Network Representation Learning: From Preprocessing, Feature Extraction to Node Embedding

JINGYA ZHOU, Soochow University, China Georgia Institute of Technology, USA State Key Laboratory of Mathematical Engineering and Advanced Computing
LING LIU and WENQI WEI, Georgia Institute of Technology
JIANXI FAN, Soochow University

Network representation learning (NRL) advances the conventional graph mining of social networks, knowledge graphs, and complex biomedical and physics information networks. Dozens of NRL algorithms have been reported in the literature. Most of them focus on learning node embeddings for homogeneous networks, but they differ in the specific encoding schemes and specific types of node semantics captured and used for learning node embedding. This article reviews the design principles and the different node embedding techniques for NRL over homogeneous networks. To facilitate the comparison of different node embedding algorithms, we introduce a unified reference framework to divide and generalize the node embedding learning process on a given network into preprocessing steps, node feature extraction steps, and node embedding model training for an NRL task such as link prediction and node clustering. With this unifying reference framework, we highlight the representative methods, models, and techniques used at different stages of the node embedding model learning process. This survey not only helps researchers and practitioners gain an in-depth understanding of different NRL techniques but also provides practical guidelines for designing and developing the next generation of NRL algorithms and systems.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Information systems** → **Data mining**;

Additional Key Words and Phrases: Network representation learning, data preprocessing, feature extraction, node embedding

J. Zhou conducted this work during his 1-year visit at Georgia Institute of Technology under the Jiangsu Overseas Visiting Scholar Program for University Prominent Young & Middle-Aged Teachers and Presidents. J. Zhou and J. Fan were supported by the National Natural Science Foundation of China under grants 61972272, 62172291, 62072321, and U1905211, the Open Project Program of the State Key Laboratory of Mathematical Engineering and Advanced Computing under grant 2019A04, and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under grant 21KJA520008. L. Liu and W. Wei were partially supported by the National Science Foundation under grants 2038029 and 1564097 and an IBM faculty award.

Authors' addresses: J. Zhou, Soochow University, 1 Shizi Street, Suzhou, Jiangsu, 215006, China, Georgia Institute of Technology, Atlanta, GA, State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi, Jiangsu, 214125, China; email: jy_zhou@suda.edu.cn; L. Liu and W. Wei, Georgia Institute of Technology, 801 Atlantic Drive, Atlanta, GA 30332; emails: lingliu@cc.gatech.edu, wenqiwei@gatech.edu; J. Fan, Soochow University, 1 Shizi Street, Suzhou, Jiangsu, China, 215006; email: jxfan@suda.edu.cn.

0360-0300/2022/01-ART38 $15.00
https://doi.org/10.1145/3491206

ACM Reference format:
Jingya Zhou, Ling Liu, Wenqi Wei, and Jianxi Fan. 2022. Network Representation Learning: From Preprocessing, Feature Extraction to Node Embedding. *ACM Comput. Surv.* 55, 2, Article 38 (January 2022), 35 pages.
https://doi.org/10.1145/3491206

1 INTRODUCTION

Recent advances in deep learning and the convolutional neural network (CNN) [52] have made remarkable breakthroughs in many fields, such as machine translation [94] and reading comprehension in natural language processing (NLP) [99], object detection [79] and image classification [59] in computer vision (CV). In addition to text, audio, image, and video data, information networks (or graphs) represent another type of natural and complex data structure representing a set of entities and their relationships. A wide variety of real-world data in business, science, and engineering domains are best captured as information networks, such as protein interaction networks, citation networks, and social media networks like Facebook and LinkedIn, to name a few.

Network representation learning (NRL), also known as network embedding, is to train a neural network to represent an information network as a collection of node-embedding vectors in a latent space such that the desired network features are preserved, which enables the well-trained NRL model to perform network analytics, such as link prediction or node cluster, as shown in Figure 1. The goal of NRL is to employ deep learning algorithms to encode useful network information into the latent semantic representations, which can be deployed for performing popular network analytics, such as node classification, link prediction, community detection, and domain-specific network mining, such as social recommendation [28, 91], protein to protein interaction prediction [30], disease-gene association identification [43], automatic molecule optimization [31], and Bitcoin transaction forecasting [92].

Different from traditional feature engineering that relies heavily on handcrafted statistics to extract structural information, NRL introduces a new data-driven deep learning paradigm to capture, encode, and embed structural features along with non-structural features into a latent space represented by dense and continuous vectors. By embedding edge semantics into node vectors, a variety of network operations can be carried out efficiently, such as computing the similarity between a pair of nodes, visualizing a network in a two-dimensional space. Moreover, parallel processing on large-scale networks can be naturally supported with the node embedding learned from NRL.

Most of the existing NRL efforts are targeted on learning node embeddings of a homogeneous network, in which all nodes are homogeneous and all edges belong to a single type of node relationships—for example, a social network is considered homogeneous when we only consider users and their friendship relationships [66]. A heterogeneous information network consists of nodes and edges of heterogeneous types, corresponding to different types of entities and different kinds of relations respectively. Knowledge graphs [38, 47] and RDF graphs [101] are known examples of heterogeneous information networks.

DeepWalk [66] is the first node embedding algorithm that learns to encode the neighborhood features of each node in a homogeneous graph through learning the encoding of its scoped random walk properties using the autoencoder algorithms in conjunction with node2vec [36]. Inspired by DeepWalk design, dozens of node embedding algorithms have been proposed [8, 14, 15, 17, 23, 25, 27, 36, 37, 41, 51, 55, 57, 67, 69–72, 77, 81, 83, 87, 88, 93, 97, 102, 103]. Although most of them focus on learning node embeddings for homogeneous networks, they differ in terms of the specific encoding schemes and the specific types of node semantics captured and used for learning node embedding. This article mainly reviews the design principles and the different node embedding techniques developed for NRL over homogeneous networks. To facilitate the comparison of

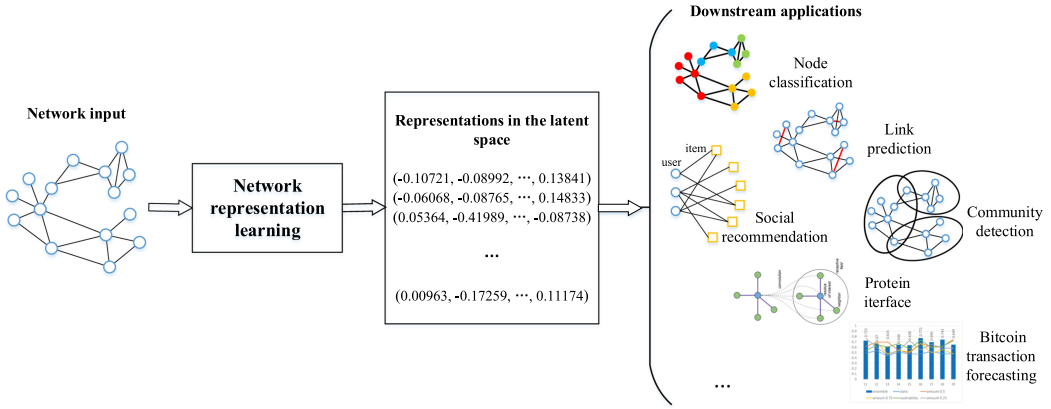

Fig. 1. An illustration of NRL and its deployment in example applications [28, 30, 31, 43, 91, 92].

different node embedding algorithms, we introduce a unified reference framework to divide and generalize the node embedding learning process on a given network into preprocessing steps, node feature extraction steps, and a node embedding model that can be used for link prediction and node clustering. With this unifying reference framework, we highlight the most representative methods, models, and techniques used at different stages of the node embedding model learning process.

We argue that an in-depth understanding of different node embedding methods/models/techniques is also essential for other types of NRL approaches that are built on top of node embedding techniques, such as edge embedding [2, 32], subgraph embedding [9, 16], and entire-graph embedding [5, 63]. For example, an edge can be represented by a Hadamard product of its two adjacent nodes' vectors. Similarly, graph coarsening mechanisms [19, 100] may create a hierarchy by successively clustering the nodes in the input graph into smaller graphs connected in a hierarchical manner, which can be used to generate representations for subgraphs and even for the entire graph.

We conjecture that this work not only helps researchers and practitioners gain an in-depth understanding of different NRL techniques but also provides practical guidelines for designing and developing the next generation of NRL algorithms and systems.

Current surveys [13, 20, 40] primarily focus on presenting a taxonomy to review the existing work on NRL. Concretely, Cai et al. [13] propose two taxonomies of graph embedding based on problem settings and techniques, respectively, and their work first appeared in 2017 on ArXiv and was published in 2018. Cui et al. [20] propose a taxonomy of network embedding according to the types of information preserved. Hamilton et al. [40] appeared in 2017 in the *IEEE Data Engineering Bulletin*. Their work describes a set of conventional node embedding methods with the focus on pairwise proximity methods and neighborhood aggregation based methods. In contrast, our unified reference framework provides a broader and more comprehensive comparative review of the state of the art in NRL. In our three-stage reference framework, each stage serves as a categorization of the set of technical solutions dedicated to the tasks respective to this stage. For example, we not only provide a review of node embedding models using the unified framework but also describe a set of optimization techniques that are commonly used in different node embedding methods, and also an overview of recent advances in NRL.

We list the mathematical notations used throughout the article in Table 1 for reference convenience.

The rest of this survey is structured as follows. In Section 2, we describe the basic steps of NRL to generate node embeddings using the autoencoder approach. In Section 3, we present an

Table 1. List of Notations and Symbols Used in the Article and Their Meaning

Notation/Symbol	Meaning	Notation/Symbol	Meaning
$\mathbf{W}, \mathbf{U}, \mathbf{M}$	Bold capital letters represent matrices	$\mathbf{h}, \mathbf{u}_k, \mathbf{v_I}$	Bold lowercase letters represent vectors
$\mathrm{u}_k, \mathrm{v_I}, \mathrm{v}_i, \mathrm{v}_j$	Lowercase letters represent nodes	v_i, h_i, y_i	Italic lowercase letters represent vector elements
$\Phi(\cdot), f_{\theta_1}(\cdot)$	Mapping function, i.e., encoder	$\Psi(\cdot), g_{\theta_2}(\cdot)$	Decoder
$\mathbf{A}$	Adjacency matrix	$\mathbf{L}$	Laplacian matrix
N_u	Node u's neighborhood	G_p	Persona graph
$G[N_\mathrm{u}]$	Ego-network of node u	$\mathbf{P^k}, \Gamma^\mathbf{k}$	k-Step transition matrix
$\mathbf{X}$	Additional information matrix	$\hat{\Psi}(\mathbf{v_i}, \mathbf{v_j})$	Empirical probability
$\mathbf{g}_\theta$	Filter of a convolution operation	$T_k(\mathbf{x})$	Chebyshev polynomials
Θ	Matrix of filter parameters	AGG_k	Aggregation function in the kth layer
V_{neg}	Set of negative samples	η	Learning rate
L_{1st}, L_{2nd}, L	First loss, second loss, loss	$\mathbf{b}$	Bias vector
$\mathbf{x}^{(i)}, \tilde{\mathbf{x}}^{(i)}$	ith instance and its corrupted form	$(\mathbf{H^{(i)}})_\mathbf{l}$	Matrix of representations in the lth layer of network i
$\mathbf{S^{RPR}}$	Rooted PageRank matrix	$\bar{t}_p$	Timestamp before the current event
$\mathbf{h}, \mathbf{r}, \mathbf{t}$	Representations of head, translation, and tail	$d_r(\mathbf{h}, \mathbf{t})$	Distance function of a triple

overview of the unifying three-stage reference framework for NRL, and discuss representative methods, models, and optimization techniques used at each stage. In Section 4, we review recent advances in conventional NRL, distributed NRL, multi-NRL, dynamic NRL, and **knowledge graph representation learning (KRL)** by using the proposed reference framework and discuss several open challenges. In Section 5, we conclude our survey.

2 NRL: WHAT AND HOW

To establish a common ground for introducing design principles of NRL methods and techniques, we first provide a walkthrough example to illustrate how NRL works from the autoencoder perspective. We first briefly describe DeepWalk [66], as it will be used as the reference NRL model in this section.

DeepWalk [66] is generalized from the advancements in language modeling (e.g., Word2vec [61]). In a language model, the corpus is built by collecting sentences from many documents. If we regard node traveling paths as sentences, we can build corpus for NRL. Given an input network, we use the random walk method with parameters γ, t, and α to generate multiple node traveling paths, where γ decides how many times to issue random walks from a node, t is the path length, and α is the probability of stopping walk and restarting from the initial node.

In the language processing field, skip-gram and continuous bag-of-words (CBOW) [61] are two commonly used models for estimating the likelihood of co-occurrence among words in the training set. For NRL, training instances are extracted from node traveling paths based on a sliding window. An instance consists of a target node and its context located within a fixed window size, such as (3, (1,7)). In the meantime, we also need to build a vocabulary to index and sort all nodes by their frequency in the corpus, and then build a Huffman tree based on the frequency for hierarchical softmax.

The learning model shown in Figure 2 contains three layers, and it belongs to a typical autoencoder paradigm. The input vectors are encoded into latent representation vectors by means of a matrix $\mathbf{W}$, which is known as encoding, and then the latent representation vectors are reconstructed into output vectors by means of a matrix $\mathbf{U}$, which is known as decoding. Given a training instance, the skip-gram model is mainly used to forecast the context given a target node, whereas the continuous bag-of-words model predicts the target node given its context. The skip-gram model is widely adopted in NRL, since the conditional probability can be decomposed into multiple simple conditional probabilities under independence assumption,

$$\Pr(context(\mathrm{v_I})|\Phi(\mathrm{v_I})) = \prod_{\mathrm{v}_j \in context(\mathrm{v_I})} \Pr(\mathrm{v}_j|\Phi(\mathrm{v_I})), \tag{1}$$

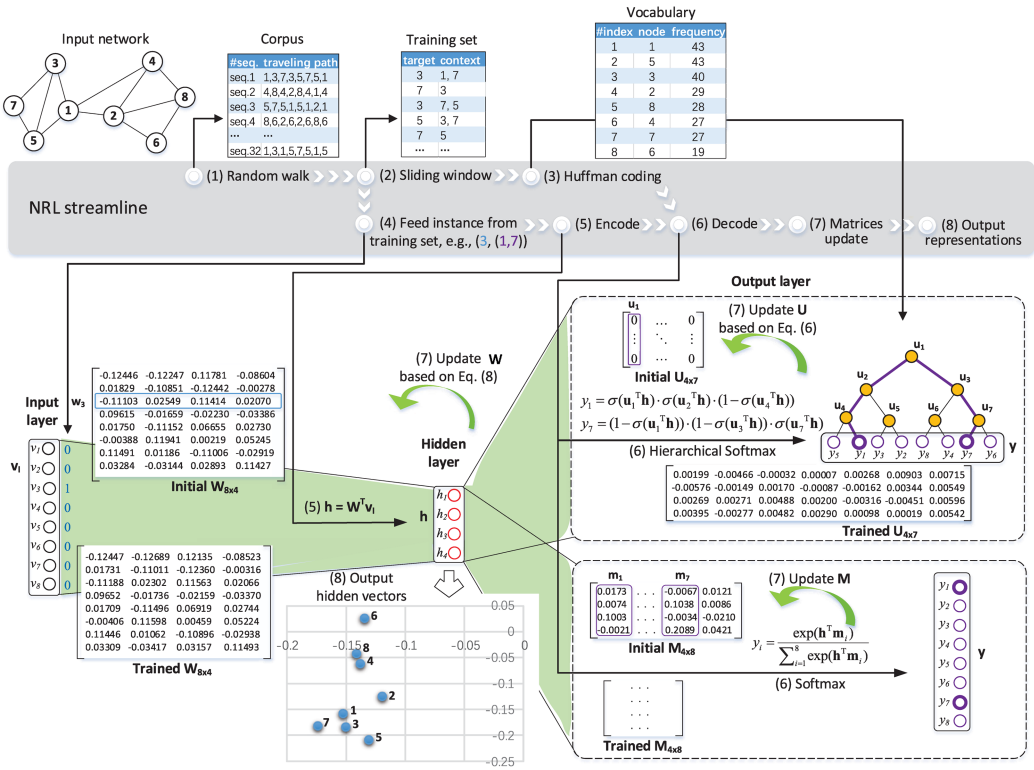

Fig. 2. An example of a three-layer neural network used to realize NRL. (1) Build the corpus based on random walk ($\gamma = 4, t = 8, \alpha = 0$). (2) Build the training set based on a sliding window ($s = 1$). (3) Build the vocabulary to store the set of nodes and their frequency in corpus, and build a Huffman tree; note that this step is optional and only necessary for hierarchical softmax. (4) Feed the training instance to the input layer. (5) Encode the input vector into the hidden vector. (6) Decode the hidden vector into conditional probabilities of context nodes based on softmax or hierarchical softmax. (7) Update matrices $\mathbf{W}$ and $\mathbf{M}$ or $\mathbf{U}$ based on the back-propagation method ($\eta = 0.025$). (8) Output the hidden vectors as the learned representations.

where $\Phi(\cdot)$ is a mapping function that embeds nodes into the low-dimensional vector space, i.e., $\Phi(\mathrm{v_I})$ refers to the target node $\mathrm{v_I}$'s learned representation. Here, $\Phi(\cdot)$ acts as the encoder from the autoencoder perspective. Given the embedding of the target node, the conditional probability acts as the decoder that captures the reconstruction of the target node and its context nodes in the original network. A loss function is defined accordingly to measure the reconstruction error, and the learning model is trained by minimizing the loss. To be specific, for each training instance, the target node $\mathrm{v_I}$ is initially encoded to be a one-hot vector $\mathbf{v_I}$ and its dimension equals vocabulary size n, and $\mathbf{W}$ is a $n \times m$ matrix initialized by letting its entries randomly falling in a range $[-1/2n, 1/2n]$. One-hot encoding implies that $\mathbf{h}$ is a m-dimensional vector simply copying a row of $\mathbf{W}$ associated with the target node $\mathrm{v_I}$. A $m \times n$ matrix $\mathbf{M}$ is set to decode the encoded vector $\mathbf{v_I}$, where the conditional probability is obtained by doing softmax, i.e., $y_i = \frac{\exp(\mathbf{h}^\mathrm{T}\mathbf{m}_i)}{\sum_{i=1}^{n}\exp(\mathbf{h}^\mathrm{T}\mathbf{m}_i)}$. But softmax is not scalable, because for each training instance, softmax requires to repeat vector multiplication for n times to obtain the denominator.

