedding matrices only contain relevant information regarding each relation type, and therefore fail to take advantage of the multiplexity of the network. This motivates us to develop a systematic way to jointly integrate the embeddings from different relation types, so as to facilitate them to mutually help each other learn high-quality embeddings.

To this end, we introduce the consensus embedding matrix $\mathbf{Z} \in \mathbb{R}^{n \times d}$ on which every relation-type specific node embedding matrix $\mathbf{H}^{(r)}$ can agree. More precisely, we introduce the consensus regularization framework that consists of 1) a regularizer minimizing the disagreements between the set of original node embeddings, i.e., $\{\mathbf{H}^{(r)} | r \in \mathcal{R}\}$ and the consensus embedding $\mathbf{Z}$, and 2) another regularizer maximizing the disagreement between the corrupted node embeddings, i.e., $\{\tilde{\mathbf{H}}^{(r)} | r \in \mathcal{R}\}$, and the consensus embedding $\mathbf{Z}$, which are formulated as follows:

$$\ell_{cs} = \left[ \mathbf{Z} - \mathbf{Q} \left( \{\mathbf{H}^{(r)} | r \in \mathcal{R}\} \right) \right]^2 - \left[ \mathbf{Z} - \mathbf{Q} \left( \{\tilde{\mathbf{H}}^{(r)} | r \in \mathcal{R}\} \right) \right]^2$$  \hspace{1cm} (6)$$

where $\mathbf{Q}$ is an aggregation function that combines a set of node embedding matrices from multiple relation types into a single embedding matrix, i.e., $\mathbf{H} \in \mathbb{R}^{n \times d}$. $\mathbf{Q}$ can be any pooling method that can handle permutation invariant input, such as set2set (Vinyals, Bengio, and Kudlur 2015) or Set Transformer (Lee et al. 2019). However, considering the efficiency of the method, we simply employ average pooling, i.e., computing the average of the set of embedding matrices:

$$\mathbf{H} = \mathbf{Q} \left( \{\mathbf{H}^{(r)} | r \in \mathcal{R}\} \right) = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \mathbf{H}^{(r)}$$  \hspace{1cm} (7)$$

It is important to note that the scoring matrix $\mathbf{M}^{(r)}$ in Eqn. 5 is shared among all the relations $r \in \mathcal{R}$, i.e., $\mathbf{M} = \mathbf{M}^{(1)} = \mathbf{M}^{(2)} = ... = \mathbf{M}^{(|\mathcal{R}|)}$. The intuition is to learn the universal discriminator that is capable of scoring the true pairs higher than the negative pairs regardless the relation types. We argue that the universal discriminator facilitates the joint modeling of different relation types together with the consensus regularization.

Finally, we jointly optimize the sum of all the relation-type specific loss in Eqn. 4, and the consensus regularization in Eqn. 6 to obtain the final objective $\mathcal{J}$ as follows:

$$\mathcal{J} = \sum_{r \in \mathcal{R}} \mathcal{L}^{(r)} + \alpha \ell_{cs} + \beta ||\Theta||^2$$  \hspace{1cm} (8)$$

where $\alpha$ controls the importance of the consensus regularization, $\beta$ is a coefficient for L2 regularization on $\Theta$, which is a set of trainable parameters, i.e., $\Theta = \{\{\mathbf{W}^{(r)} | r \in \mathcal{R}\}, \mathbf{M}, \mathbf{Z}\}$, and $\mathcal{J}$ is optimized by Adam optimizer. Figure 1 illustrates the overview of DMGI.

