Representation Learning on Graphs: Methods and Applications

William L. Hamilton
wleif@stanford.edu

Rex Ying
rexying@stanford.edu

Jure Leskovec
jure@cs.stanford.edu

Department of Computer Science
Stanford University
Stanford, CA, 94305

Abstract

Machine learning on graphs is an important and ubiquitous task with applications ranging from drug design to friendship recommendation in social networks. The primary challenge in this domain is finding a way to represent, or encode, graph structure so that it can be easily exploited by machine learning models. Traditionally, machine learning approaches relied on user-defined heuristics to extract features encoding structural information about a graph (e.g., degree statistics or kernel functions). However, recent years have seen a surge in approaches that automatically learn to encode graph structure into low-dimensional embeddings, using techniques based on deep learning and nonlinear dimensionality reduction. Here we provide a conceptual review of key advancements in this area of representation learning on graphs, including matrix factorization-based methods, random-walk based algorithms, and graph convolutional networks. We review methods to embed individual nodes as well as approaches to embed entire (sub)graphs. In doing so, we develop a unified framework to describe these recent approaches, and we highlight a number of important applications and directions for future work.

1 Introduction

Graphs are a ubiquitous data structure, employed extensively within computer science and related fields. Social networks, molecular graph structures, biological protein-protein networks, recommender systems—all of these domains and many more can be readily modeled as graphs, which capture interactions (*i.e.*, edges) between individual units (*i.e.*, nodes). As a consequence of their ubiquity, graphs are the backbone of countless systems, allowing relational knowledge about interacting entities to be efficiently stored and accessed [2].

However, graphs are not only useful as structured knowledge repositories: they also play a key role in modern machine learning. Machine learning applications seek to make predictions, or discover new patterns, using graph-structured data as feature information. For example, one might wish to classify the role of a protein in a biological interaction graph [28], predict the role of a person in a collaboration network, recommend new friends to a user in a social network [3], or predict new therapeutic applications of existing drug molecules, whose structure can be represented as a graph [21].

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

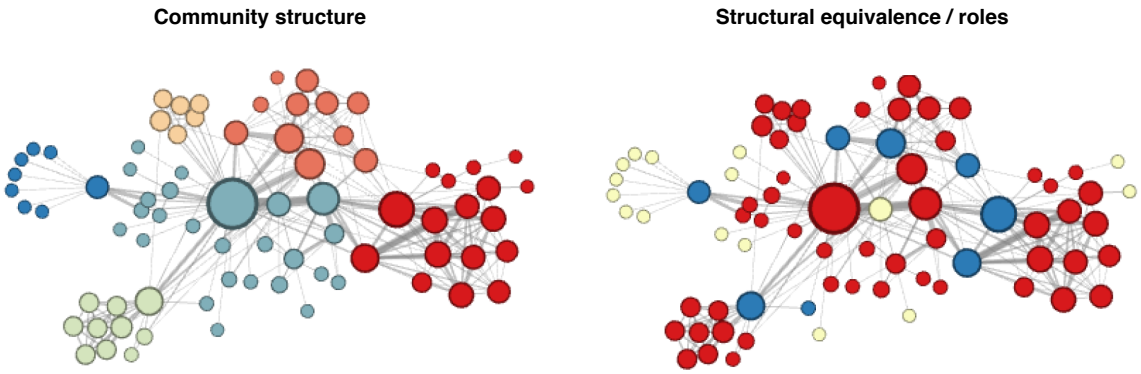


Figure 1: Two different views of a character-character interaction graph derived from the Les Misérables novel, where two nodes are connected if the corresponding characters interact. The coloring in the left figure emphasizes differences in the nodes' global positions in the graph: nodes have the same color if they belong to the same community, at a global level. In contrast, the coloring in the right figure denotes structural equivalence between nodes, or the fact that two nodes play similar roles in their local neighborhoods (*e.g.*, "bridging nodes" are colored blue). The colorings for both figures were generated using different settings of the node2vec node embedding method [27], described in Section 2. Reprinted from [27] with permission.[1]

The central problem in machine learning on graphs is finding a way to incorporate information about the structure of the graph into the machine learning model. For example, in the case of link prediction in a social network, one might want to encode pairwise properties between nodes, such as relationship strength or the number of common friends. Or in the case of node classification, one might want to include information about the global position of a node in the graph or the structure of the node's local graph neighborhood (Figure 1), and there is no straightforward way to encode this information into a feature vector.