To improve computation efficiency of the decoder, DeepWalk uses hierarchical softmax [60, 62] instead of softmax to implement the conditional probability factorization. The hierarchical softmax

model builds a binary tree and places all network nodes on the leaf layer. Then there will be $n-1$ branch nodes, and each of them has an associated m-dimensional vector. For the output node $\mathrm{v_j}$ in a training instance, it corresponds to a leaf node y_j in the tree representing the probability of $\mathrm{v_j}$ in the output layer given target node $\mathrm{v_I}$. It is easy to identify a unique path from the root node to node y_i, and the conditional probability can be computed based on the path, i.e.,

$$y_j = \Pr(\mathrm{v}_j|\Phi(\mathrm{v_I})) = \prod_{k=1}^{l_j-1} \Pr(\mathbf{u}_k|\Phi(\mathrm{v_I})), \tag{2}$$

where l_j is the length of the path toward y_j. DeepWalk uses the Huffman tree to implement hierarchical softmax due to its optimal property on average path length. On the path toward leaf node y_j, a binary classifier is used to compute the probability of going left or right at each branch node, i.e.,

$$\Pr(\mathbf{u}_k|\Phi(\mathrm{v_I})) = \begin{cases} \sigma({\mathbf{u}_k}^{\mathrm{T}} \cdot \mathbf{h}), \text{go left} \\ 1-\sigma({\mathbf{u}_k}^{\mathrm{T}} \cdot \mathbf{h}), \text{go right} \end{cases}, \text{where } \sigma({\mathbf{u}_k}^{\mathrm{T}} \cdot \mathbf{h}) = \frac{1}{1+\exp(-{\mathbf{u}_k}^{\mathrm{T}} \cdot \mathbf{h})}. \tag{3}$$

The model's goal is to obtain the maximized conditional probability, which is equivalent to minimize the following loss function

$$L = -\log \Pr(context(\mathrm{v_I})|\Phi(\mathrm{v_I})) = \sum_{\mathrm{v}_j \in context(\mathrm{v_I})} -\log y_j. \tag{4}$$

To this end, the back-propagation method is used to update two weight matrices $\mathbf{W}$ and $\mathbf{U}$ with gradient descent. First, we take the derivative of loss with regard to each $\mathbf{u_k}$ on the path toward the context node and obtain

$$\frac{\partial L}{\partial \mathbf{u}_k} = \frac{\partial L}{\partial {\mathbf{u}_k}^{\mathrm{T}} \cdot \mathbf{h}} \cdot \frac{\partial {\mathbf{u}_k}^{\mathrm{T}} \cdot \mathbf{h}}{\partial \mathbf{u}_k} = \sum_{\mathrm{v}_j \in context(\mathrm{v_I})} \left(\sigma\left({\mathbf{u}_k}^{\mathrm{T}} \cdot \mathbf{h}\right) - \pi_k\right) \cdot \mathbf{h}, \tag{5}$$

where $\pi_k = 1$ if going left and $\pi_k = 0$ otherwise. The corresponding vectors in matrix $\mathbf{U}$ are updated by

$$\mathbf{u}_k \leftarrow \mathbf{u}_k - \eta \sum_{\mathrm{v}_j \in context(\mathrm{v_I})} \left(\sigma\left({\mathbf{u}_k}^{\mathrm{T}} \cdot \mathbf{h}\right) - \pi_k\right) \cdot \mathbf{h}, \text{ where } \eta \text{ is the learning rate.} \tag{6}$$

Then for each context node in an instance, we take the derivative of loss with regard to the hidden vector $\mathbf{h}$ and obtain

$$\frac{\partial L}{\partial \mathbf{h}} = \sum_{\mathrm{v}_j \in context(\mathrm{v_I})} \sum_{k=1}^{l_j-1} \frac{\partial L}{\partial {\mathbf{u}_k}^{\mathrm{T}} \cdot \mathbf{h}} \cdot \frac{\partial {\mathbf{u}_k}^{\mathrm{T}} \cdot \mathbf{h}}{\partial \mathbf{h}} = \sum_{\mathrm{v}_j \in context(\mathrm{v_I})} \sum_{k=1}^{l_j-1} \left(\sigma\left({\mathbf{u}_k}^{\mathrm{T}} \cdot \mathbf{h}\right) - \pi_k\right) \cdot \mathbf{u}_k. \tag{7}$$

The vector $\mathbf{w_I}$ in matrix $\mathbf{W}$ is updated accordingly by

$$\mathbf{w}_{\mathrm{I}} \leftarrow \mathbf{w}_{\mathrm{I}} - \eta \sum_{\mathrm{v}_j \in context(\mathrm{v_I})} \sum_{k=1}^{l_j-1} \left(\sigma\left({\mathbf{u}_k}^{\mathrm{T}} \cdot \mathbf{h}\right) - \pi_k\right) \cdot \mathbf{u}_k. \tag{8}$$

The model will learn the latent representation for every node by updating matrices iteratively and eventually stabilize. The hidden vectors are the node representations learned from the network.

Figure 2 shows a simple network with 8 nodes, and we build a corpus by running random walk four times for each node and obtain 32 node sequences. We also generate multiple instances for training by means of a sliding window with $s = 1$, which means a target node may have 2 context nodes at most. Instance (3, (1,7)) is the first one used for the following training, and the target node's input vector is $\mathrm{v_3}$'s one-hot vector (0,0,1,0,0,0,0,0). After encoded by weight matrix $\mathbf{W}$, we

obtain v_3's hidden vector that is $\mathbf{w_3}$ here. To obtain the conditional probability, we build a Huffman tree in the output layer. The weight matrix $\mathbf{U}$ is a collection of all the branch nodes' vectors, and they are initialized to be zero vectors to make sure that the probabilities of going left and going right at each branch are initially identical, i.e., 1/2. Paths ($\mathbf{u_1}$, $\mathbf{u_2}$, $\mathbf{u_4}$) and ($\mathbf{u_1}$, $\mathbf{u_3}$, $\mathbf{u_7}$) are two unique paths toward leaf nodes y_1 and y_7, respectively, and then we have $y_1 = y_7 = 1/8$ based on Equation (2). In the following process, we need to update the correlated vectors in weight matrices reversely to minimize the loss, so we obtain $\mathbf{u_1} = \mathbf{0}$, $\mathbf{u_2} = \mathbf{u_7} = \eta\mathbf{h}/2$, $\mathbf{u_4} = \mathbf{u_3} = -\eta\mathbf{h}/2$ based on Equation (6), and then $\mathbf{w_3} = \mathbf{w_3}^{(\mathbf{old})} - \eta^2\mathbf{h}/2$ based on Equation (8). After we finish the training against all instances, the representation learned from the model is the hidden vector. Since the input layer uses one-hot encoding, the network representation is weight matrix $\mathbf{W}$. We plot these representations in a two-dimensional coordinate system for visualization, where community structure can be observed.

3 THE REFERENCE FRAMEWORK FOR NRL

We present a unified reference framework illustrated in Figure 3 to capture the workflow of NRL, and it contains three consecutive stages. The first stage is the network data preprocessing, which is responsible for obtaining the desired network structure information from the original input network. During this stage, the prime aim is to employ a learning-task suitable network preprocessing method to transform the input network into a set of internal data structures, which is more suitable for structural feature extraction in the next stage. Node state and edge state provide additional context information other than network topological structure, which are useful and can be leveraged for learning network representations. Different NRL algorithms tend to have different design choices on which additional information will be utilized to augment the node context information in addition to network topological structure. The second stage is the network feature extraction, which is responsible for sampling training instances from the input network structure. Prior to sampling, it should choose the source of raw features that helps preserve the expected network properties. These properties may optionally be inferred from a specific learning task. The source of raw features can be classified into local structure (e.g., node degree and node neighbors) and global structure (e.g., multi-hop neighborhoods and node rank) with respect to every node in the raw input graph. Different sampling methods are used for extracting features from different structures. The third stage is the learning node embeddings over the training set. Different embedding models can be leveraged to learn hidden features for node embedding, such as matrix factorization, the probabilistic model, and **graph neural networks (GNNs)**. These representative embedding models are often coupled with optimization techniques, such as hierarchical softmax, negative sampling, and the attention mechanism, for better embedding effects.

3.1 Network Data Preprocessing

The network processing method is the data preparation stage for NRL. When end users have different applications in mind for deploying NRL models, different data preprocessing methods should be employed. Hence, specific learning tasks should be discussed in the section on network data preprocessing. For example, when NRL trained models are used for node classification, if the node classification or node clustering aims to categorize new nodes based on their node state information and node edge neighbor information, then the preprocessing stage should employ techniques that can preprocess the raw graph input to obtain those required node state properties and node linkage properties for deep feature extraction (the second stage) before entering the NRL model training, the third stage of the NRL workflow. However, if the end users prefer to perform node clustering or classification based on only network topology and traversal patterns over the network structure rather than node state information, then the pairwise node relationships over the

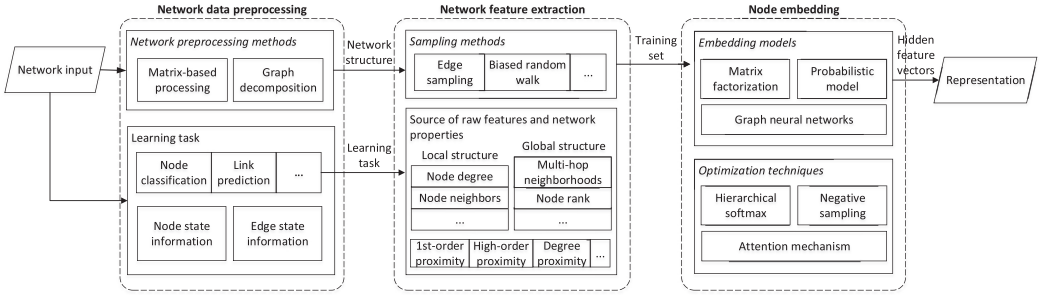

Fig. 3. A unified reference framework for NRL.

entire network and their hop counts are critical in the preprocessing stage to learn the node distance features in terms of graph traversal semantics in stage 2. These two steps will ensure that the NRL model training in stage 3 will deliver a high-quality NRL model for end users to perform their task-specific node classification or node clustering, which are network data and domain specific in real-world applications.

3.1.1 Network Preprocessing Methods. To effectively capture useful features, the network structure is usually preprocessed before feature extraction. We categorize current preprocessing methods into two types: matrix-based processing and graph decomposition.

Matrix-based processing. In most cases, we are using an adjacency matrix $\mathbf{A}$ to represent a network G, where its entries could directly describe the connection between arbitrary two nodes and the connection is also the basic unit of network structure. According to the hypothesis in DeepWalk that nodes with similar contexts are similar, a node's contexts are defined as the set of nodes arrived. To reflect transitions between nodes, a transition matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$ is proposed in the work of Cao et al. [14], where $\mathbf{D}$ is the diagonal matrix such that $\mathbf{D_{ii}} = \sum_j \mathbf{A_{ij}}$, and $\mathbf{P_{ij}}$ refers to the one-step transition probability from v_i to v_j. Accordingly, the transition matrix can be generalized to high steps, and the transition matrix within k steps [68] is obtained by $\mathbf{P}^* = \frac{1}{k}\sum_{m=1}^{k}\mathbf{P^m}$. Yang and Yang [98] defined two proximity matrices: the first-order proximity matrix $\mathbf{X^{(1)}}$ is the adjacency matrix, and the second-order proximity matrix $\mathbf{X^{(2)}}$ consists of $\mathbf{X_{ij}^{(2)}} = \cos(\mathbf{X_i^{(1)}}, \mathbf{X_j^{(1)}})$, where $\mathbf{X_i^{(1)}}$, $\mathbf{X_j^{(1)}}$ are the corresponding rows of $\mathbf{X^{(1)}}$. As an important branch of embedding, spectral methods require to convert adjacency matrix $\mathbf{A}$ into Laplacian matrix $\mathbf{L}$ before the Laplacian Eigenmaps [6], where $\mathbf{L} = \mathbf{D} - \mathbf{A}$.

Graph decomposition. Graph decomposition is another important type of methods for data preprocessing, by which the original network is decomposed into multiple graphs and each of them consists of a subset of nodes and a group of edges that correspond to connections in the original network or are connected based on certain rules. These graphs may be connected together to form a new network to extract features for a specific network property.

To capture the structural proximity, the context graph [70] is proposed by leveraging the decomposition idea. The context graph is a multi-layer weighted graph G_c, and each layer is a complete graph of all nodes. In particular, the edges in layer k are linked by nodes that are k-hop away from each other, and the edge weight in the layer is defined as $w_k(i,j) = e^{-f_k(i,j)}$, where $f_k(i,j)$ represents the k-structural distance between node v_i and v_j calculated based on their k-hop neighborhoods. For adjacent layers, the corresponding nodes are connected by directed edges, and the edge weights in both directions are defined as $w_{k,k+1}(v) = \log(\Gamma_k(v) + e)$ and $w_{k,k-1}(v) = 1$,

respectively, where $\Gamma_k(\mathrm{v})$ denotes the number of edges incident to node v such that their weights are higher than the layer k's average weight.

In social networks, ego-network $G[N_\mathrm{u}]$ induced on user u's neighborhood N_u is often used to denote her/his social circle. If the interactions of every node in the original network G are divided into couples of semantic subgroups, these subgroups will capture different components of user's network behavior. Based on the idea, Epasto and Perozzi [27] propose a method to convert G into its persona graph G_p, where each original node corresponds to multiple personas. Formally, given the original network $G = \{V, E\}$ and a clustering algorithm $\mathbb{A}$, the method consists of three steps: (1) for each node u ∈ V, its ego-network $G[N_\mathrm{u}]$ is partitioned into t_u disjoint components N_u^k via $\mathbb{A}$, denoted by $\mathbb{A}(G([N_\mathrm{u}])) = \{N_\mathrm{u}^k | k \in [1, t_\mathrm{u}]\}$; (2) collect a set V_p of personas, where each original node u induces t_u personas; and (3) add connections between personas if and only if (u, v) ∈ E, u ∈ N_v^i and v ∈ N_u^j. Social users often participate in different communities, the persona graph obtained via the preceding procedure presents a much better community structure for further embeddings.

For very large graphs whose scales exceed the capability of single machine, they must be decomposed into multiple partitions before further execution. Lerer et al. [54] propose a block decomposition method that first splits entity nodes into P parts and then divides edges into buckets based on their source and destination entity nodes' partitions. For example, for an edge (u, v), if source u and destination v are in partitions p_1 and p_2, respectively, then it should be placed into bucket (p_1, p_2). For each bucket (p_i, p_j), source and destination partitions are swapped from disk, respectively, and the edges are loaded accordingly for training. To ensure that embeddings in all partitions are aligned in the same latent space, an "inside-out" ordering is adopted to require that each bucket has at least one previously trained embedding partition.

3.1.2 Learning Task. Node classification aims to assign each node to a suitable group such that nodes in a group have some similar features. Link prediction aims to find out pairs of nodes that are most likely to be connected. Node classification and link prediction are two most basic fundamental tasks for network analytics. Both tasks could be further instantiated into many practical applications, such as social recommendation [28], knowledge completion [56], and disease-gene association identification [43]. Therefore, we mainly focus on the two basic tasks. For both tasks, a node's topological structure is the important basis for classification and prediction. For example, nodes with more common neighbors often have higher probability to be assigned to the same group or to be connected. This type of structure can be generalized to multi-hop neighborhoods [103], which requires computing the transition matrix of the network.