**Discussion.** Despite its efficiency, the above average pooling scheme in Eqn. 7 treats all the relations equally, whereas, as will be shown in the experiments, some relation type is more beneficial for a certain downstream task than others. For example, the co-authorship information between two papers plays a more significant role in predicting the topic of a
paper compared with their citation information; eventually, these two information mutually help each other to more accurately predict the topic of a paper. Therefore, we can adopt the attention mechanism (Bahdanau, Cho, and Bengio 2014) to distinguish between different relation types as follows:

$$h_i = Q \left( (h^{(r)} | r \in R) \right) = \sum_{r \in R} a_i^{(r)} h^{(r)}$$  \hspace{1cm} (9)

where $$a_i^{(r)}$$ denotes the importance of relation $$r$$ in generating the final embedding of node $$v_i$$ defined as:

$$a_i^{(r)} = \frac{\exp(q^{(r)} \cdot h_i^{(r)})}{\sum_{r' \in R} \exp(q^{(r')} \cdot h_i^{(r')})}$$  \hspace{1cm} (10)

where $$q^{(r)} \in \mathbb{R}^d$$ is the feature vector of relation $$r$$.

**Extension to Semi-Supervised Learning.** It is important to note that DMGI is trained in a *fully unsupervised* manner. However, in reality, nodes are sometimes associated with label information, which can guide the training of node embeddings even with a small amount (Kipf and Welling 2016; Qu et al. 2017). To this end, we introduce a *semi-supervised module* into our framework that predicts the labels of labeled nodes from the consensus embedding $$Z$$. More precisely, we minimize the cross-entropy error over the labeled nodes:

$$\ell_{sup} = -\frac{1}{|Y_L|} \sum_{i \in Y_L} \sum_{c} Y_{li} \ln \hat{Y}_{li}$$  \hspace{1cm} (11)

where $$Y_L$$ is the set of node indices with labels, $$Y \in \mathbb{R}^{n \times c}$$ is the ground truth label, $$\hat{Y} = \text{softmax}(f(Z))$$ is the output of a softmax layer, and $$f : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{n \times c}$$ is a classifier that predicts the label of a node from its embedding, which is a single fully connected layer in this work. The final objective function with the semi-supervised module is:

$$J_{semi} = \sum_{r \in R} L^{(r)} + \alpha \ell_{cls} + \beta ||\Theta|| + \gamma \ell_{sup}$$  \hspace{1cm} (12)

where $$\gamma$$ the coefficient of the semi-supervised module.

### 4 Experiments

**Dataset.** To make fair comparisons with HAN (Wang et al. 2019), which is the most relevant baseline method, we evaluate our proposed method on the datasets used in their original paper (Wang et al. 2019), i.e., ACM, DBLP, and IMDB. We used publicly available ACM dataset (Wang et al. 2019), and preprocessed DBLP and IMDB datasets. For ACM and DBLP datasets, the task is to classify the papers into three classes (Database, Wireless Communication, Data Mining), and four classes (DM, AI, CV, NLP), respectively, according to the research topic. For IMDB dataset, the task is to classify the movies into three classes (Action, Comedy, Drama). We note that the above datasets used by previous work are not truly multiplex in nature because the multiplexity between nodes is inferred via intermediate nodes (e.g., ACM: Paper-Paper relationships are inferred via Authors and Subjects that connect two Papers, i.e., “PAP” and “PSP”). Thus, to make our evaluation more practical, we used Amazon dataset (He and McAuley 2016) that genuinely contains a multiplex network of items, i.e., also-viewed, also-bought, and bought-together relations between items. We used datasets from four categories, i.e., Beauty, Automotive, Patio Lawn and Garden, and Baby, and the task is to classify items into the four classes. For ACM and IMDB datasets, we used the same number of labeled data as in (Wang et al. 2019) for fair comparisons, and for the remaining datasets, we used 20 labeled data for each class. Table 1 summarizes the data statistics.