To extract structural information from graphs, traditional machine approaches often rely on summary graph statistics (*e.g.*, degrees or clustering coefficients) [6], kernel functions [57], or carefully engineered features to measure local neighborhood structures [39]. However, these approaches are limited because these hand-engineered features are inflexible—*i.e.*, they cannot adapt during the learning process—and designing these features can be a time-consuming and expensive process.

More recently, there has been a surge of approaches that seek to *learn* representations that encode structural information about the graph. The idea behind these *representation learning* approaches is to learn a mapping that embeds nodes, or entire (sub)graphs, as points in a low-dimensional vector space, $\mathbb{R}^d$. The goal is to optimize this mapping so that geometric relationships in this learned space reflect the structure of the original graph. After optimizing the embedding space, the learned embeddings can be used as feature inputs for downstream machine learning tasks. The key distinction between representation learning approaches and previous work is how they treat the problem of capturing structural information about the graph. Previous work treated this problem as a pre-processing step, using hand-engineered statistics to extract structural information. In contrast, representation learning approaches treat this problem as machine learning task itself, using a data-driven approach to learn embeddings that encode graph structure.

Here we provide an overview of recent advancements in representation learning on graphs, reviewing techniques for representing both nodes and entire subgraphs. Our survey attempts to merge together multiple, disparate lines of research that have drawn significant attention across different subfields and venues in recent

[1]For this and all subsequent reprinted figures, the original authors retain their copyrights, and permission was obtained from the corresponding author.

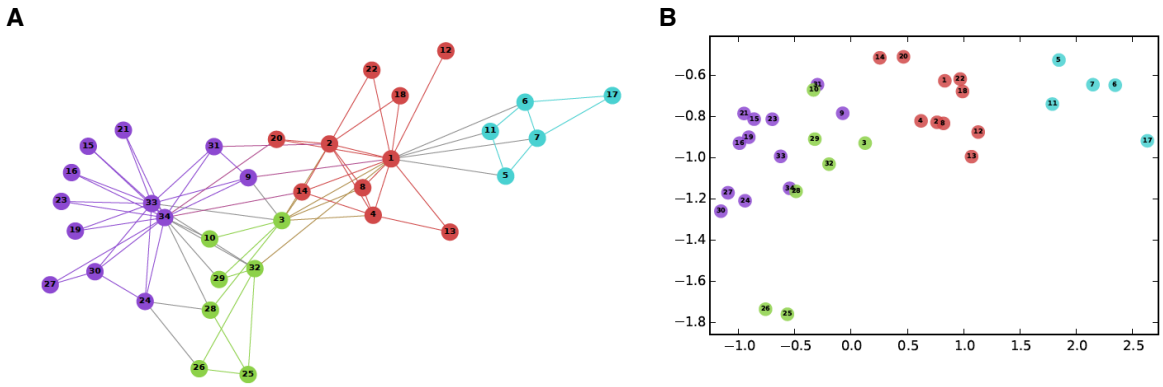


Figure 2: **A**, Graph structure of the Zachary Karate Club social network, where nodes are connected if the corresponding individuals are friends. The nodes are colored according to the different communities that exist in the network. **B**, Two-dimensional visualization of node embeddings generated from this graph using the DeepWalk method (Section 2.2.2) [46]. The distances between nodes in the embedding space reflect proximity in the original graph, and the node embeddings are spatially clustered according to the different color-coded communities. Reprinted with permission from [46, 48].

years—*e.g.*, node embedding methods, which are a popular object of study in the data mining community, and graph convolutional networks, which have drawn considerable attention in major machine learning venues. In doing so, we develop a unified conceptual framework for describing the various approaches and emphasize major conceptual distinctions.