In addition to pure structural information, there is other useful information available for learning tasks, such as node state information and edge state information. In many real-world networks, the node itself may contain some state information, such as node attribute and node label, and this information may be essential for some tasks. For example, in social networks, besides embedding social connections, we can also encode user attributes to obtain more comprehensive representations for individuals [55, 106]. Node attributes are still an important source of features to be aggregated for inductive embeddings [8, 41]. Nodes with similar attributes and/or structures are more likely to be connected or classified together. Meanwhile, some node labels are usually fed into supervised models to boost the task of node classification [51, 104]. As the most common edge state information, edge weights can be integrated with topological structure to achieve more accurate classification and prediction. In addition, edges in some networks may have signs. Taking Epinions and Slashdot as examples, users in these two social network sites are allowed to connect to other users with either positive or negative edges, where positive edges represent trust and the like while negative edges convey distrust and dislike. For link prediction on such a signed network, we have to predict not only possible edges but also signs of those edges [49, 89].

Table 2. Classification of Source of Raw Features, Network Property, and Sampling Method

Category	Source of Raw Features	Network Property	Sampling Method
Local structure	Degree/in-degree/out-degree [70]	Degree proximity	Directly from adjacency matrix
	Neighbors/in-degree neighbors/out-degree neighbors [81]	First-order proximity	Directly from adjacency matrix
	Node state [41]	Node state proximity	Directly from node state information
	Adjacent edge state [49, 89]	First-order proximity	Weighted edge sampling
Global structure	Multi-hop neighborhoods [105]	High-order proximity	Random/Anonymous walk
	Node community membership [75, 90]	Community structure	Random/Anonymous walk
	Node connectivity pattern [4, 36]	Node role proximity	Random walk
	Node degree distribution [29]	Scale-free	Degree penalty based random walk
	Node rank [53]	Node importance ranking	Random walk
	Node identity [70, 84]	Node identity proximity	Random walk

3.2 Network Feature Extraction

The main task of NRL is to find out the hidden network features and encode such features as node embedding vectors in a low-dimensional space. Network properties are used to analyze and compare different network models. For an NRL task, the learned hidden features should preserve network properties so that advanced network analytics can be performed accurately and at the same time the original network properties can be preserved. For example, nodes that are closer in the original network should also be closer in the latent representation space in which the node embedding vectors are defined. Most of the NRL methods focus on preserving topological structures of nodes, such as in-degree or out-degree neighbors (degree proximity), first-order proximity, and random walk distance, among others. We categorize the node structural properties into local structure and global structure.

Embeddings from local structure focus on the preservation of local network properties such as degree proximity and first-order proximity. In comparison, global structure provides rich choices of sources of raw features to be extracted so as to preserve even more network properties. The classification of source of raw features as well as network properties and sampling methods are summarized in Table 2.

3.2.1 Local Structure Extraction. Local structure reflects a node's local view about the network, which includes node degree (in-degree and out-degree), neighbors (in-degree neighbors and out-degree neighbors), node state, and adjacent edge (in-edge and out-edge) state. To preserve degree proximity, Ribeiro et al. [70] define a proximity function $f_k(i, j)$ between two nodes where each node's neighbors are sorted based on their degrees and the proximity is measured by the distance between the sorted degree sequences. The first-order proximity [81] assumes nodes that are neighbors to each other (i.e., connected via an edge) are similar in vector space, whereas the degree of similarity may depend on the edge state. For example, in a signed social network, neighbors with positive and negative edges are often called *friends* and *foes*, respectively. From the perspective of social psychology [49, 89], when we take edge sign into account, nodes should be more similar to friends than foes in the representation space. Hamilton et al. [41] present an inductive representation learning by aggregating features extracted from neighboring node states.

Sampling methods. For the source of raw features like degree and neighbors, training instances can be calculated or fetched directly from adjacency matrix. For adjacent edges, training instances can be generated by edge sampling. The simple edge sampling is also to fetch entries from adjacency matrix. When applied to weighted networks where the pairwise proximity has close relationship with edge weights, if edge weights have a high variance, the learning model will suffer from gradient explosion or disappearance. To address the problem, LINE [81] designs an optimized edge sampling method that fetches edges with the probabilities proportional to their weights. For a node state like node attribute, its features can be extracted by leveraging existing embedding techniques, such as Word2vec [61]. If a node state is given as a node label, it usually works as the supervised item to train the embedding model.

3.2.2 Global Structure Extraction. Structures that transcend local views can be considered global, such as multi-hop neighborhoods, community, and connectivity pattern. Considering that the first-order proximity matrix may not be dense enough to model the pairwise proximity between nodes, as a global view [14], the pairwise proximity is generalized to high-order form by using k-step transition matrix (i.e., $\mathbf{P^k}$). Instead of preserving a fixed high-order proximity, Zhang et al. [105] define the arbitrary-order proximity by calculating the weighted sum of all k-order proximities. Community structure is an important network property with dense intra-community connections and sparse inter-community connections, and it has been observed in many domain-specific networks, such as social networks, co-authoring networks, and language networks. Wang et al. [90] introduce a community representation matrix by means of the modularity-driven community detection method and use it to enable each node's representation similar to the representation of its community. Considering that a node may belong to multiple communities, Sun et al. [75] define a $n \times m$ basis matrix $\mathbf{W}$ to reveal nodes' community memberships, where n is the node set size, m is the total number of communities, and $\mathbf{W_{ij}}$ indicates the propensity of node $\mathrm{v_i}$ to community C_j. The basis matrix is learned and preserved during the representation learning process.

Network nodes usually act as various structural roles [4, 36] and appear as different connectivity patterns such as hubs, star-edge nodes, and bridges connecting different clusters. Node role proximity assumes that nodes with similar roles have similar vector representations, and it is a global structural property that is different from community structure, as it primarily focuses on the connection patterns between nodes and their neighbors.

As a global structural property, node rank is always used to denote a node's importance in the network. PageRank [65] is a well-known approach to evaluate the rank of a node by means of its connections to others. Specifically, the ranking score of a node is measured by the probability of visiting it, whereas the probability is obtained from the ranking score accumulated from its direct predecessors weighted by the reciprocal of its out-degree. Lai et al. [53] demonstrate that node representations with global ranking preserved can potentially improve both results of ranking-oriented tasks and classification tasks. Node degree distribution is also a global inherent network property. For example, scale-free property refers to a fact that node degrees follow a heavy-tailed distribution, and it has proven to be ubiquitous across many real networks (Internet, social networks, etc.). The representation learning for scale-free is explored in the work of Feng et al. [29].

Another global structural property is proposed in Struc2vec [70] that considers structural similarity from network equivalence perspective without requiring two nodes being nearby (i.e., independent of nodes' network positions). To reflect this property, Struc2vec presents a notion of node structural identity that refers to a node's global sense. Struc2vec uses the multi-layer graph output from the data preprocessing stage to measure node similarity at different scales. In the bottom layer, the similarity exclusively depends on node degrees. In the top layer, the similarity lies in the entire network. Tu et al. [84] propose a similar concept (i.e., regular equivalence) to describe the similarity between nodes that may not be directly connected or not having common neighbors. According to its recursive definition, neighbors of regularly equivalent nodes are also regularly equivalent. To ensure the property of regular equivalence, each node's embedding is approximated by aggregating its neighbors' embeddings. After updating the learned representations iteratively, the final node embedding is capable of preserving the property in a global sense.

Different types of global structure are used to reflect different network properties where multi-hop neighborhoods, node connectivity pattern, and node identity are used to preserve pairwise proximity, which reflects a pairwise relationship between nodes, including high-order proximity, node role proximity, and node identity proximity. For example, the node community

membership reflects a relationship between a node and a group of nodes, which share some common network properties. Furthermore, node rank and node degree distribution are used to preserve a kind of distribution-based network property, including node importance ranking, or a relationship between a node and the entire network, such as the scale-free network whose degree distribution follows a power law.

Sampling methods. For source of raw features like multi-hop neighborhoods, they can be obtained by matrix power operation (i.e., $\mathbf{A^k}$), but the computation suffers from high complexity. Random walk and its variants are widely explored to capture the desirable network properties with high confidence. For example, DeepWalk [66] presents a truncated random walk to generate node traveling paths. It uses co-occurrence frequencies between node and its multi-hop neighborhoods along these paths to reflect their similarity and capture the high-order proximity accordingly.

From the perspective of community structure, due the dense intra-community connections, nodes within the same community have higher probability to co-occur on the traveling paths than nodes in different communities. Hence, random walk can also be used to capture the community structure. When we consider the hierarchy of communities, different communities may have different scales. The regular random walk makes the training set having more entries from $\mathbf{A^i}$ than from $\mathbf{A^j}$ ($1 \leq i < j$), and then it is biased toward preserving small-scale community structure. Walklets [67] presents a skipped random walk to sample multi-scale node pairs by skipping over steps in each traveling path.

Another drawback of random walk is that it requires too many steps or restarts to cover a node's neighborhoods. To improve its coverage, Diff2Vec [73] presents a diffusion-based node sequence generating method that consists of two steps. The first step is diffusion graph generation, which is in charge of generating a diffusion graph DG_i for each node $\mathrm{v_i}$. DG_i is initialized with $\mathrm{v_i}$, then randomly fetches node $\mathrm{v_j}$ from DG_i and node $\mathrm{v_k}$ from $\mathrm{v_j}$'s neighborhoods in the original graph, and appends two nodes and the edge e_{jk} to DG_i. The preceding process is repeated until DG_i grows to the predefined size. The second step is node sequence sampling, which generates Euler walk from DG_i as the node sequence. To make sure DG_i is Eulerian, DG_i is converted to a multi-graph by doubling every edge into two edges.

Real-world networks often exhibit a mixture of multiple network properties. To capture both community structure and node role proximity, node2vec [36] designs a flexible biased random walk that generates traveling paths in an integrated fashion of breadth-first sampling and depth-first sampling. To this end, two parameters p and q are introduced to smoothly interpolate between two sampling methods, where p decides the probability of re-fetching a node in the path while q allows the sampling to discriminate between inward and outward nodes.

In scale-free networks, a tiny fraction of "big hubs" usually attracts most edges. Considering that connecting to "big hubs" does not imply proximity as strong as connecting to nodes with mediocre degrees, a degree penalty based random walk [29] is proposed. For a pair of connected nodes ($\mathrm{v_i}$, $\mathrm{v_j}$), its principle is to reduce the likelihood of $\mathrm{v_j}$ being sampled as $\mathrm{v_i}$'s context when $\mathrm{v_i}$ has a high degree and they do not share many common neighbors. To this end, the jumping probability from $\mathrm{v_i}$ to $\mathrm{v_j}$ is defined as $\frac{C_{ij}}{(D_{ii}D_{jj})^{\beta}}$, where C_{ij} denotes the first and second order of proximity between two nodes, D_{ii} and D_{jj} are their degrees, and β is a parameter.

As an anonymous version of random walk, anonymous walk [46] provides a flexible way to reconstruct a network. For a random walk $\mathrm{rw} = (\mathrm{v_1}, \mathrm{v_2}, \ldots, \mathrm{v_k})$, its corresponding anonymous walk is the sequence of node's first occurrence positions—that is, $\mathrm{aw} = (\mathrm{f}(\mathrm{v_1}), \mathrm{f}(\mathrm{v_2}), \ldots, \mathrm{f}(\mathrm{v_k}))$, where $\mathrm{f}(\mathrm{v_i}) = \min_{\mathrm{p_j} \in \mathrm{pos}(\mathrm{rw}, \mathrm{v_i})} \mathrm{pos}(\mathrm{rw}, \mathrm{v_i})$. For an arbitrary node $\mathrm{v} \in \mathrm{G}$, a known distribution D_l over anonymous walks of length l is sufficient to reconstruct a subgraph of G that is induced by all nodes located within r hops away from v. Therefore, anonymous walk can be used to preserve

global network properties like high-order proximity and community structure by approximating the distributions.

One of the baseline approaches to extracting global structure is to use random walk as the sampling method. For complex types of global structure, such as multi-hop neighborhoods and node community membership, an integrated sampling method is often recommended, which combines random walk with other types of graph traversal methods, such as anonymous walk. An advantage of using anonymous walk as the sampling approach is that it is sufficient to reconstruct the topology around a node by utilizing distribution of anonymous walks of a single node, because anonymous walk captures richer semantics than random walk.

3.3 Node Embedding

In recent years, many efforts have been devoted to the design of the node embedding model. We are trying to review those works from a universal perspective of the autoencoder. In general, node embedding is equivalent to an optimization problem that encodes network nodes into latent vectors by means of an encoding function $\Phi : V \rightarrow \mathbb{R}^d$. Meanwhile, the objective is to ensure that the results decoded from vectors preserve the network properties we intend to incorporate, where the decoder is represented by a function of encoding results—that is, $\Psi : [\mathbb{R}^d]^t \rightarrow \mathbb{R}$, where t denotes the number of input arguments. The output vectors are the latent representations of hidden features learned by Φ, and they are also the expected node representations.

3.3.1 Embedding Models. We classify the current embedding models into the following three types: matrix factorization, the probabilistic model, and GNNs.

Matrix factorization. Given the matrix of the input network, matrix factorization embedding factorizes the matrix to get a low-dimensional matrix as the output collection of node representations. Its basic idea can be traced back to the matrix dimensionality reduction techniques [18]. According to the matrix type, we categorize the current work into relational matrix factorization and the spectral model.

(1) Relational matrix factorization. Matrix analysis often requires figuring out the inherent structure of a matrix by a fraction of its entries, and it is always based on an assumption that a matrix $\mathbf{M_{n\times n}}$ admits an approximation of low rank $d \ll n$. Under this assumption, the objective of matrix factorization corresponds to finding a matrix $\mathbf{W_{d\times n}}$ such that $\mathbf{W^T W}$ approximates $\mathbf{M}$ with the lowest loss L, where $L : [\mathbb{R}]^2 \rightarrow \mathbb{R}$ is a user-defined loss function. In the autoencoder paradigm, the encoder is defined by the matrix $\mathbf{W_{d\times n}}$ (e.g., $\Phi(\mathrm{v_i}) = \mathbf{w_i}$, where each column $\mathbf{w_i}$ represents a vector). The decoder is defined by the inner product of two node vectors (e.g., $\Psi(\mathbf{w_i}, \mathbf{w_j}) = \mathbf{w_i^T w_j}$) so as to infer the reconstruction of the proximity between node $\mathrm{v_i}$ and $\mathrm{v_j}$. When we focus on preserving the first-order proximity μ (i.e., $\mu(\mathrm{v_i}, \mathrm{v_j}) = \mathbf{A_{ij}}$) and let Ω be the training set sampled from $\mathbf{A}$, then the objective is to find $\mathbf{W}$ to minimize the reconstruction loss—that is,

$$\arg\min_{\mathbf{W}} L = \sum_{(\mathrm{v_i}, \mathrm{v_j}) \in \Omega} \left\| \Psi(\mathbf{w_i}, \mathbf{w_j}) - \mu(\mathbf{v_i}, \mathbf{v_j}) \right\|_F^2 = \left\| \mathbf{W^T W} - \mathbf{A} \right\|_F^2, \tag{9}$$

where Frobenius norm $\|\cdot\|_F^2$ is often used to define the loss function. Singular value decomposition (SVD) [34] is a well-known matrix factorization approach that can find the optimal rank d approximation of proximity matrix $\mathbf{A}$. If the network property focuses on high-order proximity, the proximity matrix can be replaced by a power matrix (e.g., $\mathbf{A^k}$, $\mathbf{P^k}$, $k > 1$). For example, GraRep [14] implements node embedding by factorizing $\mathbf{P^k}$ into two matrices $\mathbf{W}$ and $\mathbf{M}$,

$$\arg\min_{\mathbf{W}} \left\| \mathbf{W^T M} - \mathbf{P^k} \right\|_F^2, \tag{10}$$

where $\mathbf{W}$ denotes the representation matrix and $\mathbf{M}$ denotes the parameter matrix in decoder—for example, the unary decoder $\Psi(\mathbf{w_i}) = \mathbf{w_i^T M}$ computes the reconstructed proximities between v_i and other nodes.

In the autoencoder paradigm, the equivalence between DeepWalk and matrix factorization can be proved by making the following analogies: (1) define the pairwise proximity $\mu(v_i, v_j)$ as the co-occurrence probability inferred from the training set, and (2) let $\mathbf{W}$ and $\mathbf{M}$ correspond to $\mathbf{W^T}$ and $\mathbf{M^T}$ in DeepWalk, where each column $\mathbf{w_i}$, $\mathbf{m_i}$ of them refers to the representation of node v_i acting as target node and context node, respectively.

In addition, matrix factorization can incorporate additional information into node embeddings. For example, given a text feature matrix $\mathbf{T_{x\times n}}$ where n denotes the node set size and x denotes the feature vector dimension, TADW [97] applies inductive matrix completion to the node embedding and defines the following matrix factorization problem:

$$\arg\min_{\mathbf{W},\mathbf{M}} \left\|\mathbf{P}^* - \mathbf{W}^T\mathbf{M}\mathbf{T}\right\|_F^2 + \frac{\lambda}{2}\left(\|\mathbf{W}\|_F^2 + \|\mathbf{M}\|_F^2\right), \tag{11}$$

where λ is a harmonic factor. The output matrices $\mathbf{W}$ and $\mathbf{MT}$ factorized from the transition matrix $\mathbf{P}^*$ can be regarded as the collection of node embeddings and concatenated as an $n \times 2d$ representation matrix.