**Methods Compared.**

1. **Embedding methods for a single network**
   - No attributes: Deepwalk (Perozzi, Al-Rfou, and Skiena 2014), node2vec (Grover and Leskovec 2016): They learn node embeddings by random walks and skip-gram.
   - Attributed network embedding: GCN (Kipf and Welling 2016), GAT (Veličković et al. 2017): They learn node embeddings based on local neighborhood structures. As they perform similarly, we report the best performing method among them; DGI (Veličković et al. 2019): It maximizes the MI between the graph-level summary representation and the local patches; ANRL (Zhang et al. 2018b): It uses neighbor enhancement autoencoder to model the node attribute information, and skip-gram model to capture the network structure; CAN (Meng et al. 2019): It learns embeddings of both attributes and nodes in the same semantic

**Table 1: Statistics of the datasets. The node attributes are bag-of-words of text associated with each node.**

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>Num. Relations</th>
<th>Num. Node Attributes</th>
<th>Num. Labeled Data</th>
<th>Num. Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM Paper-Author</td>
<td>3,025</td>
<td>1,726</td>
<td>144,783</td>
<td>3</td>
</tr>
<tr>
<td>ACM Paper-Subject</td>
<td>3,025</td>
<td>1,726</td>
<td>144,783</td>
<td>3</td>
</tr>
<tr>
<td>IMDB Movie-Actor</td>
<td>3,500</td>
<td>1,726</td>
<td>144,783</td>
<td>3</td>
</tr>
<tr>
<td>DBLP Paper-Paper</td>
<td>7,907</td>
<td>1,726</td>
<td>144,783</td>
<td>3</td>
</tr>
<tr>
<td>DBLP Author-Term</td>
<td>1,960</td>
<td>1,726</td>
<td>144,783</td>
<td>3</td>
</tr>
<tr>
<td>Amazon Item-Item</td>
<td>7,621</td>
<td>1,726</td>
<td>144,783</td>
<td>3</td>
</tr>
</tbody>
</table>

**The Academic Citation Network.** We used the same number of labeled data as in the original paper (Wang et al. 2019), i.e., ACM, DBLP, and IMDB. We used publicly available ACM dataset (Wang et al. 2019), and preprocessed DBLP and IMDB datasets. For ACM and DBLP datasets, the task is to classify the papers into three classes (Database, Wireless Communication, Data Mining), and four classes (DM, AI, CV, NLP), respectively, according to the research topic. For IMDB dataset, the task is to classify the movies into three classes (Action, Comedy, Drama). We note that the above datasets used by previous work are not truly multiplex in nature because the multiplexity between nodes is inferred via intermediate nodes (e.g., ACM: Paper-Paper relationships are inferred via Authors and Subjects that connect two Papers, i.e., “PAP” and “PSP”). Thus, to make our evaluation more practical, we used Amazon dataset (He and McAuley 2016) that genuinely contains a multiplex network of items, i.e., also-viewed, also-bought, and bought-together relations between items. We used datasets from four categories, i.e., Beauty, Automotive, Patio Lawn and Garden, and Baby, and the task is to classify items into the four classes. For ACM and IMDB datasets, we used the same number of labeled data as in (Wang et al. 2019) for fair comparisons, and for the remaining datasets, we used 20 labeled data for each class. Table 1 summarizes the data statistics.

**Methods Compared.**

1. **Embedding methods for a single network**
   - No attributes: Deepwalk (Perozzi, Al-Rfou, and Skiena 2014), node2vec (Grover and Leskovec 2016): They learn node embeddings by random walks and skip-gram.
   - Attributed network embedding: GCN (Kipf and Welling 2016), GAT (Veličković et al. 2017): They learn node embeddings based on local neighborhood structures. As they perform similarly, we report the best performing method among them; DGI (Veličković et al. 2019): It maximizes the MI between the graph-level summary representation and the local patches; ANRL (Zhang et al. 2018b): It uses neighbor enhancement autoencoder to model the node attribute information, and skip-gram model to capture the network structure; CAN (Meng et al. 2019): It learns embeddings of both attributes and nodes in the same semantic