We focus our review on recent approaches that have garnered significant attention in the machine learning and data mining communities, especially methods that are scalable to massive graphs (*e.g.*, millions of nodes) and inspired by advancements in deep learning. Of course, there are other lines of closely related and relevant work, which we do not review in detail here—including latent space models of social networks [32], embedding methods for statistical relational learning [42], manifold learning algorithms [37], and geometric deep learning [7]—all of which involve representation learning with graph-structured data. We refer the reader to [32], [42], [37], and [7] for comprehensive overviews of these areas.

1.1 Notation and essential assumptions

We will assume that the primary input to our representation learning algorithm is an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with associated binary adjacency matrix, $\mathbf{A}$.[2] We also assume that the methods can make use of a real-valued matrix of node attributes $\mathbf{X} \in \mathbb{R}^{m \times |\mathcal{V}|}$ (*e.g.*, representing text or metadata associated with nodes). The goal is to use the information contained in $\mathbf{A}$ and $\mathbf{X}$ to map each node, or a subgraph, to a vector $\mathbf{z} \in \mathbb{R}^d$, where $d << |\mathcal{V}|$.

Most of the methods we review will optimize this mapping in an *unsupervised* manner, making use of only information in $\mathbf{A}$ and $\mathbf{X}$, without knowledge of the downstream machine learning task. However, we will also discuss some approaches for *supervised* representation learning, where the models make use of classification or regression labels in order to optimize the embeddings. These classification labels may be associated with individual nodes or entire subgraphs and are the prediction targets for downstream machine learning tasks (*e.g.*, they might label protein roles, or the therapeutic properties of a molecule, based on its graph representation).

[2] Most of the methods we review are easily generalized to work with weighted or directed graphs, and we will explicitly describe how to generalize certain methods to the multi-modal setting (*i.e.*, differing node and edge types).

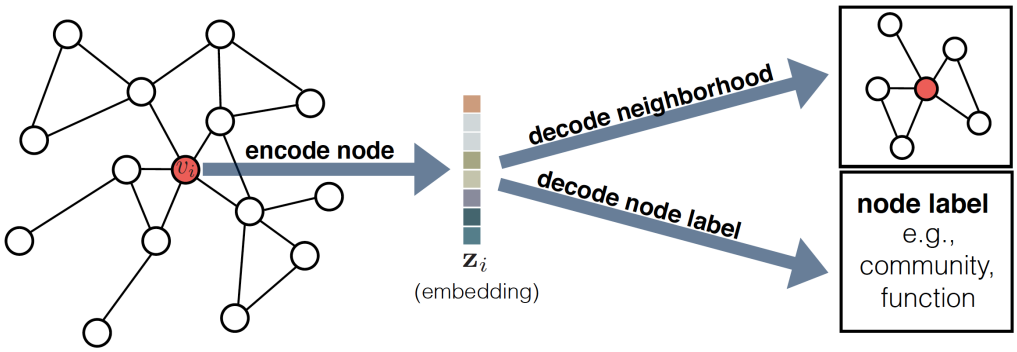


Figure 3: Overview of the encoder-decoder approach. First the encoder maps the node, v_i, to a low-dimensional vector embedding, $\mathbf{z}_i$, based on the node's position in the graph, its local neighborhood structure, and/or its attributes. Next, the decoder extracts user-specified information from the low-dimensional embedding; this might be information about v_i's local graph neighborhood (*e.g.*, the identity of its neighbors) or a classification label associated with v_i (*e.g.*, a community label). By jointly optimizing the encoder and decoder, the system learns to compress information about graph structure into the low-dimensional embedding space.

2 Embedding nodes

We begin with a discussion of methods for *node embedding*, where the goal is to encode nodes as low-dimensional vectors that summarize their graph position and the structure of their local graph neighborhood. These low-dimensional embeddings can be viewed as encoding, or projecting, nodes into a latent space, where geometric relations in this latent space correspond to interactions (*e.g.*, edges) in the original graph [32]. Figure 2 visualizes an example embedding of the famous Zachary Karate Club social network [46], where two dimensional node embeddings capture the community structure implicit in the social network.