(2) Spectral model. In the spectral model, a network is mathematically represented by a Laplacian matrix—that is, $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where adjacency matrix $\mathbf{A}$ acts as the proximity matrix, and the entry $\mathbf{D_{ii}}$ of diagonal matrix $\mathbf{D}$ describes the importance of node v_i. $\mathbf{L}$ is a symmetric positive semi-definite matrix that can be thought of as an operator on functions defined on the original network. Let $\mathbf{W_{n\times d}}$ be the representation matrix, and it also acts as the encoder that maps the network to a d-dimensional space. The representations should assure that neighboring nodes stay as close as possible. As a result, the decoder can be defined as $\mathbf{w_i^T L w_i}$ according to Laplacian eigenmaps [7], where $\mathbf{w_i}$ is the ith column of $\mathbf{W}$. Then the node embedding problem is defined as follows:

$$\arg\min_{\mathbf{W^TDW=I}} \mathbf{W}^T\mathbf{L}\mathbf{W}, \tag{12}$$

where $\mathbf{W^TDW = I}$ is imposed as a constraint to prevent collapse onto a subspace of dimension less than $d - 1$. The solution $\mathbf{W}$ consists of the eigenvectors $\mathbf{w_i}$ corresponding to the lowest d eigenvalues of the generalized eigenvalue problem $\mathbf{Lw} = \lambda\mathbf{Dw}$. Furthermore, given an additional information matrix $\mathbf{X}$ describing node attribute features, the node embedding can be derived by following the idea of locality preserving projection (LPP) [44] that introduces a transformation matrix $\mathbf{M}$ to realize the mapping $\mathbf{w_i} = \mathbf{M^T x_i}$, where $\mathbf{w_i}$ and $\mathbf{x_i}$ are the ith column of $\mathbf{W}$ and $\mathbf{X}$. Similarly, the solution $\mathbf{M}$ is obtained by computing the mapping vectors $\mathbf{m}$ corresponding to d lowest eigenvalues of the problem $\mathbf{XLX^Tm} = \lambda\mathbf{XDX^Tm}$, and the node embedding incorporates additional information by $\mathbf{W} = \mathbf{M^TX}$.

The probabilistic model. The probabilistic model is specifically designed for node embedding via preserving the pairwise proximity measured by a flexible probabilistic manner.

(1) The skip-gram model. Skip-gram is the classic embedding model that converts pairwise proximity to the conditional probability between a node and its context. For example, DeepWalk [66] relies on the co-occurrence probability derived from random traveling paths to preserve the second-order proximity during node embedding.

(2) The edge probabilistic model. The edge probabilistic model enforces node embeddings designed primarily for network reconstruction. For example, LINE [81] relies on the idea of edge probability reconstruction without assistance of random walk sampling. It focuses on preserving

both first-order and second-order proximities by defining two decoders—that is,

$$\Psi_1(\mathbf{v_i}, \mathbf{v_j}) = \Pr(v_i, v_j) = \frac{1}{1 + \exp(-\mathbf{v_i}^T \mathbf{v_j})}, \ \Psi_2(\mathbf{v_i}, \mathbf{m_j}) = \Pr(v_j|v_i) = \frac{\exp(\mathbf{m_j}^T \mathbf{v_i})}{\sum_{k=1}^{n} \exp(\mathbf{m_k}^T \mathbf{v_i})}, \tag{13}$$

where $\mathbf{m_k}$ refers to the representation of v_k acting as a context node. The objectives to be optimized are based on the loss functions derived from the distance $Dis(\cdot, \cdot)$ between two distributions—that is,

$$\arg\min_{\mathbf{v_i}} \sum_{i=1}^{n} \lambda_i Dis(\Psi_1(\mathbf{v_i}, \cdot), \hat{\Psi}_1(\mathbf{v_i}, \cdot)), \ \arg\min_{\mathbf{v_i}} \sum_{i=1}^{n} \lambda_i Dis(\Psi_2(\mathbf{v_i}, \cdot), \hat{\Psi}_2(\mathbf{v_i}, \cdot)), \tag{14}$$

where $\hat{\Psi}_1(\mathbf{v_i}, \mathbf{v_j}) = \frac{s_{ij}}{\sum_{e_{xy} \in E} w_{xy}}$, $\hat{\Psi}_2(\mathbf{v_i}, \mathbf{v_j}) = \frac{s_{ij}}{\sum_{e_{ik} \in E} s_{ik}}$ are the empirical probabilities and s_{ij} is the weight of edge e_{ij}.

Graph neural networks. It is not difficult to find that each node's embedding via the preceding models requires the participation of its neighboring nodes—for example, building k-order proximity matrix with nodes and their k-step neighborhoods in matrix factorization model, and counting node's co-occurrence frequency with its context nodes in the skip-gram model. The common idea can be intuitively generalized to a more general model—GNNs that follow a recursive neighborhood aggregation or message passing scheme. The **graph convolutional network (GCN)** [11] is a very popular variant of GNNs, where each node's representation is generated by a convolution operation that aggregates its own features and its neighboring nodes' features. The **graph isomorphism network (GIN)** [96] is another recently proposed variant with a relatively strong expressive power.

(1) The GCN. In terms of the definition of convolution operation [95], GCNs are often grouped into two categories: spectral GCNs and spatial GCNs. Spectral GCNs [11, 22] define the convolution as conducting the eigendecomposition of the normalized Laplacian matrix $\mathbf{L} = \mathbf{I_n} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ in the Fourier domain. An intuitive way to explain spectral convolution is to regard it as an operation that uses a filter $\mathbf{g}$ to remove noise from a graph signal $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$. The graph signal denotes a feature vector of the graph with entry $\mathbf{x_i}$ representing node v_i's value. Let $\mathbf{U}$ be the matrix of eigenvectors of $\mathbf{L}$, and graph Fourier transform is defined as a function that projects input signal $\mathbf{x}$ to the orthonormal space formed by $\mathbf{U}$ (i.e., $\mathcal{F}(\mathbf{x}) = \mathbf{U}^T\mathbf{x}$). Let $\hat{\mathbf{x}}$ be the transformed signal, and entries of $\hat{\mathbf{x}}$ correspond to the coordinates of the input signal in the new generated space, and the inverse Fourier transform is defined as $\mathcal{F}^{-1}(\hat{\mathbf{x}}) = \mathbf{U}\hat{\mathbf{x}}$. When the filter $\mathbf{g}$ is parameterized by $\theta \in \mathbb{R}^{\mathbf{n}}$ and defined as $\mathbf{g}_\theta = diag(\mathbf{U}^T\mathbf{g})$, the convolution operation against signal $\mathbf{x}$ with filter $\mathbf{g}_\theta$ is represented by

$$\mathbf{g}_\theta \star \mathbf{x} = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{g}) \circ \mathcal{F}(\mathbf{x})) = \mathbf{U}\left(\mathbf{U}^T\mathbf{g} \circ \mathbf{U}^T\mathbf{x}\right) = \mathbf{U}\mathbf{g}_\theta\mathbf{U}^T\mathbf{x}, \tag{15}$$

where $\circ$ represents the Hadamard product. Nevertheless, The eigendecomposition of Laplacian matrix $\mathbf{L}$ is expensive for large graphs, and the complexity of multiplication with $\mathbf{U}$ is $O(n^2)$. To address the complexity problem, Hammond et al. [42] suggest that the filter $\mathbf{g}_\theta$ can be approximated by an abridged expansion based on Chebyshev polynomials $T_k(\mathbf{x})$ up to Kth order. The convolution is redefined as

$$\mathbf{g}_{\theta'} \star \mathbf{x} \approx \sum_{k=0}^{K} \theta'_k T_k(\tilde{\mathbf{L}})\mathbf{x}, \tag{16}$$

where $\tilde{\mathbf{L}} = \frac{2}{\lambda_{\max}}\mathbf{L} - \mathbf{I_n}$ and $T_k(\mathbf{x}) = 2\mathbf{x}\mathbf{T_{k-1}}(\mathbf{x}) - \mathbf{T_{k-2}}(\mathbf{x})$ with $T_1(\mathbf{x}) = \mathbf{x}$ and $T_0(\mathbf{x}) = 1$. The preceding convolution operation is K-localized since it requires the participation of the neighboring nodes within K-hop away from the central node. The spectral GCN model consists of multiple convolution layers of the form Equation (16) where each layer is followed by a pointwise non-linearity.

From the perspective of the autoencoder, each node's feature embedding is encoded by the convolution operation so that the graph convolution actually acts as the encoder, and we call it *spectral convolution encoder* or *filter encoder* here. To deal with multi-dimensional input graph signals, the spectral convolution encoder is generalized to account for the signal matrix $\mathbf{X} \in \mathbb{R}^{\mathbf{n}\times\mathbf{c}}$ with c input channels and d filters—that is,

$$\Phi(\mathbf{X}, \mathbf{A}) = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}\Theta, \tag{17}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I_n}$, $\tilde{\mathbf{D}}_{\mathbf{ii}} = \sum_{\mathbf{j}} \tilde{\mathbf{A}}_{\mathbf{ij}}$ and $\Theta \in \mathbb{R}^{c\times d}$ is the matrix of filter parameters. The signal matrix is often initialized by the input graph information such as node attributes. It is noted that the filter parameters Θ can be shared over the whole graph, which significantly decreases the amount of parameters as well as improves the efficiency of filter encoder.

Similar to convolutional neural networks on images, spatial GCNs [41] consider graph nodes as image pixels and directly define convolution operation in the graph domain as the feature aggregation from neighboring nodes. To be specific, the convolution acts as the encoder to take the aggregation of the central node representation and its neighbors' representations to generate an updated representation, and here we call it the *spatial convolution encoder* or *aggregation encoder*. To explore the depth and breadth of a node's receptive field, spatial GCNs usually stack multiple convolution layers. For a K-layer spatial GCN, its aggregation encoder Φ is defined as follows:

$$\begin{cases} \Phi(\mathbf{X}, \mathbf{A}) = NORM(\mathbf{H}^{\mathrm{K}}), \mathbf{H}^{\mathrm{K}} = (\mathbf{h}_1^{\mathrm{K}}, \ldots, \mathbf{h}_{\mathrm{n}}^{\mathrm{K}}), \\ \mathbf{h}_{\mathrm{v}}^{\mathrm{k}} = \sigma\left(\mathbf{W}^{\mathrm{k}} \cdot AGG_{\mathrm{k}}\left(\mathbf{h}_{\mathrm{v}}^{\mathrm{k}-1}, \left\{\mathbf{h}_{\mathrm{u}}^{\mathrm{k}-1}\right\}\right)\right), \forall \mathrm{u} \in N_{\mathrm{v}}, \forall \mathrm{k} \in [0, \mathrm{K}], \end{cases} \tag{18}$$

where $\mathbf{h}_{\mathrm{v}}^{\mathrm{k}}$ is node v's representation (also called *hidden state* in some other literature) in the kth layer with $\mathbf{h}_{\mathrm{v}}^{0} = \mathbf{x}_{\mathrm{v}}$. AGG_{k} is an aggregation function responsible for assembling a node's neighborhood information in the kth layer, and parameter matrix $\mathbf{W}^{\mathbf{k}}$ specifies how to do aggregation from neighborhoods, like filters in spectral GCNs, and $\mathbf{W}^{\mathbf{k}}$ is shared across all graph nodes for generating their representations. As the layer deepens, the node representations will contain more information aggregated from wider coverage of nodes. The final node representations are output as the normalized matrix of node representations on layer K (i.e., $NORM(\mathbf{H}^{\mathrm{K}})$).

(2) The GIN. The representational power of a GNN primarily depends on whether it maps two multisets to different representations, where a multiset is a generalized concept of a set that allows multiple instances for its elements. The aggregation operation can be regarded as a class of functions over multisets that their neural networks can represent. To maximize the representational power of a GNN, its multiset functions must be injective. GCN has been proven to be not able to distinguish certain simple structural features, as its aggregation operation is inherently not injective. To model injective multiset functions for the neighbor aggregation, GIN [96] presents a theory of deep multisets that parameterizes universal multiset functions with neural networks. With the help of multi-layer perceptrons (MLPs), GIN updates node representations as follows:

$$\mathbf{h}_{\mathrm{v}}^{\mathrm{k}} = MLP^{\mathrm{k}}\left(\left(1 + \varepsilon^{\mathrm{k}}\right)\mathbf{h}_{\mathrm{v}}^{\mathrm{k}-1} + \sum_{\mathrm{u}\in N_{\mathrm{v}}} \mathbf{h}_{\mathrm{u}}^{\mathrm{k}-1}\right), \tag{19}$$

where ε is a learnable parameter or a fixed scalar. The aggregation encoder Φ can be defined in the same way of Equation (18).

For both GCN and GIN, the decoder can be designed in any form of the previously discussed decoders. Recently, many efforts have been devoted to designing task-driven GCNs so that the decoder has to incorporate supervision from a specific task. For instance, assume that each node has a label chosen from a label set $\mathbf{y}$. A function like sigmoid, $y_i = \Psi_\theta(\Phi(\mathbf{x_i}, \mathbf{A}))$, can be defined as a decoder to realize the mapping from a node representation $\Phi(\mathbf{x_i}, \mathbf{A})$ to its corresponding label y_i, where parameters θ are learnable throughout the embedding process. We use function

Table 3. Classification of Node Embedding Models

Category	Sub-Category	Encoder	Decoder	Optimization Objective	Reference
Matrix Factorization	Relational matrix factorization	$\Phi(\mathrm{v_i}) = \mathbf{w_i}$	$\Psi(\mathbf{w_i}, \mathbf{w_j}) = \mathbf{w_i^T w_j}$, $\Psi(\mathbf{w_i}) = \mathbf{w_i^T M}$, $\Psi(\mathbf{w_i}) = \mathbf{w_i^T MT}$	Equation (9), Equation (10), Equation (11)	[14, 97]
	Spectral model	$\Phi(V) = \mathbf{W}$, $\Phi(V) = \mathbf{M^T X}$	$\mathbf{w_i^T L w_i}$, $\mathbf{x_i L x_i^T m_i}$	Equation (12)	[7, 44]
Probabilistic Model	Skip-gram model	$\Phi(\mathrm{v_i}) = \mathbf{w_i}$	Equation (2)	Equation (4)	[66]
	Edge probabilistic model	$\Phi(\mathrm{v_i}) = \mathbf{w_i}$	Equation (13)	Equation (14)	[81]
Graph Neural Networks	Spectral graph convolutional network	Equation (17)	$\Psi_\theta(\Phi(\mathbf{x_i}, \mathbf{A}))$	Equation (20)	[42, 50, 51]
	Spatial graph convolutional network	Equation (18)	$\Psi_\theta(\Phi(\mathbf{x_i}, \mathbf{A}))$	Equation (20)	[41]
	Graph isomorphism network	Equation (18), Equation (19)	$\Psi_\theta(\Phi(\mathbf{x_i}, \mathbf{A}))$	Equation (20)	[96]

Table 4. Comparison of Node Embedding Models

Category	Advantage	Disadvantage
Matrix Factorization	(1) Has the solid theoretical foundation available to enhance its interpretability, and (2) additional information (e.g., node state and edge state) can be easily incorporated into the matrix factorization model	(1) High computation cost and (2) hard to be applied to dynamic embedding
Probabilistic Model	Relatively efficient compared to other models, especially when it was designed for network reconstruction	Hard to be applied to dynamic embedding
Graph Neural Networks	(1) Naturally supports embedding of both structural and additional information and (2) easy to be applied to dynamic embedding	High computation cost

$f(\hat{y}_i, \Psi_\theta(\Phi(\mathbf{x_i}, \mathbf{A})))$ to measure the loss between the true label $\hat{y}_i$ and the predicted one y_i, then the objective is concluded as

$$\arg\min_{\Phi(\mathbf{x_i}, \mathbf{A})} \sum_{\mathrm{i}=1}^{n} f(\hat{y}_\mathrm{i}, \Psi_\theta(\Phi(\mathbf{x_i}, \mathbf{A}))). \tag{20}$$

We summarize the specific classification of node embedding models outlined in Table 3 and make comparisons among them as shown in Table 4.

3.3.2 Optimization Techniques. Many of the preceding models have achieved non-trivial embedding results. The success of these models inevitably relies on some optimization techniques that have been proposed to assist the node embedding from various aspects such as the computational complexity reduction, the acceleration of embedding process, and the enhancement of training efficiency and effectiveness. In this section, we will show several popular optimization techniques with a focus on how they work at different stages of NRL, and summarize these techniques in Table 5.

Hierarchical softmax. As a variant of softmax, hierarchical softmax [62, 66] has been proposed to speed up the training process. To this end, hierarchical softmax leverages a well-designed tree structure with multiple binary classifiers to compute the conditional probabilities for each training instance (see Section 2 for details). Therefore, it helps the decoder reduce the complexity from $O(n)$ to $O(\log n)$.