**Table 1: Statistics of the datasets. The node attributes are bag-of-words of text associated with each node.**

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>Num. Relations</th>
<th>Num. Node Attributes</th>
<th>Num. Labeled Data</th>
<th>Num. Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM Paper-Author</td>
<td>3,025</td>
<td>1,726</td>
<td>144,783</td>
<td>3</td>
</tr>
<tr>
<td>ACM Paper-Subject</td>
<td>3,025</td>
<td>1,726</td>
<td>144,783</td>
<td>3</td>
</tr>
<tr>
<td>IMDB Movie-Actor</td>
<td>3,500</td>
<td>1,726</td>
<td>144,783</td>
<td>3</td>
</tr>
<tr>
<td>DBLP Paper-Paper</td>
<td>7,907</td>
<td>1,726</td>
<td>144,783</td>
<td>3</td>
</tr>
<tr>
<td>DBLP Author-Term</td>
<td>1,960</td>
<td>1,726</td>
<td>144,783</td>
<td>3</td>
</tr>
<tr>
<td>Amazon Item-Item</td>
<td>7,621</td>
<td>1,726</td>
<td>144,783</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 2: Properties of the compared methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dw/n2v</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>GCN/GAT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DGI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ANRL</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CAN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DGCN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CMNA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MNE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>mGCN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HAN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DMGI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

2) Multiplex embedding methods

- **No attributes**: CMNA (Chu et al. 2019): It leverages the cross-network information to refine inter-vector for network alignment and intra-vector for other downstream tasks. We use the intra-vector for our evaluations; MNE (Zhang et al. 2018a): It jointly models multiple networks by introducing a common embedding, and an additional embedding for each relation type.

- **Attributed multiplex network embedding**: mGCN (Ma et al. 2019), HAN (Wang et al. 2019): They apply GCNs, and GATs on multiplex network considering the inter- and intra-network interactions. For fair comparisons, we initialized the initial node embeddings of mGCN by using the node attribute matrix, although the node attributes information is ignored in the original mGCN; DMGI with the attention mechanism (Eqn. 9).

For the sake of fair comparisons with DMGI, which considers the node attributes, we concatenated the raw attribute matrix \( X \) to the learned node embeddings \( Z \) of the methods that ignore the node attributes, i.e., Deepwalk, node2vec, CMNA, and MNE. i.e., \( Z \leftarrow [Z; X] \). Moreover, regarding the embedding methods for a single network, i.e., the methods that belong to the first category in the above list, we obtain the final node embedding matrix \( Z \) by computing the average of the node embeddings obtained from each single graph. i.e., \( Z = \frac{1}{|R|} \sum_{r \in R} H^{(r)} \). We provide a summary of the properties of the compared methods in Table 2.

**Evaluation Metrics.** Recall that DMGI is an unsupervised method that does not require any labeled data for training. Therefore, we evaluate the performance of DMGI in terms of **node clustering** and **similarity search**, both of which are classical performance measures for unsupervised methods. For node clustering, we use the most commonly used metric (Wang et al. 2019), i.e., Normalized Mutual Information (NMI). For similarity search, we compute the cosine similarity scores of the node embeddings between all pairs of nodes, and for each node, we rank the nodes according to the similarity score. Then, we calculate the ratio of the nodes that belong to the same class within top-5 ranked nodes (Sim@5). Moreover, we also evaluate DMGI on the performance in terms of **node classification**. More precisely, after learning the node embeddings, we train a logistic regression classifier on the learned embeddings in the training set, and then evaluate on the nodes in the test set. We use Macro-F1 (MaF1) and Micro-F1 (MiF1) (Wang et al. 2019).

**Experimental Settings.** We randomly split our dataset into train/validation/test, and we have the equal number of labeled data for training and validation datasets. We report the test performance when the performance on validation data gives the best result. For DMGI, we set the node embedding dimension \( d = 64 \), self-connection weight \( w = 3 \), tune \( \alpha, \beta, \gamma \in \{0.0001, 0.001, 0.01, 0.1\} \). We implement DMGI in PyTorch\(^3\), and for all other methods, we used the source codes published by the authors, and tried to tune them to their best performance. More precisely, apart from the guidelines provided by the original papers, we tuned learning rate, and the coefficients for regularization from \( \{0.0001, 0.0005, 0.001, 0.005\} \) on the validation dataset. After learning the node embeddings, for fair comparisons, we conducted the evaluations within the same platform.