2.1 Overview of approaches: An encoder-decoder perspective

Recent years have seen a surge of research on node embeddings, leading to a complicated diversity of notations, motivations, and conceptual models. Thus, before discussing the various techniques, we first develop a unified *encoder-decoder* framework, which explicitly structures this methodological diversity and puts the various methods on equal notational and conceptual footing.

In this framework, we organize the various methods around two key mapping functions: an *encoder*, which maps each node to a low-dimensional vector, or embedding, and a *decoder*, which decodes structural information about the graph from the learned embeddings (Figure 3). The intuition behind the encoder-decoder idea is the following: if we can learn to decode high-dimensional graph information—such as the global positions of nodes in the graph and the structure of local graph neighborhoods—from encoded low-dimensional embeddings, then, in principle, these embeddings should contain all information necessary for downstream machine learning tasks.

Formally, the *encoder* is a function,

$$\text{ENC} : \mathcal{V} \rightarrow \mathbb{R}^d, \tag{1}$$

that maps nodes to vector embeddings, $\mathbf{z}_i \in \mathbb{R}^d$ (where $\mathbf{z}_i$ corresponds to the embedding for node $v_i \in \mathcal{V}$). The *decoder* is a function that accepts a set of node embeddings and decodes user-specified graph statistics from these embeddings. For example, the decoder might predict the existence of edges between nodes, given their embeddings [1, 35], or it might predict the community that a node belongs to in the graph [28, 34] (Figure 3). In principle, many decoders are possible; however, the vast majority of works use a basic *pairwise decoder*,

$$\text{DEC} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+, \tag{2}$$

that maps pairs of node embeddings to a real-valued graph proximity measure, which quantifies the proximity of the two nodes in the original graph.

When we apply the pairwise decoder to a pair of embeddings ($\mathbf{z}_i$,$\mathbf{z}_j$) we get a *reconstruction* of the proximity between v_i and v_j in the original graph, and the goal is optimize the encoder and decoder mappings to minimize the error, or loss, in this reconstruction so that:

$$\text{DEC}(\text{ENC}(v_i), \text{ENC}(v_j)) = \text{DEC}(\mathbf{z}_i, \mathbf{z}_j) \approx s_{\mathcal{G}}(v_i, v_j), \tag{3}$$

where $s_{\mathcal{G}}$ is a user-defined, graph-based proximity measure between nodes, defined over the graph, $\mathcal{G}$. For example, one might set $s_{\mathcal{G}}(v_i, v_j) \triangleq \mathbf{A}_{i,j}$ and define nodes to have a proximity of 1 if they are adjacent and 0 otherwise [1], or one might define s_G according to the probability of v_i and v_j co-occurring on a fixed-length random walk over the graph $\mathcal{G}$ [27, 46]. In practice, most approaches realize the reconstruction objective (Equation 3) by minimizing an empirical loss, $\mathcal{L}$, over a set of training node pairs, $\mathcal{D}$:

$$\mathcal{L} = \sum_{(v_i,v_j)\in\mathcal{D}} \ell\left(\text{DEC}(\mathbf{z}_i, \mathbf{z}_j), s_{\mathcal{G}}(v_i, v_j)\right), \tag{4}$$

where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a user-specified loss function, which measures the discrepancy between the decoded (*i.e.*, estimated) proximity values, $\text{DEC}(\mathbf{z}_i, \mathbf{z}_j)$, and the true values, $s_{\mathcal{G}}(v_i, v_j)$.

Once we have optimized the encoder-decoder system, we can use the trained encoder to generate embeddings for nodes, which can then be used as a feature inputs for downstream machine learning tasks. For example, one could feed the learned embeddings to a logistic regression classifier to predict the community that a node belongs to [46], or one could use distances between the embeddings to recommend friendship links in a social network [3, 27] (Section 2.7 discusses further applications).