Negative sampling. As an alternative to hierarchical softmax, noise contrastive estimation [39] suggests that logistic regression can be used to help models distinguish data from noise. Negative sampling is a simplified version of noise contrastive estimation that was first leveraged by Word2vec [61] to help the skip-gram model deal with the difficulty of high computational complexity with respect to model training. Specifically, negative sampling needs a noise distribution $P_n(\mathrm{v})$ to generate k negative samples for every positive one. $P_n(\mathrm{v})$ can be arbitrarily chosen, and more often it is empirically set by raising the frequency to the 3/4 power for the best quality of

Table 5. Summary of Optimization Techniques

Technique	Working Stage	TechnicalPrinciple	Optimization Goal	Reference
Hierarchical Softmax	Node embedding	Leverage a tree structure to minimize the computation of conditional probabilities	Computational complexity reduction, acceleration of the embedding process	[62, 66]
Negative Sampling	Feature extraction	Reduce the number of output vectors that need to be updated	Computational complexity reduction, acceleration of the embedding process	[39, 61]
Attention Mechanism	(1) Feature extraction and (2) node embedding	(1) Replace previously fixed hyperparameters with trainable ones, and(2) distinguish the neighborhood's importance via trainable weights	Enhancement of training efficiency and effectiveness	[3, 86]

embeddings. The training loss of negative sampling is formally defined as

$$L = -\log \sigma(\mathbf{v_O}^{\mathrm{T}} \Phi(\mathrm{v_I})) - \sum_{\mathrm{v_j} \in \mathrm{V}_{neg}} E_{\mathrm{v_j} \sim P_{n(\mathrm{v})}} \left[\log \sigma \left(-\mathbf{v_j}^{\mathrm{T}} \Phi(\mathrm{v_I}) \right) \right], \tag{21}$$

where V_{neg} is the set of negative samples and $\mathbf{v_O}$ is the output vector of node $\mathrm{v_O}$ that corresponds to the column O of matrix $\mathbf{M}$ in the skip-gram model. Then the training objective is simplified to be able to discriminate the output node $\mathrm{v_O}$ from nodes draws from $P_n(\mathrm{v})$ using logistic regression. To update node embeddings under negative sampling, the derivative of loss L with regard to the input of output $\mathrm{v_j}$ is given by

$$\frac{\partial L}{\partial \mathbf{v_j}^{\mathrm{T}} \Phi(\mathrm{v_I})} = \sigma \left(\mathbf{v_j}^{\mathrm{T}} \Phi(\mathrm{v_I}) \right) - s_j, \tag{22}$$

where s_j is a binary indicator of samples—that is, $s_j = 0$ if $\mathrm{v_j}$ is a negative sample and otherwise $s_j = 1$. The output vector is updated by

$$\mathbf{v_j} \leftarrow \mathbf{v_j} - \eta \left(\sigma \left(\mathbf{v_j}^{\mathrm{T}} \Phi(\mathrm{v_I}) \right) - s_j \right) \Phi(\mathrm{v_I}). \tag{23}$$

By using negative sampling, we just need to update the output vectors of nodes from $\{\mathrm{v_O}\} \cup \mathrm{V}_{neg}$ instead of the entire node set V. The computational effort is therefore saved significantly. Finally, the node vector is updated accordingly by the error backpropagation to the hidden layer—that is,

$$\Phi(\mathrm{v_I}) \leftarrow \Phi(\mathrm{v_I}) - \eta \sum_{\mathrm{v_O} \in context(\mathrm{v_I})} \sum_{\mathrm{v_j} \in \{\mathrm{v_O}\} \cup \mathrm{V}neg} \left(\sigma \left(\mathbf{v_j}^{\mathrm{T}} \Phi(\mathrm{v_I}) \right) - s_j \right) \mathbf{v_j}. \tag{24}$$

Attention mechanism. Ever since the attention mechanism was proposed, it has become an effective way to help models focus on the most important part of data. NRL also benefits from the attention mechanism by conducting attention-guided random walks in the feature extraction stage and designing an attention-based encoder in the node embedding stage.

In the stage of feature extraction, the attention mechanism [3] is borrowed to lead random walk to optimize an upstream objective. Let Γ be the transition matrix, $\tilde{\mathbf{P}}^{(0)}$ be the initial positions matrix with $\tilde{\mathbf{P}}^{(0)}_{\mathrm{vv}}$ set to the number of walks starting at node v, and C be the walk length, and the context distribution can be represented by a C-dimensional vector $\mathbf{Q} = (\mathbf{Q_1}, \mathbf{Q_2}, ..., \mathbf{Q_C})$. To obtain the expectation on the co-occurrence matrix, $\mathbb{E}[\mathbf{D}]$, $\mathbf{Q_k}$ need to be assigned to $\Gamma^{\mathbf{k}}$ as a coefficient. An attention model is proposed to learn $\mathbf{Q}$ automatically, where $\mathbf{Q} = \mathrm{softmax}((\mathrm{q_1}, \mathrm{q_2}, ..., \mathrm{q_C}))$, and all $\mathrm{q_k}$ can be trained by backpropagation. The attention model aims to guide the random surfer on "where to attend to" as a function of distance. To this end, the model on Γ^{∞} is trained according to the expectation on the random walk matrix—that is,

$$\mathbb{E}[\mathbf{D}^{\mathrm{softmax}[\infty]}; \mathrm{q_1}, \mathrm{q_2}, ..., \mathrm{q_\infty}] = \tilde{\mathbf{P}}^{(0)} \lim_{C \to \infty} \sum_{\mathrm{k}=1}^{C} \mathrm{softmax}(\mathrm{q_1}, \mathrm{q_2}, ..., \mathrm{q}_k) \Gamma^{\mathrm{k}}. \tag{25}$$

For many random walk based methods like DeepWalk, they are special cases of the preceding equation where C is not infinite and $\mathbf{Q}$ are fixed *a priori* (i.e., $\mathbf{Q_k} = [1 - \frac{\mathrm{k}-1}{C}]$).

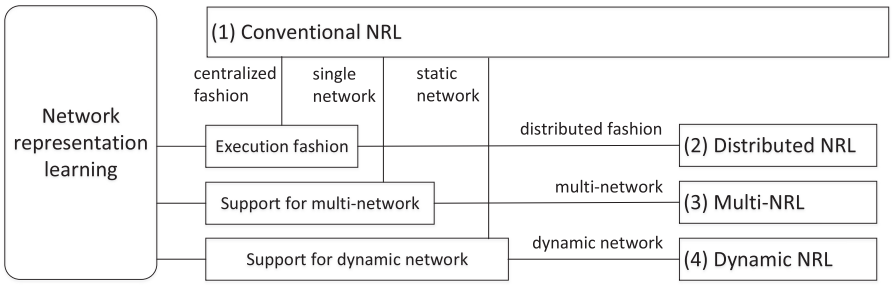

Fig. 4. A taxonomy of NRLs based on its evolution directions.

In the stage of node embedding, the **graph attention network (GAT)** [86] proposes to incorporate the attention mechanism into a spatial GCN for providing differentiated weights of neighborhoods. Specifically, GAT defines a graph attention layer parameterized by a weight vector $\mathbf{a}$ and builds a GAT by stacking the layers. The attention function $\alpha(\cdot)$ is defined to measure the importance of neighbor u to the central node v,

$$\alpha_{vu} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_v||\mathbf{W}\mathbf{h}_u]))}{\sum_{w\in N_v}\exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_v||\mathbf{W}\mathbf{h}_w]))}, \tag{26}$$

where $\mathbf{h}_v$ refers to the feature of input node and $\mathbf{W}$ denotes the weight matrix of a shared linear transformation that is applied to every node. The convolution operation (aggregation encoder) is defined as

$$\mathbf{h}_v^t = \sigma\left(\sum_{u\in N_v}\alpha\left(\mathbf{h}_v^{t-1}, \mathbf{h}_u^{t-1}\right)\mathbf{W}^{t-1}\mathbf{h}_u^{t-1}\right). \tag{27}$$

4 RECENT ADVANCES IN NRL

Here we review existing NRL studies with a focus on recent methods that have achieved significant advances in machine learning and/or data mining. These studies are classified into four categories, and the detailed taxonomy is shown in Figure 4, which enables us to classify current NRL methods based on whether the learning is (1) performed in centralized or distributed fashion, (2) using a single network or multiple networks, and (3) applied to a static network or dynamic evolving network. In addition, we include some works on the knowledge graph as one of important extensions of NRL. We review these works according to the proposed reference framework and present a brief summary, as shown in Table 6.

4.1 Conventional NRL Methods

Methods in this category share some common features, such as performing in a centralized fashion, applying to only a single NRL, and only learning from a static network. We introduce recent advances in this category according to two main types of network properties: pairwise proximity and community structure.

4.1.1 Pairwise Proximity. MVE [69] presents a multi-view embedding framework where pairs of nodes with different views are sampled as instances. Each view corresponds to a type of proximity between nodes, such as, the following-followee, reply, retweet relationships in many social networks. These views are usually complementary to each other. MVE uses the skip-gram model to yield the view-specific representations that preserve the first-order proximities encoded in

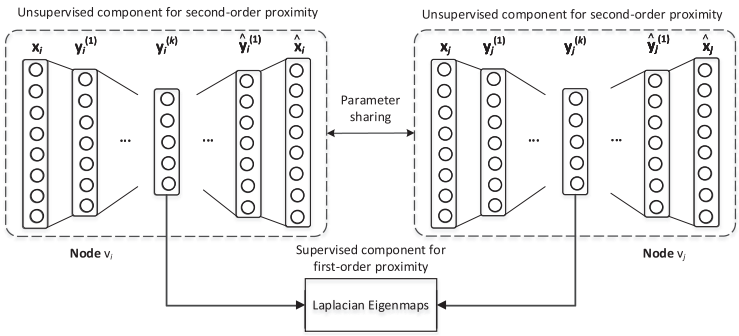

Fig. 5. The semi-supervised model of SDNE (adapted from Figure 2 [87]).

different views. Moreover, an attention-based voting scheme is proposed to identify important views by learning different weights of views.

SDNE [87] presents a semi-supervised learning model to learn node representations by capturing both local and global structure. The model architecture is illustrated in Figure 5 that consists of two components: the unsupervised component and supervised component. The former is designed to learn node representations by preserving the second-order proximity, and the latter utilizes node connections as the supervised information to exploit the first-order proximity and refine node representations. Specifically, given the adjacency matrix $\mathbf{A}$, the first component utilizes a deep autoencoder to reconstruct the neighborhood structure of each node. The encoder relies on multiple non-linear functions to encode each node $\mathrm{v_i}$ into a vector representation. The hidden representation in the kth layer is defined as

$$\mathbf{y}_i^{(k)} = \sigma\left(\mathbf{W}^{(k)}\mathbf{y}_i^{(k-1)} + \mathbf{b}^{(k)}\right), \tag{28}$$

where $\mathbf{x_i}$ is the input vector of $\mathrm{v_i}$, $\mathbf{y}_i^{(0)} = \mathbf{x}_i$, $\mathbf{W^{(k)}}$ and $\mathbf{b^{(k)}}$ are weights and biases, respectively, in the kth layer. The decoder reconstructs the input vectors (e.g., $\hat{\mathbf{x}}_i$) from the most hidden vectors (e.g., $\mathbf{y}_i^{(k)}$) by means of non-linear functions. Note that the number of zero elements in $\mathbf{A}$ is far less than that of zero elements. The autoencoder is prone to reconstruct the zero elements. To avoid this situation, SDNE imposes more penalty to the reconstruction error of non-zero elements and the loss function is defined as

$$L_{2nd} = \sum_{i=1}^{n} \|(\hat{\mathbf{x}}_i - \mathbf{x}_i) \circ \mathbf{b}_i\|_2^2, \tag{29}$$

where $\circ$ denotes the Hadamard product. The second component enhances the first-order proximity by borrowing the idea of Laplacian eigenmaps to incur a penalty once neighboring nodes are embedded far away. Consequently, the loss function is defined as

$$L_{1st} = \sum_{i,j=1}^{n} \mathbf{A}_{ij} \left\|(\mathbf{y}_i^{(k)} - \mathbf{y}_j^{(k)}) \circ \mathbf{b}_i\right\|_2^2. \tag{30}$$

In addition, to avoid falling to local optima in the parameter space, SDNE leverages the deep belief network to pretrain the parameters at first.

DNGR [15] designs a deep denoising autoencoder method to capture the non-linearities of network features. Its basic idea is to learn node representations from the **positive pointwise mutual positive pointwise mutual information (PPMI)** matrix of nodes and their contexts. As illustrated in Figure 6, DNGR consists of three components: random surfing, calculation of PPMI, and a **stacked denoising autoencoder (SDAE)**. In the first component, random surfing is used to

generate the **probabilistic co-occurrence (PCO)** matrix that corresponds to a transition matrix by nature. Then the second component calculates the PPMI matrix based on the PCO matrix by following the work of Bullinaria and Levy [12]. After that, SDAE is presented for highly non-linear abstractions learning. To recover the complete matrix under certain assumptions, SDAE partially corrupts the training sample $\mathbf{x}$ by randomly assigning some of $\mathbf{x}$'s entries to zero with a certain probability. As a result, the objective becomes minimizing the reconstruction loss—that is,

$$\min_{\theta_1,\theta_2} \sum_{i=1}^{n} L(\mathbf{x}^{(i)}, g_{\theta_2}(f_{\theta_1}(\tilde{\mathbf{x}}^{(i)}))), \tag{31}$$

where $f_{\theta_1}(\cdot)$ denotes an encoding function, $g_{\theta_2}(\cdot)$ denotes a decoding function, and the ith instance and its corrupted form are denoted by $\mathbf{x}^{(i)}$ and $\tilde{\mathbf{x}}^{(i)}$, respectively.

Struc2vec [70] insists that node identity similarity should be independent of network position and neighborhoods' labels. To well preserve the property, the input network G is first preprocessed into a context graph M, which is a multi-layer weighted graph described in Section 3.1.1. Then a biased random walk process is conducted to produce node traveling sequences. Each walker chooses its next step on the same layer or across different layers and the choosing probabilities are proportional to edge weights so that the structurally similar nodes are more likely to be visited. After having samples, Struc2vec uses skip-gram to train the learning model and hierarchical softmax is leveraged to minimize the complexity.

4.1.2 Community Structure. MRF [48] proposes a structured pairwise Markov random field framework. To ensure the global coherent community structure, MRF adopts the Gibbs distribution to measure the posterior probability $P(C_p|\mathbf{A})$ of community partition C_p given network adjacency matrix $\mathbf{A}$ and maximize $P(C_p|\mathbf{A})$ by optimizing a well-designed energy function $E(C_p;\mathbf{A})$. $E(C_p;\mathbf{A})$ consists of two parts: a group of unary potentials that are used to enforce node representations playing a dominant role and a group of pairwise potentials that are used to fine-tune the obtained unary potentials based on the connections of nodes. In addition, the Gaussian mixture model is utilized to approximate the probability distributions of all nodes belonging to various communities.

To capture the overlapping community structure, Epasto and Perozzi [27] propose a multi-embedding method, splitter, to learn multiple vectors for each node, representing each one's involvement in various communities. To exploit the multi-community participation, splitter preprocesses the original network into a persona graph G_p (see Section 3.1.1) having multiple personas for every node. But the persona graph may consist of many disconnected components that pose challenges for representation learning. To address the challenge, splitter adds virtual edges between each persona and its parent node in the original network so as to enforce that every original node $\mathrm{v_o}$ can be predicted given its arbitrary persona $\mathrm{v_i}$'s representation—that is, $\Pr(\mathrm{v_o}|\Phi_{G_p}(\mathrm{v_i}))$. Like DeepWalk, splitter uses the skip-gram model coupled with hierarchical softmax to learn node representations. By introducing a regularization parameter λ, the optimization objective is formalized as follows:

$$\arg\min_{\Phi_{G_p}} -\log \Pr(\{\mathrm{v_{i-w}}, \ldots, \mathrm{v_{i+w}}\} \backslash \mathrm{v_i} | \Phi_{G_p}(\mathrm{v_i})) - \lambda \Pr(\mathrm{v_o}|\Phi_{G_p}(\mathrm{v_i})). \tag{32}$$

vGraph [76] presents a probabilistic generative model where each node v is viewed as a mixture of multiple communities and represented by a multinomial distribution over communities C (i.e., $p(C|\mathrm{v})$), and meanwhile, each community C can be represented by a distribution over nodes (i.e., $p(\mathrm{v}|C)$). vGraph casts the edge generation process as an inference problem. Specifically, for each node v, vGraph first draws a community assignment C from $p(C|\mathrm{v})$ and then generates an edge $\mathrm{e_{vu}}$ by drawing another node u according to distribution $p(u|\mathrm{C})$. Both types of distributions are

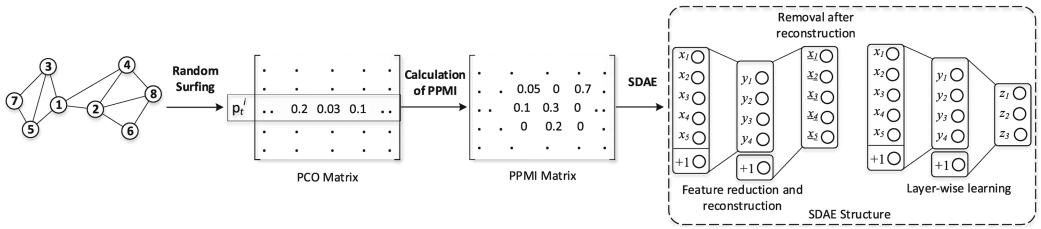

Fig. 6. The framework of DNGR (adapted from Figure 1 [15]).