**4.1 Performance Analysis**

**Overall evaluation.** Table 3 and Table 4 show the evaluation results on unsupervised and supervised task, respectively. We have the following observations: 1) Our proposed DMGI and DMGI\(_{attn}\) outperform all the state-of-the-art baselines not only on the unsupervised tasks, but also the supervised task, although the improvement is more significant in the unsupervised task as expected. This verifies the benefit of our framework that models the multiplexity and the global property of a network together with the node attributes within a single framework. 2) Although DGI shows relatively good performance, the performance is unstable (poor performance on Amazon dataset), indicating that multiple relation types should be jointly modeled. 3) Attribute-aware multiplex network embedding methods.

Table 3: Performance for node clustering and similarity search on test data.

<table>
<thead>
<tr>
<th>Method</th>
<th>ACM</th>
<th>IMDB</th>
<th>DBLP</th>
<th>Amazon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NMI</td>
<td>Sim@5</td>
<td>NMI</td>
<td>Sim@5</td>
</tr>
<tr>
<td>Deepwalk</td>
<td>0.310</td>
<td>0.710</td>
<td>0.117</td>
<td>0.490</td>
</tr>
<tr>
<td>node2vec</td>
<td>0.309</td>
<td>0.710</td>
<td>0.123</td>
<td>0.487</td>
</tr>
<tr>
<td>GCN/GAT</td>
<td>0.671</td>
<td>0.867</td>
<td>0.176</td>
<td>0.565</td>
</tr>
<tr>
<td>DGI</td>
<td>0.640</td>
<td>0.889</td>
<td>0.182</td>
<td>0.578</td>
</tr>
<tr>
<td>ANRL</td>
<td>0.515</td>
<td>0.814</td>
<td>0.163</td>
<td>0.527</td>
</tr>
<tr>
<td>CAN</td>
<td>0.504</td>
<td>0.836</td>
<td>0.074</td>
<td>0.544</td>
</tr>
<tr>
<td>DGCN</td>
<td>0.691</td>
<td>0.690</td>
<td>0.143</td>
<td>0.179</td>
</tr>
<tr>
<td>CMNA</td>
<td>0.498</td>
<td>0.363</td>
<td>0.152</td>
<td>0.069</td>
</tr>
<tr>
<td>MNE</td>
<td>0.545</td>
<td>0.791</td>
<td>0.013</td>
<td>0.482</td>
</tr>
<tr>
<td>mGCN</td>
<td>0.668</td>
<td>0.873</td>
<td>0.183</td>
<td>0.550</td>
</tr>
<tr>
<td>HAN</td>
<td>0.658</td>
<td>0.872</td>
<td>0.164</td>
<td>0.561</td>
</tr>
<tr>
<td>DMGI</td>
<td>0.687</td>
<td>0.898</td>
<td>0.196</td>
<td>0.605</td>
</tr>
<tr>
<td>DMGI(_{attn})</td>
<td>0.702</td>
<td>0.901</td>
<td>0.185</td>
<td>0.586</td>
</tr>
</tbody>
</table>

\(^3\)https://github.com/pcy1302/DMGI
that methods (GCN/GAT, DGI, and ANRL). We observe

gather with the performance of single network embed-

distribution of the attention weights on DBLP dataset
IMDB dataset. To analyze the reason for this, we first plot

show the performance of

Table 5: Performance of similarity search (Sim@5) of em-
bedding methods for a single network. (Merged denotes the
average of all the relation-type specific embeddings.)