Adopting this encoder-decoder view, we organize our discussion of the various node embedding methods along the following four methodological components:

1. A **pairwise proximity function** $s_{\mathcal{G}} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}^+$, defined over the graph, $\mathcal{G}$. This function measures how closely connected two nodes are in $\mathcal{G}$.

2. An **encoder function**, ENC, that generates the node embeddings. This function contains a number of trainable parameters that are optimized during the training phase.

3. A **decoder function**, DEC, which reconstructs pairwise proximity values from the generated embeddings. This function usually contains no trainable parameters.

4. A **loss function**, ℓ, which determines how the quality of the pairwise reconstructions is evaluated in order to train the model, *i.e.*, how $\text{DEC}(\mathbf{z}_i, \mathbf{z}_j)$ is compared to the true $s_{\mathcal{G}}(v_i, v_j)$ values.

As we will show, the primary methodological distinctions between the various node embedding approaches are in how they define these four components.

2.1.1 Notes on optimization and implementation details

All of the methods we review involve optimizing the parameters of the encoder algorithm, Θ_{ENC}, by minimizing a loss analogous to Equation (4).[3] In most cases, stochastic gradient descent is used for optimization, though some algorithms do permit closed-form solutions via matrix decomposition (*e.g.*, [9]). However, note that we will not focus on optimization algorithms here and instead will emphasize high-level differences that exist across different embedding methods, independent of the specifics of the optimization approach.

[3] Occasionally, different methods will add additional auxiliary objectives or regularizers beyond the standard encoder-decoder objective, but we will often omit these details for brevity. A few methods also optimize parameters in the decoder, Θ_{DEC}.

Table 1: A summary of some well-known direct encoding embedding algorithms. Note that the decoders and proximity functions for the random-walk based methods are asymmetric, with the proximity function, $p_{\mathcal{G}}(v_j|v_i)$, corresponding to the probability of visiting v_j on a fixed-length random walk starting from v_i.

Type	Method	Decoder	Proximity measure	Loss function (ℓ)
Matrix factorization	Laplacian Eigenmaps [4]	$\|\mathbf{z}_i - \mathbf{z}_j\|_2^2$	general	$\text{DEC}(\mathbf{z}_i, \mathbf{z}_j) \cdot s_{\mathcal{G}}(v_i, v_j)$
	Graph Factorization [1]	$\mathbf{z}_i^\top \mathbf{z}_j$	$\mathbf{A}_{i,j}$	$\|\text{DEC}(\mathbf{z}_i, \mathbf{z}_j) - s_{\mathcal{G}}(v_i, v_j)\|_2^2$
	GraRep [9]	$\mathbf{z}_i^\top \mathbf{z}_j$	$\mathbf{A}_{i,j}, \mathbf{A}_{i,j}^2, ..., \mathbf{A}_{i,j}^k$	$\|\text{DEC}(\mathbf{z}_i, \mathbf{z}_j) - s_{\mathcal{G}}(v_i, v_j)\|_2^2$
	HOPE [44]	$\mathbf{z}_i^\top \mathbf{z}_j$	general	$\|\text{DEC}(\mathbf{z}_i, \mathbf{z}_j) - s_{\mathcal{G}}(v_i, v_j)\|_2^2$
Random walk	DeepWalk [46]	$\frac{e^{\mathbf{z}_i^\top \mathbf{z}_j}}{\sum_{k \in \mathcal{V}} e^{\mathbf{z}_i^\top \mathbf{z}_k}}$	$p_{\mathcal{G}}(v_j\|v_i)$	$-s_{\mathcal{G}}(v_i, v_j) \log(\text{DEC}(\mathbf{z}_i, \mathbf{z}_j))$
	node2vec [27]	$\frac{e^{\mathbf{z}_i^\top \mathbf{z}_j}}{\sum_{k \in \mathcal{V}} e^{\mathbf{z}_i^\top \mathbf{z}_k}}$	$p_{\mathcal{G}}(v_j\|v_i)$ (biased)	$-s_{\mathcal{G}}(v_i, v_j) \log(\text{DEC}(\mathbf{z}_i, \mathbf{z}_j))$