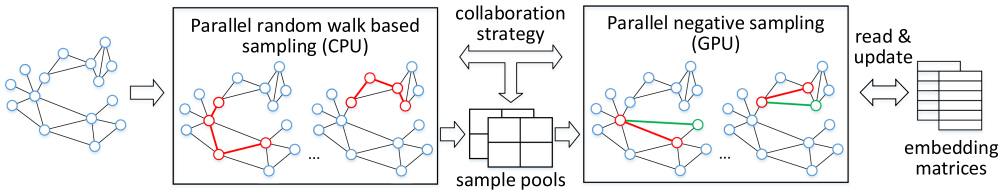

Fig. 7. Overview of the CPU-GPU hybrid system (adapted from Figure 1 [110]).

parameterized by the representations of nodes and communities. In addition, a smoothness regularizer is borrowed by vGraph to ensure the community memberships of neighborhoods to be similar.

4.2 Distributed NRL Methods

Considering the recent advances in GPU-enabled neural network training, Zhu et al. [110] propose a CPU-GPU hybrid system, GraphVite, which is a hardware coupling system specifically designed for large-scale network embedding. To leverage distinct advantages of CPUs and GPUs, GraphVite focuses on the parallelization of instance sampling inside of feature extraction and node embedding, and its overview is shown in Figure 7. The random walk based sampling procedure involves excessive random access and the random walks initiated from different starting nodes are independent, so the positive instance sampling is suitable for parallel execution on multiple CPUs. The node embedding procedure involves excessive matrix computation, which is the advantage of GPUs. Due to the limited GPU memory, it is impossible to place all sampling instances and parameter matrices of node embeddings on a GPU, and GraphVite organizes a grid sample pool. The pool is divided into multiple blocks, and each one corresponds to a subset of the original network. The training task is also divided into many subtasks, and they are assigned with varied blocks for training. Due to the sparse nature of networks, many block pairs are gradient exchangeable, which means that exchanging the order of gradient descent steps does not result in a vector difference. As a result, subtasks on different GPUs are capable of performing gradient updates currently in their own subsets without any synchronization. At the same time, the sampling pool is shared between CPUs and GPUs. To reduce the synchronization cost, GraphVite presents a collaboration strategy that maintains two sample pools in the main memory so that CPUs fill up a pool and pass it to GPUs, and they always work on different pools in a parallel fashion. GraphVite reports about 50 times faster than the current fastest system and takes around 20 hours to embed a network as large as 66 million nodes and 1.8 billion edges.

Lerer et al. [54] also present a distributed embedding system, **PyTorch-BigGraph (PBG)**, which incorporates several modifications to traditional embedding systems so as to allow it supports scale to multi-entity, multi-relation graphs with billions of nodes and trillions of edges. To implement

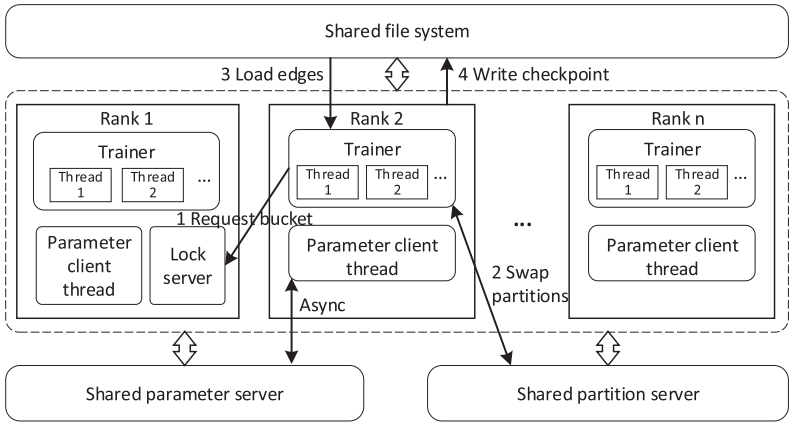

Fig. 8. A block diagram of the modules used for PBG's distributed mode (adapted from Figure 2 [54]).

parallelization, PBG partitions the adjacency matrix of the original graph into multiple buckets and feeds the edge samples from buckets distributively. An example of the training of one bucket upon rank 2 is illustrated in Figure 8. The trainer requests a bucket from the lock server residing on rank 1, where all partitions of that bucket are locked. Then the trainer saves the no longer used partitions and loads new partitions that it needs to and from the shared partition servers, at which point it drops its old partitions left on the lock server. Edge samples are loaded from a shared filesystem, and the training is performed inside multi-thread with no inter-thread synchronization required. In a single thread, only a small fraction of shared parameters will be synchronized with a shared parameter server. Note that checkpoints will be occasionally written back to the filesystem from trainers. A distributed execution of PBG on eight machines achieves 4x speedup and reduces memory consumption by 88%.

4.3 Multi-NRL Methods

Compared to single network embedding, multiple networks may contain complementary information and multi-network embedding can produce better representations. Similar examples can be observed in many fields like social networks and bioinformatics, and a node in one network may be associated with multiple nodes in another network that forms a many-to-many mapping relationship between networks. Ni et al. [64] propose a **deep multi-network embedding (DMNE)** method, which coordinates different neural networks with one co-regularized loss to manipulate cross-network correlations. An example of DMNE for two networks is illustrated in Figure 9, where $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ are network contexts derived from the two networks, $\mathbf{S}^{(12)}$ and $\mathbf{S}^{(21)}$ denote the cross-network relationship matrix, and $(\mathbf{H}^{(\mathbf{i})})_\mathbf{l}$ corresponds to the matrix of representations for all nodes in the lth layer of network i. For single network embedding, say network i, the neural network consists of L_i+1 layers built in the autoencoder fashion, where half of hidden layers act as encoders to learn node representations, whereas the others are decoders in charge of the reconstruction of input. To achieve the network information complementarity, DMNE relies on the intuition that a node's representation should be similar to the representation of its mapped node in another network, and introduces the cross-network regularization based on two kinds of loss functions: **embedding disagreement (ED)** loss function and **proximity disagreement (PD)** loss function. The former is for situation that all networks have an identical embedding dimension, whereas the latter is more flexible without dimension constraint. As a result, the unified objective is represented

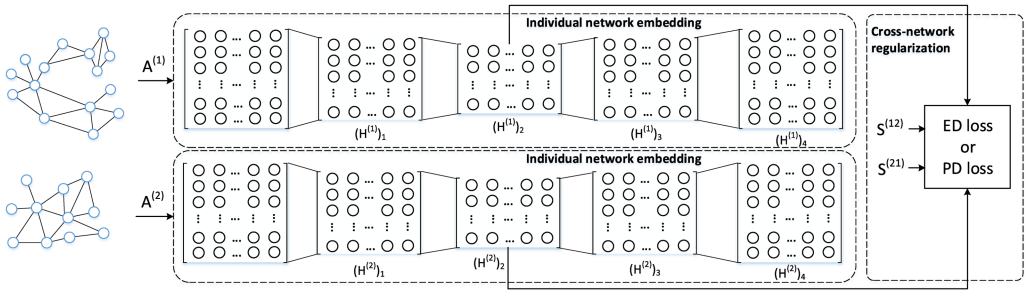

Fig. 9. The DMNE framework for two networks (adapted from Figure 2 [64]).

by

$$\arg\min_{\theta^{(i)},\mathbf{U}^{(i)}} \sum_{i=1}^{g} L_{ae}^{(i)} + \alpha \sum_{(i,j)\in I}^{g} L_R^{(i)} + \beta \sum_{i=1}^{g} \left\| \mathbf{U}^{(i)} - \mathbf{H}^{(i)} \right\|_F^2, \tag{33}$$

where $\mathbf{U}^{(i)}$ denotes network i's representation matrix, $\theta^{(i)}$ denotes the weight matrix, $L_{ae}^{(i)}$ refers to the loss of network i, $L_R^{(ij)}$ can be either ED loss or PD loss, and α and β are trade-off factors.

TransLink [108] models social network alignment as a link prediction between different networks and aims to find out all pairs of user accounts with the same identities and connects them via anchor links. TransLink embeds users and links between uses into a latent space by incorporating both network structure and user interaction meta-path, then iteratively predicts the potential anchor links between users who have similar representations over translative operations. Du et al. [26] investigate a joint framework, CENALP, that couples link prediction and network alignment together. The newly predicted links enrich the network structure information, and new potential nodes with richer features will have a greater chance to be identified and vice versa. Therefore, link prediction and network alignment are allowed to work in a mutually beneficial way. CENALP relies on a biased random walk to generate samples across different networks. For each training sample (v_i, v_j), CENALP adopts a product layer to decode the latent interaction between two nodes and sets up a logistic regression layer for prediction.

4.4 Dynamic NRL Methods

The methods introduced so far mainly focus on exploring NRL for static networks. However, real-world networks like social networks and biological networks are dynamically evolving over time, which means that not all nodes are available during the training. Inductive learning is an effective way to support the representation learning for unseen nodes. As the first inductive framework, GraphSage [41] derives from the idea of spatial GCN to generate node embeddings. It designs an aggregation function to essentially assemble features from each node's neighborhood, such as text attributes and node degrees. The aggregation encoder is defined as follows:

$$\mathbf{h}_v^k = \sigma\left(\mathbf{W}^k \cdot AGG_k\left(\mathbf{h}_v^{k-1}, \left\{\mathbf{h}_u^{k-1}, \forall u \in N_v\right\}\right)\right), \tag{34}$$

where the function $AGG_k(\cdot,\cdot)$ should be symmetric so as to ensure the framework can be trained and applied to neighborhood feature sets with arbitrary order. GraphSage examines three types of functions: the mean aggregator, LSTM aggregator, and pooling aggregator. Note that LSTM is not inherently permutation invariant, so it is applied to a random permutation of neighborhoods for implementation. The learning procedure of GraphSage is illustrated by an example in Figure 10. For each node to be embedded, GraphSage first samples a fixed number of neighborhoods within k hops, then obtains the central node's state by aggregating its neighborhoods' features, and finally

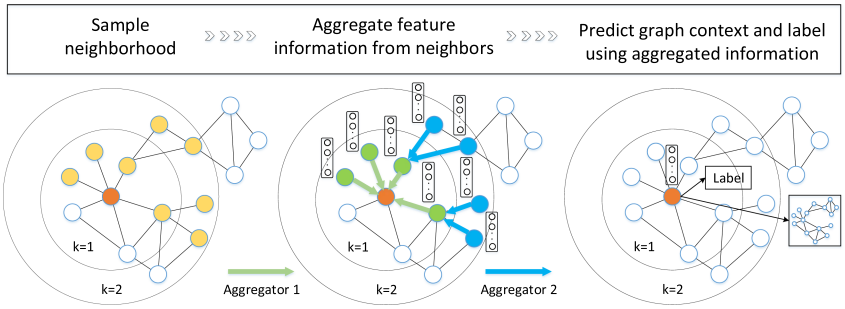

Fig. 10. An example illustrating the three-step learning procedure of GraphSage (adapted from Figure 1 [41]).

makes predictions and backpropagates errors based on the node's state. To learn useful predictive representations, negative sampling is introduced to enhance performance, then a graph-based loss function is given by

$$L(\mathbf{z}_u) = -\log\left(\sigma\left(\mathbf{z}_u^T \mathbf{z}_v\right)\right) - Q \cdot E_{v_n \sim P_n(v)}\left[\log \sigma\left(-\mathbf{z}_u^T \mathbf{z}_n\right)\right], \tag{35}$$

where $\mathbf{z_u}$ denotes the representation of node v_u, Q denotes the amount of negative samples and P_n represents the negative sampling distribution. As an inductive representation learning method, GraphSage enforces that every representation v_u fed into the preceding function is generated from the features of node u's neighborhoods rather than a unique representation trained by an embedding look-up.

Graph2Gauss (G2G) [8] proposes a novel unsupervised inductive learning method that embeds nodes in (un)directed attributed graphs into Gaussian distributions rather than conventional vector representations so that it can capture the uncertainty of each node's embedding. G2G first uses a deep encoder to yield the parameters associated with the node's embedding distribution from the node attributes. The mean and diagonal representations for a node are learned as functions of the node attributes. After that, G2G optimizes a ranking loss function that incorporates the ranking of nodes derived from the network structure. Given an anchor node v, nodes at distance 1 of v are closer than nodes at distance 2, and so on. The distance between node representations is measured by the asymmetric KL divergence. The ranking loss is a square-exponential loss proposed in energy-based models. To avoid the situation that low-degree nodes are less often updated, G2G presents a node-anchored sampling method. For each node v, the method randomly samples one other node from its neighborhoods and then optimizes over all the corresponding pairwise constraints. To enable the inductive learning for unseen nodes, G2G passes the attributes of these unseen nodes through the learned deep encoder.

To incorporate global structural information, SPINE [37] proposes a structural identity preserved method. Rooted PageRank matrix $\mathbf{S^{RPR}}$ is used as the indicator of pairwise proximity where $\mathbf{S_i^{RPR}}$ represents v_i's global feature. As an inductive method, the length of a node's structural description should be independent of the network size so that the structural feature $\mathbf{T_i}$ of v_i is defined as the top-k values of $\mathbf{S_i^{RPR}}$. Given the content feature matrix $\mathbf{F_i^k}$ of k nodes, node v_i's embedding is generated by aggregating the k vectors with respect to the corresponding weights in $\mathbf{T_i}$. To encode the similarity in terms of both structural identities and local proximities simultaneously, a biased positive sampling strategy is proposed to collaborate with negative sampling.

Zhou et al. [109] propose a dynamic NRL method, DynamicTriad, to preserve both structural information and evolution patterns during the learning process. The evolution of a dynamic network is described by a series of static network snapshots over discrete time. A triad of three nodes

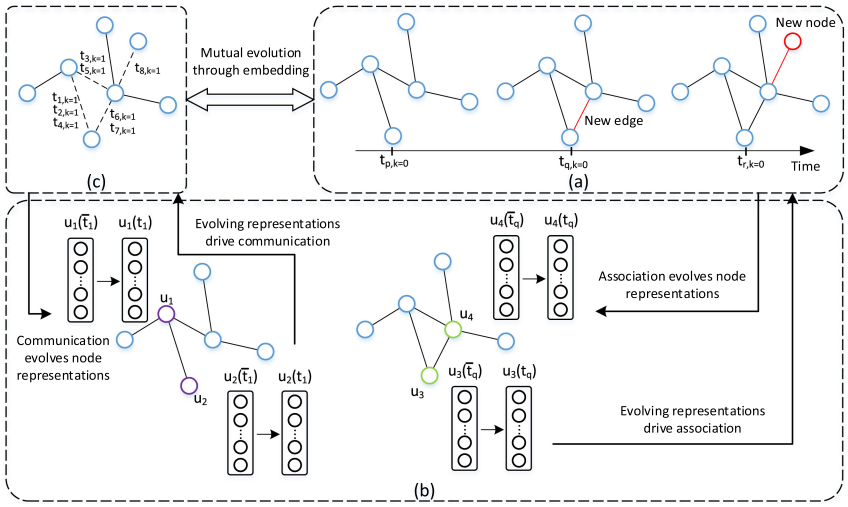

Fig. 11. An illustration of evolution for representation learning. (a) Association events (k = 0). (b) Evolving representations. (c) Communication events (k = 1). (Adapted from Figure 1 [82].)

works as the basic units of networks, and its closure process is used to capture the network dynamics. Specifically, the triad closure process reflects how a closed triad develops from an open triad over time by means of an evolutionary probability, where a closed triad is a complete graph of three nodes while an open triad misses a connection between any two nodes. For sampling, considering the expensive computation on the combination of positive and negative samples, DynamicTriad adopts an idea of sample corruption to generate negative samples by replacing nodes in a positive triad with varied nodes. In addition, DynamicTriad enforces the evolutionary smoothness by minimizing the distance of node representations in adjacent timestamps. In the dynamic settings, nodes not seen at the current timestamp are regarded as out-of-sample nodes and their embeddings can be inferred by exploring the idea of inductive learning. To ensure the inferred embeddings preserve intricate network properties, DepthLGP [58] designs a high-order Laplacian Gaussian process to encode these properties and employs a DNN to learn a non-linear transformation from the high-order Laplacian Gaussian process.