<table>
<thead>
<tr>
<th>Rel. Type</th>
<th>ACM</th>
<th>GIN</th>
<th>DGI</th>
<th>ANRL</th>
<th>DMGI</th>
<th>DMGI attn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merged</td>
<td>0.867</td>
<td>0.889</td>
<td>0.814</td>
<td>0.898</td>
<td>0.901</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rel. Type</th>
<th>IMDB</th>
<th>GIN</th>
<th>DGI</th>
<th>ANRL</th>
<th>DMGI</th>
<th>DMGI attn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merged</td>
<td>0.566</td>
<td>0.578</td>
<td>0.527</td>
<td>0.605</td>
<td>0.586</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rel. Type</th>
<th>DBLP</th>
<th>GIN</th>
<th>DGI</th>
<th>ANRL</th>
<th>DMGI</th>
<th>DMGI attn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merged</td>
<td>0.724</td>
<td>0.786</td>
<td>0.720</td>
<td>0.766</td>
<td>0.799</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rel. Type</th>
<th>Amazon</th>
<th>GIN</th>
<th>DGI</th>
<th>ANRL</th>
<th>DMGI</th>
<th>DMGI attn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merged</td>
<td>0.624</td>
<td>0.558</td>
<td>0.764</td>
<td>0.816</td>
<td>0.825</td>
<td></td>
</tr>
</tbody>
</table>

| Rel. Type | Merged | 0.817 | 0.887 | 0.602 | 0.606 | 0.778 | 0.770 | 0.758 | 0.758 |

over the training epochs in Figure 2. The above graph in
Figure 2 demonstrates that the attention weights eventually
end up in both extremes, i.e., close to 0 or close to 1, and
the below graphs show that most of the attention weight is
dedicated to a single relation type, i.e., “PAP”, which actu-
ally turns out to be the most important relation among the
three (See Table 5); This phenomenon is common in every
dataset. Next, we look at the performance of the single net-
work embedding methods, especially DGI, on each relation
type in Table 5. We observe that the performance differences
among relation types in ACM, DBLP, and Amazon datasets are
more biased to a single relation type, whereas in IMDB
dataset, “MAM” and “MDM” relations relatively show sim-
ilar performance. To summarize our findings, since the at-
tention mechanism tends to favor the single most important
relation type (“PAP” in ACM, “MDM” in IMDB, “PAP” in
DBLP, and “Bought-together” in Amazon), DMGI attn out-
performs DMGI on datasets where one relation type signif-
cantly outperforms the other, i.e., ACM, DBLP, and Amazon,
by removing the noise from other relations. On the other
hand, for datasets where all the relations show relatively
even performance, i.e., IMDB, extremely favoring a single
well performing relation type (“MDM”) is rather detrimen-
tal to the overall performance because the relation “MAM”
should also be considered to some extent.

We also note that since the attention mechanism of DMGI attn can infer the importance of each relation type, we can filter out unnecessary relation types as a prepro-
cessing step. To verify this, we evaluated on all possible
combinations of relation types in DBLP dataset (Ta-
ble 6). We observe that by removing the relation “PATAP”,
which turned out to be the most useless relation type in Ta-
ble 5, DMGI attn obtains even better results than using all the
relation types, whereas for GCN and DGI, still considering
all the relation types shows the best performance. This indi-
cates that the attention mechanism can be useful to filter
out unnecessary relation types, which will especially come
in handy when the number of relation types is large.

Ablation study. To measure the impact of each com-
ponent of DMGI attn, we conduct ablation studies on the largest
dataset, i.e., DBLP, in Table 7. We have the following obser-
vations: 1) As expected, the semi-supervised module specif-
Table 6: NMI on various combinations of relation types.

<table>
<thead>
<tr>
<th>DBLP dataset</th>
<th>GCN/GAT</th>
<th>DGI</th>
<th>DMGI _attn</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAP+PPP</td>
<td>0.464</td>
<td>0.543</td>
<td><strong>0.565</strong></td>
</tr>
<tr>
<td>PAP+PAP–PAP</td>
<td>0.458</td>
<td>0.535</td>
<td>0.017</td>
</tr>
<tr>
<td>PPP–PAP+PAP</td>
<td>0.332</td>
<td>0.237</td>
<td>0.201</td>
</tr>
<tr>
<td>All</td>
<td>0.465</td>
<td>0.551</td>
<td>0.554</td>
</tr>
</tbody>
</table>

Table 7: Result for ablation studies of DMGI \_attn.