2.2 Direct encoding approaches

The majority of node embedding algorithms rely on what we call *direct encoding*. For these direct encoding approaches, the encoder function—which maps nodes to vector embeddings—is simply an "embedding lookup":

$$\text{ENC}(v_i) = \mathbf{Z}\mathbf{v}_i, \tag{5}$$

where $\mathbf{Z} \in \mathbb{R}^{d \times |\mathcal{V}|}$ is a matrix containing the embedding vectors for all nodes and $\mathbf{v}_i \in \mathbb{I}_{\mathcal{V}}$ is a one-hot indicator vector indicating the column of $\mathbf{Z}$ corresponding to node v_i. The set of trainable parameters for direct encoding methods is simply $\Theta_{\text{ENC}} = \{\mathbf{Z}\}$, *i.e.* the embedding matrix $\mathbf{Z}$ is optimized directly.

These approaches are largely inspired by classic matrix factorization techniques for dimensionality reduction [4] and multi-dimensional scaling [36]. Indeed, many of these approaches were originally motivated as factorization algorithms, and we reinterpret them within the encoder-decoder framework here. Table 1 summarizes some well-known direct-encoding methods within the encoder-decoder framework. Table 1 highlights how these methods can be succinctly described according to (i) their decoder function, (ii) their graph-based proximity measure, and (iii) their loss function. The following two sections describe these methods in more detail, distinguishing between matrix factorization-based approaches (Section 2.2.1) and more recent approaches based on random walks (Section 2.2.2).

2.2.1 Factorization-based approaches

Early methods for learning representations for nodes largely focused on matrix-factorization approaches, which are directly inspired by classic techniques for dimensionality reduction [4, 36].

Laplacian eigenmaps. One of the earliest, and most well-known instances, is the Laplacian eigenmaps (LE) technique [4], which we can view within the encoder-decoder framework as a direct encoding approach in which the decoder is defined as

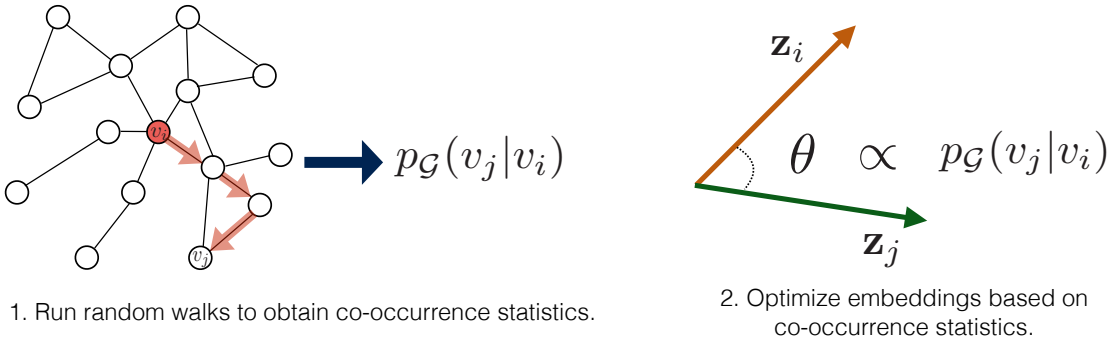
$$\text{DEC}(\mathbf{z}_i, \mathbf{z}_j) = \|\mathbf{z}_i - \mathbf{z}_j\|_2^2$$

and where the loss function weights pairs of nodes according to their proximity in the graph:

$$\mathcal{L} = \sum_{(v_i, v_j) \in \mathcal{D}} \text{DEC}(\mathbf{z}_i, \mathbf{z}_j) \cdot s_{\mathcal{G}}(v_i, v_j). \tag{6}$$

Inner-product methods. Following on the Laplacian eigenmaps technique, there are a large number of recent embedding methodologies based on a pairwise, inner-product decoder:

$$\text{DEC}(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^\top \mathbf{z}_j, \tag{7}$$