The preceding work assumes that network dynamics evolve at a single time scale; however, the evolution of real-world networks usually exhibits multiple time scales. DyRep [82] uses two distinct dynamic processes to model the dynamic evolution: the association process describes dynamics of the network, which brings structural changes caused by nodes and edges and results in long-lasting information flow associated with them, and the communication process reflects dynamics on the network, which relates to the activities between connected or non-connected nodes and results in temporary information exchange across them. As shown in Figure 11, the dynamic evolution of a network is given by a stream of events that involve both processes. To model the occurrence of event $p = (\mathrm{u}, \mathrm{v}, \mathrm{t_p}, \mathrm{k})$ between u and v at time $\mathrm{t_p}$, a conditional intensity function is defined based on the temporal point process—that is,

$$\lambda_k^{\mathrm{u,v}}(t_p) = f_k\left(g_k^{\mathrm{u,v}}(\bar{t}_p)\right), \tag{36}$$

where $\bar{t}_p$ is the timestamp before the current event and $f_k(\cdot)$ is a parameterized softplus function designed for capturing the timescale dependence. Once an event p occurs, the representation of each involved node will be updated via a deep recurrent neural network (RNN)—for example, node

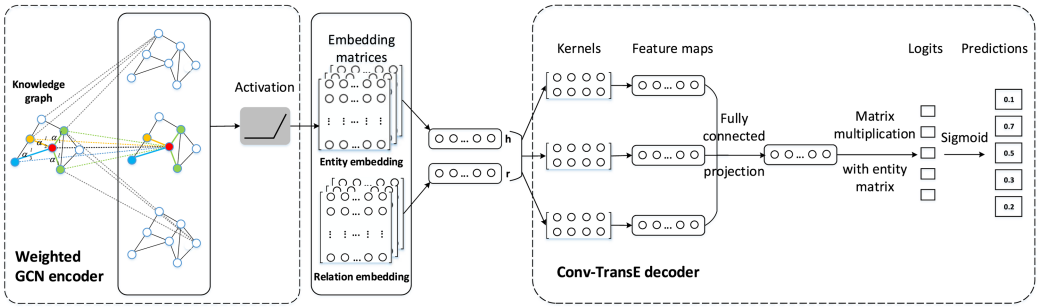

Fig. 12. An illustration of SACN (adapted from Figure 1 [74]).

v's representation $\mathbf{z}^{\mathrm{v}}(\mathbf{t_p})$ at current timestamp t_p is updated by

$$\mathbf{z}^{\mathrm{v}}(t_p) = \sigma\left(\underbrace{\mathbf{W}^{struct}\mathbf{h}^{\mathrm{u}}_{struct}(\bar{t}_p)}_{\text{Localized embedding propagation}} + \underbrace{\mathbf{W}^{rec}\mathbf{z}^{\mathrm{v}}(\bar{t}^{\mathrm{v}}_p)}_{\text{Self-propagation}} + \underbrace{\mathbf{W}^{t}(t_p - \bar{t}^{\mathrm{v}}_p)}_{\text{Exogenous drive}}\right), \tag{37}$$

where $\mathbf{h}^{\mathrm{u}}_{struct}$ is the representation aggregated from node u's neighborhoods, $\mathbf{z}^{\mathrm{v}}(\bar{t}^{\mathrm{v}}_p)$ is the recurrent state obtained from v's previous representation, and $\bar{t}^{\mathrm{v}}_p$ denotes the timestamp of v's previous event. Parameter matrices $\mathbf{W}^{struct}$, $\mathbf{W}^{rec}$, and $\mathbf{W}^{t}$ are applied to control the aggregate effect of three inputs. For a set of observed events, the learning objective is defined as negative log likelihood. Specifically, DyRep borrows the idea of GATs [86] that is the attention mechanism applied to graph data and uses it to endue neighborhoods with varied attention coefficients that also evolve over time.

4.5 Knowledge Graph Representation Learning

As a special type of networks, the knowledge graph can be viewed a structured representation of facts, denoted by a set of triples. Each triple consists of two entities h, t and a relation r between them (e.g., (h, r, t)). KRL is to map entity nodes and relation edges into low-dimensional vectors while capturing their semantic meanings [47]. The primary goal of KRL is to improve the plausibility of facts where the plausibility can be viewed as a specialization of network property. Distance proximity is often used to enhance the plausibility of a fact by minimizing the distance between entities of the corresponding triple. In the classic translating embedding model [10], the relation r corresponds to a translation from head entity to tail entity and their embeddings $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ hold when (h, r, t) is a fact. RotatE [78] expands the embedding space from real-valued space to complex space where entities and relations are mapped to low-dimensional complex vectors and each relation corresponds to a rotation from head entity to tail entity—that is,

$$\mathbf{t} \approx \mathbf{h} \circ \mathbf{r}, \tag{38}$$

where $\circ$ denotes the Hadamard product. For each element in the embedding space, $t_i \approx h_i \circ r_i$ and $|r_i| = 1$. Element r_i is of the form $e^{i\theta_{r,i}}$, which corresponds to a counterclockwise rotation by $\theta_{r,i}$ radians about the origin of the complex plane. For each triple (h, r, t), its distance is defined as

$$d_r(\mathbf{h}, \mathbf{t}) = \|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|. \tag{39}$$

The complex space allows RotatE to capture more relation patterns including symmetry/antisymmetry, inversion, and composition. During the training, traditional negative sampling samples the negative triples in a uniform way, which suffers the problem of inefficiency since

many samples are obviously false and cannot provide meaningful information. RotatE presents a variant called *self-adversarial negative sampling*, which samples negative triples from the following distribution:

$$p(h'_j, r, t'_j | \{(h_i, r_i, t_i)\}) = \frac{\exp \alpha d_r(\mathbf{h'}_j, \mathbf{t'}_j)}{\sum_i \exp \alpha d_r(\mathbf{h'}_i, \mathbf{t'}_i)}, \tag{40}$$

where α is the temperature of sampling. The above probability is treated as the weight of the negative sample. The negative sampling loss is defined as

$$L = -\log \sigma(\gamma - d_r(\mathbf{h}, \mathbf{t})) - \sum_i p(h'_i, r, t'_i) \log \sigma(d_r(\mathbf{h'}_i, \mathbf{t'}_i) - \gamma), \tag{41}$$

where γ is a fixed margin, σ is the sigmoid function, and (h'_i, r, t'_i) is the ith negative triple.

SACN [74] proposes an end-to-end structure-aware convolutional network to predict new triples for knowledge graph completion. As illustrated in Figure 12, SACN presents a weighted GCN as the encoder to learn entities' representations by aggregating connected entities as specified by the relations in the knowledge graph, where the weighted GCN weights the different types of relations differently by defining the interaction strength (e.g., α_t denotes the strength of relation type t) so that the amount of information from neighboring nodes used in aggregation can be controlled. Let $\mathbf{h}_i^l$ be the input vector of node $\mathrm{v_i}$ in the lth layer, and its output vector is given by

$$\mathbf{h}_i^{l+1} = \sigma\left(\sum_{\mathrm{v}_j \in N_i} \alpha_t^l g\left(\mathbf{h}_i^l, \mathbf{h}_j^l\right)\right), \tag{42}$$

where N_i is the neighbor set of node v_i and g is the aggregation function. With node embeddings as the input, the decoder aims to represent the relations more accurately by recovering the original triples. Based on ConvE [24], SACN develops Conv-TransE as the decoder where the translation fashion of TransE is incorporated. Conv-TransE aligns the convolutional outputs of both entity and relation embeddings with all kernels and yields a matrix $\mathbf{M}(\mathbf{h}, \mathbf{r})$. Both the encoder and decoder are jointly trained by minimizing the cross entropy between $\mathbf{h} + \mathbf{r}$ and $\mathbf{t}$ to preserve the distance proximity and the objective function is defined as

$$d_r(\mathbf{h}, \mathbf{t}) = f(vec(\mathbf{M}(\mathbf{h}, \mathbf{r}))\mathbf{W})\mathbf{t}, \tag{43}$$

where f is a non-linear function and W is a matrix for the linear transformation. $vec(\mathbf{M})$ corresponds to a reshaping operation that changes feature map matrix to a vector.

4.6 Discussions and Open Challenges

The past half-decade has witness the rapid development of NRL, and meanwhile NRL research also faces multiple challenges, among which we select some promising challenges with high attention to discuss, hoping to provide useful guidance for future study.

4.6.1 Automated Learning. We have shown that different NRL approaches tend to exhibit different performances for different scenarios. There is no single method that is the winner for all scenarios. When one method is considered to be better than another, it typically depends on which benchmark is used, including tasks and datasets available for evaluation. According to our framework, proper combinations of candidate methods, models, and techniques at different stages of NRL may have greater potential to boost the performance of NRL. For example, Struc2vec [70] preprocesses the input graph into a context graph, generates samples via a biased random walk method, and learns representations by leveraging the skip-gram model as well as the negative sampling technique to reflect features with respect to node identities. It ranks the second-best model for node classification on Wikipedia [1]. Most of the existing NRL approaches follow the data-driven

Table 6. Summary of Recent Advances in NRL

Category	NRL Work	Data Preprocessing Method, Input Data	Network Feature Extraction		Node Embedding	
			Network Property	Sampling Method	Embedding Model	Optimization Technique
Conventional NRL	MVE [69]	-, Adjacency matrix	Multi-view 1st-order proximities	Edge sampling	Skip-gram model	Attention mechanism, negative sampling
	SDNE [87]	Matrix-based processing, Adjacency matrix, Laplacian matrix	1st- and 2nd-order proximity	Fetch from adjacency matrix	Deep autoencoder, Laplacian eigenmaps	Deep belief network
	DNGR [15]	Random surfing, PCO matrix	High-order proximity	Fetch from PPMI matrix	SDAE	Negative sampling
	Struc2vec [70]	Graph decomposition, context graph	Node identity proximity	Biased random walk	Skip-gram model	Hierarchical softmax
	MRF [48]	Matrix-based processing, transition matrix	Community structure	Random pairwise sampling	Markov random field	Gaussian mixture model
	splitter [27]	Graph decomposition, persona graph	Overlapping community structure	Random walk	Skip-gram model	Hierarchical softmax
	vGraph [76]	-, Adjacency matrix	Community structure	Edge sampling	Probabilistic generative model	Negative sampling
Distributed NRL	GraphVite [110]	-, Adjacency matrix	1st- and 2nd-order proximity	Parallel random walk	Edge probabilistic model	Parallel negative sampling
	PBG [54]	Graph decomposition, multi-bucket	1st-order proximity	Distributed edge sampling	Edge probabilistic model	Batched negative sampling
Multi-NRL	DMNE [64]	-, Adjacency matrix, cross-network relationship matrix	High-order proximity under many-to-many mapping	Random walk	Deep autoencoder	Negative sampling
	TransLink [108]	Meta-path extraction, meta-path, node state, edge state	High-order proximity under translations	Fetch from adjacency matrix	Translating embedding	Negative sampling
	CENALP [26]	-, Adjacency matrix	High-order proximity	Biased random walk	Skip-gram model	Negative sampling
Dynamic NRL	GraphSage [41]	Matrix-based processing, transition matrix, node state	2nd-order proximity	Random walk	Spatial GCN	Negative sampling
	Graph2Gauss [8]	-, Adjacency matrix, node state	High-order proximity	Node-anchored sampling	Deep autoencoder	Negative sampling
	SPINE [37]	Matrix-based processing, rooted PageRank matrix, node state	Node identity proximity	Random walk	Skip-gram model	Biased positive sampling, negative sampling
	DynamicTriad [109]	-, Adjacency matrix	1st-order proximity	Edge sampling	Edge probabilistic model	Negative sampling base on corruption
	DyRep [82]	-, Adjacency matrix, node state, edge state	2nd-order proximity	Random walk	GNN	Graph attention network, negative sampling
KRL	RotatE [78]	-, Adjacency matrix	Distance proximity	Fetch from adjacency matrix	Translating embedding	Self-adversarial negative sampling
	SACN [74]	-, Adjacency matrix, node state, edge state	Distance proximity	Fetch from adjacency matrix	Weighted GCN	Negative sampling

The hyphens represent no data preprocessing method available.

paradigm and are semi-automaticthat is, given a benchmark with the specified dataset, we have to decide the following. How should the data be preprocessed? What features should be preserved? What models and techniques can be used? How to optimize the chosen model? It is not a trivial task even for an expert to answer the preceding questions. Some recent efforts on meta-learning [45] hold potential to partially address these questions. The learning-to-learn mode allows meta-learning, being useful in multi-task scenarios where task-agnostic knowledge is learned from a family of tasks and used to improve learning of new tasks from the same family. AutoNE [85] presents an automatic hyperparameter optimization, the principle behind which is transferring the knowledge regarding optimized hyperparameters from multiple subnetworks to the original network.

Nevertheless, implementing a fully automated learning of network representation remains an open problem.

4.6.2 Proximity vs. Distinguishability. Observed from Table 6, we find that most of the current studies prefer to put feature proximity preserving as the guideline to learn node representations so as to ensure that structurally similar nodes having similar representations. As a result, application tasks like node classification and social recommendation benefit from this proximity-driven embedding. On the contrary, enhancing proximity reduces the distinguishability of nodes at the same time. Here the distinguishability refers to that an arbitrary node is significantly different from others in the representation space even though there is structural proximity between nodes. The enforced proximity imposes adverse impacts on some other application tasks. For example, to accurately identify anchor users between different social networks, it is desirable to ensure that each node's representation should be explicitly separated from the representations of its proximities. Otherwise, it would be hard to discriminate the anchor nodes from their structurally similar neighborhoods. **Generative adversarial networks (GANs)** [35] may provide a possible way to enhance distinguishability—for example, dNAME [107] presents a GAN mechanism to learn the latent space of single network combined with a graph kernel based regularizer for discriminating anchor nodes from others. The GAN's function depends on capturing true data distribution. How to make use of GANs to finely distinguish each node while preserving proximity is still exploration. GCNs [11] generate node representations by combining their own features with features aggregated from their neighborhoods. A well-designed convolution operation can benefit the trade-off between proximity and distinguishability. Today, many real-world networks can be classified as scale-free networks, such as airline networks and co-authorship networks. Scale-free networks follow power-law distributions, and most of the network nodes have few neighbors; on the contrary, a tiny fraction of network nodes have huge numbers of neighbors. Can we design a proper preprocessing method to deal with such extreme imbalance and to keep the structural features of different nodes at the same time? The receptive field for convolution largely depends on the adopted sampling method. How to design a sampling method to collaborate with the preprocessing method and embedding models to handle this challenge still needs to be explored.

4.6.3 Interpretability. Compared with handcrafted feature engineering, the superiority of NRL research has been empirically verified by both visualization and benchmarks, but fewer studies on NRL can give theoretically satisfactory answers to the following fundamental questions. What exactly latent features are learned from the network? What contributes to good performance on visualization and benchmarks? The connections between the underlying working mechanism and the performance results have not been well revealed yet. NetMF [68] makes a good attempt to create intrinsic connections between four NRL methods (e.g., DeepWalk [66], LINE [81], PTE [80], and node2vec [36]) and graph Laplacian by unifying them into the matrix factorization model. Gogoglou et al. [33] try to explore the association of the embedding space with both external and internal node categorization. Abductive learning [21] provides another beneficial attempt by unifying machine learning and logic programming, where the former learns to perceive primitive logic facts from data, whereas the latter is able to exploit symbolic domain knowledge and correct the wrongly perceived facts for improving learning results including interpretability. In short, to answer the preceding questions and ensure that the learned representations truly reflect the network information, more efforts need to be devoted to the interpretable learning in the future.

5 CONCLUSIONS

We have presented a comprehensive overview of the state-of-the-art NRL techniques through a unifying reference framework. We describe and compare different NRL approaches in terms of

the network data preprocessing methods, the network feature learning models, and the node embedding and optimization techniques. We also discuss open challenges of NRL research, including automated learning, trade-off between proximity, and distinguishability and interpretability. Our unified three phase reference framework offers a number of unique benefits. First, this framework presents a new perspective to reviewing the state-of-the-art NRL methods and technical optimizations employed. Second, the unified framework promotes reviewing the state-of-the-art NRL methods using three progressive and complementary phases. This new approach provides an in-depth understanding of NRL from three important perspectives: (1) raw input preprocessing, (2) the node feature extraction task and different methods utilized for improving and optimizing node feature extraction qualities, and (3) NRL models. Third, this unified three-phase framework by design serves dual purposes: (1) to help beginner readers and practitioners understand the end-to-end NRL workflow, and (2) to help researchers and graduate students who are interested in learning the state-of-the-art NRL gain a deeper understanding through reviewing each of the three phases in the NRL workflow.

REFERENCES

[1] Wikipedia. n.d. Node Classification on Wikipedia. Retrieved October 30, 2021 from https://paperswithcode.com/sota/node-classification-on-wikipedia.

[2] Sami Abu-El-Haija, Bryan Perozzi, and Rami Al-Rfou. 2017. Learning edge representations via low-rank asymmetric projections. In *Proceedings of CIKM'17*. 1787–1796.

[3] Sami Abu-El-Haija, Bryan Perozzi, Rami Al-Rfou, and Alexander A. Alemi. 2018. Watch your step: Learning node embeddings via graph attention. In *Proceedings of NeurIPS'18*. 9198–9208.

[4] Nesreen K. Ahmed, Ryan A. Rossi, John Boaz Lee, Xiangnan Kong, Theodore L. Willke, Rong Zhou, and Hoda Eldardiry. 2018. Learning role-based graph embeddings. *CoRR* abs/1802.02896 (2018). arXiv:1802.02896 http://arxiv.org/abs/1802.02896.