<table>
<thead>
<tr>
<th>DBLP dataset</th>
<th>MaF1</th>
<th>NMI</th>
<th>Sim@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMGI _attn</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1) DMGI _attn + Semi supervised</td>
<td>0.778</td>
<td>0.554</td>
<td>0.798</td>
</tr>
<tr>
<td>2) Readout _attn (Eqn. 3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random sample</td>
<td>0.774</td>
<td>0.555</td>
<td>0.797</td>
</tr>
<tr>
<td>Maxpool</td>
<td>0.778</td>
<td>0.552</td>
<td>0.802</td>
</tr>
<tr>
<td>Linear projection</td>
<td>0.783</td>
<td>0.565</td>
<td>0.803</td>
</tr>
<tr>
<td>SAGPool</td>
<td>0.797</td>
<td>0.563</td>
<td>0.797</td>
</tr>
<tr>
<td>3) Without 2nd term of Eqn. 6</td>
<td>0.749</td>
<td>0.448</td>
<td>0.787</td>
</tr>
<tr>
<td>4) M _\neq M^{(1)} _\neq \ldots _\neq M^{(K)}</td>
<td>0.645</td>
<td>0.076</td>
<td>0.677</td>
</tr>
<tr>
<td>5) No attributes (Adj. as attribute)</td>
<td>0.377</td>
<td>0.053</td>
<td>0.763</td>
</tr>
<tr>
<td>6) Neg sample: Shuffle adj.</td>
<td>0.364</td>
<td>0.156</td>
<td>0.504</td>
</tr>
</tbody>
</table>

Table 5: Result for ablation studies of DMGI \_attn.

5 Related Work

Network embedding. Network embedding methods aim at learning low-dimensional vector representations for nodes in a graph while preserving the network structure (Perozzi, Al-Rfou, and Skiena 2014; Grover and Leskovec 2016; Tang et al. 2015), and various other properties such as node attributes (Zhang et al. 2018b; Meng et al. 2019), structural importance (Ribeiro, Saverese, and Figueiredo 2017), and node label information (Huang, Li, and Hu 2017).

Multiplex Network embedding. A multiplex network, which is also known as a multi-view network (Tang et al. 2015; Shi et al. 2018) or a multi-dimensional network (Ma et al. 2018; 2019) in the literature, consists of multiple relation types among a set of single-typed nodes. It can be thought of as a special type of heterogeneous network (Dong, Chawla, and Swami 2017; Fu, Lee, and Lei 2017) with a single type of node and multiple types of edges. Therefore, a multiplex network calls for a special attention because there is no need to consider the semantics between different types of nodes, which is often addressed by the concept of metapath (Sun et al. 2011). Distinguished from heterogeneous network, a key challenge in the multiplex network embedding is to learn a consensus embedding for each node by taking into account the interrelationship among the multiple graphs. In this regard, existing methods mainly focused on how to integrate the information from multiple graphs. HAN (Wang et al. 2019) employed graph attention network (Veličković et al. 2017) on each graph, and then applied the attention mechanism to merge the node representations learned from each graph by considering the importance of each graph. However, the existing methods either require labels for training (Wang et al. 2019; Qu et al. 2017; Schlichtkrull et al. 2018), or overlook the node attributes (Liu et al. 2017; Xu et al. 2017; Li et al. 2018; Shi et al. 2018; Zhang et al. 2018a; Ni et al. 2018; Chu et al. 2019). Most recently, Ma et al. (2019) proposed a graph convolutional network (GCN) based method called mGCN, which is not only unsupervised, but also naturally incorporates the node attributes by using GCNs. However, since it is based on GCNs that capture the local graph structure (Yadav et al. 2019), it fails to fully model the global properties of a graph (Zhuang and Ma 2018; Wang, Cui, and Zhu 2016; Veličković et al. 2019).