1. Run random walks to obtain co-occurrence statistics.

2. Optimize embeddings based on co-occurrence statistics.

Figure 4: The random-walk based methods sample a large number of fixed-length random walks starting from each node, v_i. The embedding vectors are then optimized so that the dot-product, or angle, between two embeddings, $\mathbf{z}_i$ and $\mathbf{z}_j$, is (roughly) proportional to the probability of visiting v_j on a fixed-length random walk starting from v_i.

where the strength of the relationship between two nodes is proportional to the dot product of their embeddings. The Graph Factorization (GF) algorithm[4] [1], GraRep [9], and HOPE [44] all fall firmly within this class. In particular, all three of these methods use an inner-product decoder, a mean-squared-error (MSE) loss,

$$\mathcal{L} = \sum_{(v_i,v_j)\in\mathcal{D}} \|\text{DEC}(\mathbf{z}_i, \mathbf{z}_j) - s_{\mathcal{G}}(v_i, v_j)\|_2^2, \tag{8}$$

and they differ primarily in the graph proximity measure used, *i.e.* how they define $s_{\mathcal{G}}(v_i, v_j)$. The Graph Factorization algorithm defines proximity directly based on the adjacency matrix (*i.e.*, $s_{\mathcal{G}}(v_i, v_j) \triangleq \mathbf{A}_{i,j}$) [1]; GraRep considers various powers of the adjacency matrix (*e.g.*, $s_{\mathcal{G}}(v_i, v_j) \triangleq \mathbf{A}_{i,j}^2$) in order to capture higher-order graph proximity [9]; and the HOPE algorithm supports general proximity measures (*e.g.*, based on Jaccard neighborhood overlaps) [44]. These various different proximity functions trade-off between modeling "first-order proximity", where $s_{\mathcal{G}}$ directly measures connections between nodes (*i.e.*, $s_{\mathcal{G}}(v_i, v_j) \triangleq \mathbf{A}_{i,j}$ [1]) and modeling "higher-order proximity", where $s_{\mathcal{G}}$ corresponds to more general notions of neighborhood overlap (*e.g.*, $s_{\mathcal{G}}(v_i, v_j) = \mathbf{A}_{i,j}^2$ [9]).

We refer to these methods in this section as matrix-factorization approaches because, averaging over all nodes, they optimize loss functions (roughly) of the form:

$$\mathcal{L} \approx \|\mathbf{Z}^\top\mathbf{Z} - \mathbf{S}\|_2^2, \tag{9}$$

where $\mathbf{S}$ is a matrix containing pairwise proximity measures (*i.e.*, $\mathbf{S}_{i,j} \triangleq s_{\mathcal{G}}(v_i, v_j)$) and $\mathbf{Z}$ is the matrix of node embeddings. Intuitively, the goal of these methods is simply to learn embeddings for each node such that the inner product between the learned embedding vectors approximates some deterministic measure of graph proximity.

2.2.2 Random walk approaches

Many recent successful methods that also belong to the class of *direct encoding* approaches learn the node embeddings based on random walk statistics. Their key innovation is optimizing the node embeddings so that nodes have similar embeddings if they tend to co-occur on short random walks over the graph (Figure 4). Thus, instead of using a deterministic measure of graph proximity, like the methods of Section 2.2.1, these random walk methods employ a flexible, stochastic measure of graph proximity, which has led to superior performance in a number of settings [26].

[4]Of course, Ahmed et al. [1] were not the first researchers to propose factorizing an adjacency matrix, but they were the first to present a scalable $O(|\mathcal{E}|)$ algorithm for the purpose of generating node embeddings.