[5] Yunsheng Bai, Hao Ding, Yang Qiao, Agustin Marinovic, Ken Gu, Ting Chen, Yizhou Sun, and Wei Wang. 2019. Unsupervised inductive graph-level representation learning via graph-graph proximity. In *Proceedings of IJCAI'19*. 1988–1994.

[6] Mikhail Belkin and Partha Niyogi. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of NIPS'01*. 585–591.

[7] Mikhail Belkin and Partha Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* 15, 6 (2003), 1373–1396.

[8] Aleksandar Bojchevski and Stephan Günnemann. 2018. Deep Gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *Proceedings of ICLR'18*.

[9] Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. In *Proceedings of EMNLP'14*. 615–620.

[10] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of NeurIPS'13*. 2787–2795.

[11] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2014. Spectral networks and locally connected networks on graphs. In *Proceedings of ICLR'14*.

[12] John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics. *Behavior Research Methods* 39, 3 (2007), 510–526.

[13] HongYun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1616–1637.

[14] Shaosheng Cao, Wei Lu, and Qiongkai Xu. 2015. GraRep: Learning graph representations with global structural information. In *Proceedings of CIKM'15*. 891–900.

[15] Shaosheng Cao, Wei Lu, and Qiongkai Xu. 2016. Deep neural networks for learning graph representations. In *Proceedings of AAAI'16*. 1145–1152.

[16] Zhu Cao, Linlin Wang, and Gerard de Melo. 2018. Link prediction via subgraph embedding-based convex matrix completion. In *Proceedings of AAAI'18*. 2803–2810.

[17] Yukuo Cen, Xu Zou, Jianwei Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Representation learning for attributed multiplex heterogeneous network. In *Proceedings of KDD'19*. 1358–1368.

[18] Haochen Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2018. A tutorial on network embeddings. *CoRR* abs/1808.02590 (2018). arXiv:1808.02590 http://arxiv.org/abs/1808.02590.

[19] Haochen Chen, Bryan Perozzi, Yifan Hu, and Steven Skiena. 2018. HARP: Hierarchical representation learning for networks. In *Proceedings of AAAI'18*. 2127–2134.
[20] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. 2019. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering* 31, 5 (2019), 833–852.
[21] Wang-Zhou Dai, Qiu-Ling Xu, Yang Yu, and Zhi-Hua Zhou. 2019. Bridging machine learning and logical reasoning by abductive learning. In *Proceedings of NeurIPS'19*. 2811–2822.
[22] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of NeurIPS'16*. 3837–3845.
[23] Chenhui Deng, Zhiqiang Zhao, Yongyu Wang, Zhiru Zhang, and Zhuo Feng. 2020. GraphZoom: A multi-level spectral approach for accurate and scalable graph embedding. In *Proceedings of ICLR'20*.
[24] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2D knowledge graph embeddings. In *Proceedings of AAAI'18*. 1811–1818.
[25] Claire Donnat, Marinka Zitnik, David Hallac, and Jure Leskovec. 2018. Learning structural node embeddings via diffusion wavelets. In *Proceedings of KDD'18*. 1320–1329.
[26] Xingbo Du, Junchi Yan, and Hongyuan Zha. 2019. Joint link prediction and network alignment via cross-graph embedding. In *Proceedings of IJCAI'19*. 2251–2257.
[27] Alessandro Epasto and Bryan Perozzi. 2019. Is a single embedding enough? Learning node representations that capture multiple social contexts. In *Proceedings of WWW'19*. 394–404.
[28] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Yihong Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *Proceedings of WWW'19*. 417–426.
[29] Rui Feng, Yang Yang, Wenjie Hu, Fei Wu, and Yueting Zhang. 2018. Representation learning for scale-free networks. In *Proceedings of AAAI'18*. 282–289.
[30] Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. 2017. Protein interface prediction using graph convolutional networks. In *Proceedings of NeurIPS'17*. 6530–6539.
[31] Tianfan Fu, Cao Xiao, and Jimeng Sun. 2020. CORE: Automatic molecule optimization using copy & refine strategy. In *Proceedings of AAAI'20*. 638–645.
[32] Zheng Gao, Gang Fu, Chunping Ouyang, Satoshi Tsutsui, Xiaozhong Liu, Jeremy J. Yang, Christopher Gessner, et al. 2019. edge2vec: Representation learning using edge semantics for biomedical knowledge discovery. *BMC Bioinformatics* 20, 1 (2019), Article 306, 15 pages.
[33] Antonia Gogoglou, C. Bayan Bruss, and Keegan E. Hines. 2019. On the interpretability and evaluation of graph representation learning. *CoRR* abs/1910.03081 (2019).
[34] G. H. Golub and C. Reinsch. 1970. Singular value decomposition and least square solutions. *Numerische Mathematik* 14, 5 (1970), 403–420.
[35] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of NIPS'14*. 2672–2680.
[36] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of KDD'16*. 855–864.
[37] Junliang Guo, Linli Xu, and Jingchang Liu. 2019. SPINE: Structural identity preserved inductive network embedding. In *Proceedings of IJCAI'19*. 2399–2405.
[38] Shu Guo, Quan Wang, Lihong Wang, Bin Wang, and Li Guo. 2018. Knowledge graph embedding with iterative guidance from soft rules. In *Proceedings of AAAI'18*. 4816–4823.
[39] Michael Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research* 13 (2012), 307–361.
[40] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin* 40, 3 (2017), 52–74.
[41] William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of NeurIPS'17*. 1024–1034.
[42] David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. 2011. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis* 30, 2 (2011), 129–150.
[43] Peng Han, Peng Yang, Peilin Zhao, Shuo Shang, Yong Liu, Jiayu Zhou, Xin Gao, and Panos Kalnis. 2019. GCN-MF: Disease-gene association identification by graph convolutional networks and matrix factorization. In *Proceedings of KDD'19*. 705–713.
[44] Xiaofei He and Partha Niyogi. 2003. Locality preserving projections. In *Proceedings of NIPS'03*. 153–160.
[45] Timothy M. Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. 2020. Meta-learning in neural networks: A survey. *CoRR* abs/2004.05439 (2020).
[46] Sergey Ivanov and Evgeny Burnaev. 2018. Anonymous walk embeddings. In *Proceedings of ICML'18*. 2191–2200.

[47] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2020. A survey on knowledge graphs: Representation, acquisition and applications. *CoRR* abs/2002.00388 (2020).
[48] Di Jin, Xinxin You, Weihao Li, Dongxiao He, Peng Cui, Françoise Fogelman-Soulié, and Tanmoy Chakraborty. 2019. Incorporating network embedding into Markov random field for better community detection. In *Proceedings of AAAI'19*. 160–167.
[49] Junghwan Kim, Haekyu Park, Ji-Eun Lee, and U. Kang. 2018. SIDE: Representation learning in signed directed networks. In *Proceedings of WWW'18*. 509–518.
[50] Thomas N. Kipf and Max Welling. 2016. Variational graph auto-encoders. In *Proceedings of the NeurIPS'16 Workshop on BDL*.
[51] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of ICLR'17*.
[52] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of NeurIPS'12*. 1106–1114.
[53] Yi-An Lai, Chin-Chi Hsu, Wen-Hao Chen, Mi-Yen Yeh, and Shou-De Lin. 2017. PRUNE: Preserving proximity and global ranking for network embedding. In *Proceedings of NeurIPS'17*. 5257–5266.
[54] Adam Lerer, Ledell Wu, Jiajun Shen, Timothée Lacroix, Luca Wehrstedt, Abhijit Bose, and Alexander Peysakhovich. 2019. PyTorch-BigGraph: A large-scale graph embedding system. In *Proceedings of SysML'19*.
[55] Lizi Liao, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2018. Attributed social network embedding. *IEEE Transactions on Knowledge and Data Engineering* 30, 12 (2018), 2257–2270.
[56] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of AAAI'15*. 2181–2187.
[57] Ninghao Liu, Qiaoyu Tan, Yuening Li, Hongxia Yang, Jingren Zhou, and Xia Hu. 2019. Is a single vector enough? Exploring node polysemy for network embedding. In *Proceedings of KDD'19*. 932–940.
[58] Jianxin Ma, Peng Cui, and Wenwu Zhu. 2018. DepthLGP: Learning embeddings of out-of-sample nodes in dynamic networks. In *Proceedings of AAAI'18*. 370–377.
[59] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2017. The more you know: Using knowledge graphs for image classification. In *Proceedings of CVPR'17*. 20–28.
[60] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR'13*.
[61] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS'13*. 3111–3119.
[62] Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In *Proceedings of AISTATS'05*.
[63] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. 2017. graph2vec: Learning distributed representations of graphs. *CoRR* abs/1707.05005 (2017). http://arxiv.org/abs/1707.05005.
[64] Jingchao Ni, Shiyu Chang, Xiao Liu, Wei Cheng, Haifeng Chen, Dongkuan Xu, and Xiang Zhang. 2018. Co-regularized deep multi-network embedding. In *Proceedings of WWW'18*. 469–478.
[65] Larry Page, Sergey Brin, R. Motwani, and T. Winograd. 1998. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford University, Stanford, CA.
[66] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online learning of social representations. In *Proceedings of KDD'14*. 701–710.
[67] Bryan Perozzi, Vivek Kulkarni, Haochen Chen, and Steven Skiena. 2017. Don't walk, skip! Online learning of multi-scale network embeddings. In *Proceedings of ASONAM'17*. 258–265.
[68] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. 2018. Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and node2vec. In *Proceedings of WSDM'18*. 459–467.
[69] Meng Qu, Jian Tang, Jingbo Shang, Xiang Ren, Ming Zhang, and Jiawei Han. 2017. An attention-based collaboration framework for multi-view network representation learning. In *Proceedings of CIKM'17*. 1767–1776.
[70] Leonardo Filipe Rodrigues Ribeiro, Pedro H. P. Saverese, and Daniel R. Figueiredo. 2017. *Struc2vec*: Learning node representations from structural identity. In *Proceedings of KDD'17*. 385–394.
[71] Ryan A. Rossi, Rong Zhou, and Nesreen K. Ahmed. 2018. Deep inductive network representation learning. In *Proceedings of WWW'18*. 953–960.
[72] Benedek Rozemberczki, Ryan Davies, Rik Sarkar, and Charles A. Sutton. 2019. GEMSEC: Graph embedding with self clustering. In *Proceedings of ASONAM'19*. 65–72.
[73] Benedek Rozemberczki and Rik Sarkar. 2020. Fast sequence-based embedding with diffusion graphs. *CoRR* abs/2001.07463 (2020). arXiv:2001.07463 https://arxiv.org/abs/2001.07463.

[74] Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. 2019. End-to-end structure-aware convolutional networks for knowledge base completion. In *Proceedings of AAAI'19*. 3060–3067.
[75] Bing-Jie Sun, Huawei Shen, Jinhua Gao, Wentao Ouyang, and Xueqi Cheng. 2017. A non-negative symmetric encoder-decoder approach for community detection. In *Proceedings of CIKM'17*. 597–606.
[76] Fan-Yun Sun, Meng Qu, Jordan Hoffmann, Chin-Wei Huang, and Jian Tang. 2019. vGraph: A generative model for joint community detection and node representation learning. In *Proceedings of NeurIPS'19*. 512–522.
[77] Yiwei Sun, Suhang Wang, Tsung-Yu Hsieh, Xianfeng Tang, and Vasant G. Honavar. 2019. MEGAN: A generative adversarial network for multi-view network embedding. In *Proceedings of IJCAI'19*. 3527–3533.
[78] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge graph embedding by relational rotation in complex space. In *Proceedings of ICLR'19*.
[79] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. 2013. Deep neural networks for object detection. In *Proceedings of NeurIPS'13*. 2553–2561.
[80] Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. PTE: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of KDD'15*. 1165–1174.
[81] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale information network embedding. In *Proceedings of WWW'15*. 1067–1077.
[82] Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. 2019. DyRep: Learning representations over dynamic graphs. In *Proceedings of ICLR'19*.
[83] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, and Emmanuel Müller. 2018. VERSE: Versatile graph embeddings from similarity measures. In *Proceedings of WWW'18*. 539–548.
[84] Ke Tu, Peng Cui, Xiao Wang, Philip S. Yu, and Wenwu Zhu. 2018. Deep recursive network embedding with regular equivalence. In *Proceedings of KDD'18*. 2357–2366.
[85] Ke Tu, Jianxin Ma, Peng Cui, Jian Pei, and Wenwu Zhu. 2019. AutoNE: Hyperparameter optimization for massive network embedding. In *Proceedings of the 25th ACM SIGKDD'19*. 216–225.
[86] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of ICLR'18*.
[87] Daixin Wang, Peng Cui, and Wenwu Zhu. 2016. Structural deep network embedding. In *Proceedings of KDD'16*. 1225–1234.
[88] Hongwei Wang, Jia Wang, Jialin Wang, Miao Zhao, Weinan Zhang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. GraphGAN: Graph representation learning with generative adversarial nets. In *Proceedings of AAAI'18*. 2508–2515.
[89] Suhang Wang, Jiliang Tang, Charu C. Aggarwal, Yi Chang, and Huan Liu. 2017. Signed network embedding in social media. In *Proceedings of SDM'17*. 327–335.
[90] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. 2017. Community preserving network embedding. In *Proceedings of AAAI'17*. 203–209.
[91] Zekai Wang, Hongzhi Liu, Yingpeng Du, Zhonghai Wu, and Xing Zhang. 2019. Unified embedding model over heterogeneous information network for personalized recommendation. In *Proceedings of IJCAI'19*. 3813–3819.
[92] Wenqi Wei, Qi Zhang, and Ling Liu. 2021. Bitcoin transaction forecasting with deep network representation learning. *IEEE Transactions on Emerging Topics in Computing* 9, 3 (2021), 1359–1371.
[93] Jun Wu, Jingrui He, and Jiejun Xu. 2019. DEMO-Net: Degree-specific graph neural networks for node and graph classification. In *Proceedings of KDD'19*. 406–415.
[94] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144 (2016).
[95] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2019. A comprehensive survey on graph neural networks. *CoRR* abs/1901.00596 (2019).
[96] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How powerful are graph neural networks? In *Proceedings of ICLR'19*.
[97] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y. Chang. 2015. Network representation learning with rich text information. In *Proceedings of IJCAI'15*. 2111–2117.
[98] Shuang Yang and Bo Yang. 2018. Enhanced network embedding with text information. In *Proceedings of ICPR'18*. 326–331.
[99] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of ICLR'18*.
[100] Lei Yu, Qi Zhang, Donna Dillenberger, Ling Liu, Calton Pu, Ka Ho Chow, Mehmet Emre Gursoy, Stacey Truex, Hong Min, Arun Iyengar, and Gong Su. 2019. GRAHIES: Multi-scale graph representation learning with latent hierarchical structure. In *Proceedings of CogMI'19*. 8–15.

[101] Pingpeng Yuan, Pu Liu, Buwen Wu, Hai Jin, Wenya Zhang, and Ling Liu. 2013. TripleBit: A fast and compact system for large scale RDF data. *Proceedings of the VLDB Endowment* 6, 7 (2013), 517–528.
[102] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. 2018. SINE: Scalable incomplete network embedding. In *Proceedings of ICDM'18*. 737–746.
[103] Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. In *Processings of NeurIPS'18*. 5171–5181.
[104] Yingxue Zhang, Soumyasundar Pal, Mark Coates, and Deniz Üstebay. 2019. Bayesian graph convolutional neural networks for semi-supervised classification. In *Proceedings of AAAI'19*. 5829–5836.
[105] Ziwei Zhang, Peng Cui, Xiao Wang, Jian Pei, Xuanrong Yao, and Wenwu Zhu. 2018. Arbitrary-order proximity preserved network embedding. In *Proceedings of KDD'18*. 2778–2786.
[106] Zhen Zhang, Hongxia Yang, Jiajun Bu, Sheng Zhou, Pinggang Yu, Jianwei Zhang, Martin Ester, and Can Wang. 2018. ANRL: Attributed network representation learning via deep neural networks. In *Proceedings of IJCAI'18*. 3155–3161.
[107] Fan Zhou, Zijing Wen, Goce Trajcevski, Kunpeng Zhang, Ting Zhong, and Fang Liu. 2019. Disentangled network alignment with matching explainability. In *Proceedings of INFOCOM'19*. 1360–1368.
[108] Jingya Zhou and Jianxi Fan. 2019. TransLink: User identity linkage across heterogeneous social networks via translating embeddings. In *Proceedings of INFOCOM'19*. 2116–2124.
[109] Le-Kui Zhou, Yang Yang, Xiang Ren, Fei Wu, and Yueting Zhuang. 2018. Dynamic network embedding by modeling triadic closure process. In *Proceedings of AAAI'18*. 571–578.
[110] Zhaocheng Zhu, Shizhen Xu, Jian Tang, and Meng Qu. 2019. GraphVite: A high-performance CPU-GPU hybrid system for node embedding. In *Proceedings of WWW'19*. 2494–2504.

Received May 2020; revised June 2021; accepted October 2021