Attributed Network Embedding. Nodes in a network are often affiliated with various contents, such as abstract text in the publication network, user profiles in social networks, and item description text in movie database or item networks. Such networks are called attributed networks, and have been extensively studied (Li et al. 2017; Hamilton, Ying, and Leskovec 2017; Yang et al. 2015; Zhang et al. 2018b; Gao and Huang 2018; Zhou et al. 2018; Veličković et al. 2019; Meng et al. 2019). Their goal is to preserve not only the network structure, but also the node attribute proximity in learning representations. Recently, GCNs (Kipf and Welling 2016; Veličković et al. 2017; 2019) have been widely praised for its seamless integration of the network structure, and node attributes into a single framework.

Mutual Information. It has been recently made possible to compute the MI between high dimensional input/output pairs of deep neural networks (Belghazi et al. 2018). Several recent work adopted the infomax principle (Linsker 1988) to learn the unsupervised representations in different domains, such as images (Hjelm et al. 2019), speech (Ravanelli and Bengio 2018) and graphs (Veličković et al. 2019). More precisely, Veličković et al. (2019) proposed Deep Graph Infomax (DGI) for learning representations of graph structured inputs by maximizing the MI between a high-level global representation, and the local patches of a graph.

6 Conclusion

We presented a simple yet effective unsupervised method for embedding attributed multiplex network. DMGI can jointly integrate the embeddings from multiple types of relations between nodes through the consensus regularization framework, and the universal discriminator. Moreover, the attention mechanism of DMGI can infer the importance of each relation type, which facilitates the preprocessing of the multiplex network. Experimental results on not only unsupervised tasks, but also a supervised task verify the superiority of our proposed framework.

References


Dong, Y.; Chawla, N. V.; and Swami, A. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In KDD. ACM.


Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In NIPS.

He, R., and McAuley, J. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In WWW.


Huang, X.; Li, J.; and Hu, X. 2017. Label informed attributed network embedding. In WSDM. ACM.


Lee, J.; Lee, I.; and Kang, J. 2019. Self-attention graph pooling. ICLR.


Li, J.; Chen, C.; Tong, H.; and Liu, H. 2018. Multi-layered network embedding. In SDM. SIAM.


Ma, Y.; Ren, Z.; Jiang, Z.; Tang, J.; and Yin, D. 2018. Multi-dimensional network embedding with hierarchical structure. In WSDM. ACM.


Qu, M.; Tang, J.; Shang, J.; Ren, X.; Zhang, M.; and Han, J. 2017. An attention-based collaboration framework for multi-view network representation learning. In CIKM. ACM.


Sun, Y.; Han, J.; Yan, X.; Yu, P. S.; and Wu, T. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. VLDB.


Velicković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. ICLR.


Vinyals, O.; Bengio, S.; and Kudlur, M. 2015. Order matters: Sequence to sequence for sets. NIPS.

Wang, X.; Cui, P.; Wang, J.; Pei, J.; Zhu, W.; and Yang, S. 2017. Community preserving network embedding. In AAAI.


Wang, D.; Cui, P.; and Zhu, W. 2016. Structural deep network embedding. In KDD. ACM.

Xu, L.; Wei, X.; Cao, J.; and Philip, S. Y. 2017. Multi-task network embedding. In DSAA. IEEE.


Order matters: Sequence to sequence for sets. NIPS.

Wang, X.; Cui, P.; Wang, J.; Pei, J.; Zhu, W.; and Yang, S. 2017. Community preserving network embedding. In AAAI.


Wang, D.; Cui, P.; and Zhu, W. 2016. Structural deep network embedding. In KDD. ACM.

Xu, L.; Wei, X.; Cao, J.; and Philip, S. Y. 2017. Multi-task network embedding. In DSAA. IEEE.