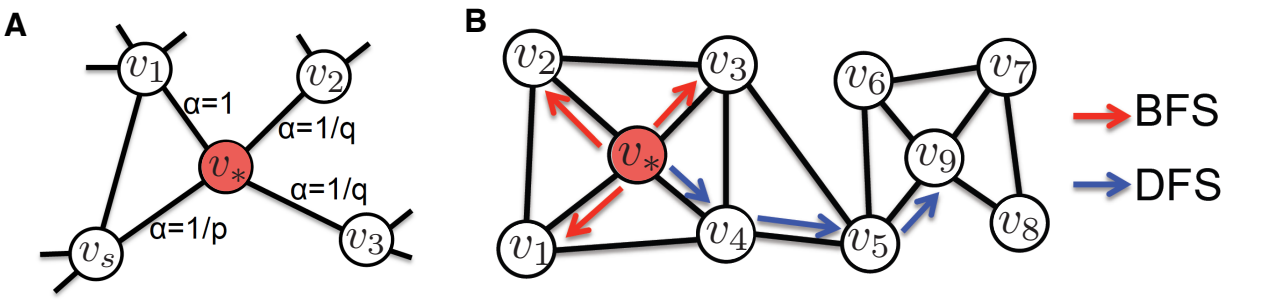


Figure 5: **A**, Illustration of how node2vec biases the random walk using the p and q parameters. Assuming that the walk just transitioned from v_s to v_*, the edge labels, α, are proportional to the probability of the walk taking that edge at next time-step. **B**, Difference between random-walks that are based on breadth-first search (BFS) and depth-first search (DFS). BFS-like random walks are mainly limited to exploring a node's immediate (*i.e.*, one-hop) neighborhood and are generally more effective for capturing structural roles. DFS-like walks explore further away from the node and are more effective for capturing community structures. Adapted from [27].

DeepWalk and node2vec. Like the matrix factorization approaches described above, DeepWalk and node2vec rely on direct encoding and use a decoder based on the inner product. However, instead of trying to decode a fixed deterministic distance measure, these approaches optimize embeddings to encode the statistics of random walks. The basic idea behind these approaches is to learn embeddings so that (roughly):

$$\begin{aligned} \text{DEC}(\mathbf{z}_i, \mathbf{z}_j) &\triangleq \frac{e^{\mathbf{z}_i^\top \mathbf{z}_j}}{\sum_{v_k \in \mathcal{V}} e^{\mathbf{z}_i^\top \mathbf{z}_k}} \\ &\approx p_{\mathcal{G},T}(v_j|v_i), \end{aligned} \tag{10}$$

where $p_{\mathcal{G},T}(v_j|v_i)$ is the probability of visiting v_j on a length-T random walk starting at v_i, with T usually defined to be in the range $T \in \{2, ..., 10\}$. Note that unlike the proximity measures in Section 2.2.1, $p_{\mathcal{G},T}(v_j|v_i)$ is both stochastic and asymmetric.

More formally, these approaches attempt to minimize the following cross-entropy loss:

$$\mathcal{L} = \sum_{(v_i, v_j) \in \mathcal{D}} -\log(\text{DEC}(\mathbf{z}_i, \mathbf{z}_j)), \tag{11}$$

where in this case the training set, $\mathcal{D}$, is generated by sampling random walks starting from each node (*i.e.*, where N pairs for each node, v_i, are sampled from the distribution $(v_i, v_j) \sim p_{\mathcal{G},T}(v_j|v_j)$). However, naively evaluating this loss is prohibitively expensive—in particular, $O(|\mathcal{D}||\mathcal{V}|)$—since evaluating the denominator of Equation (10) has time complexity $O(|\mathcal{V}|)$. Thus, DeepWalk and node2vec use different optimizations and approximations to compute the loss in Equation (11). DeepWalk employs a "hierarchical softmax" technique to compute the normalizing factor, using a binary-tree structure to accelerate the computation [46]. In contrast, node2vec approximates Equation (11) using "negative sampling": instead of normalizing over the full vertex set, node2vec approximates the normalizing factor using a set of random "negative samples" [27].

Beyond these algorithmic differences, the key distinction between node2vec and DeepWalk is that node2vec allows for a flexible definition of random walks, whereas DeepWalk uses simple unbiased random walks over the graph. In particular, node2vec introduces two random walk hyperparameters, p and q, that bias the random walk (Figure 5.A). The hyperparameter p controls the likelihood of the walk immediately revisiting a node, while q controls the likelihood of the walk revisiting a node's one-hop neighborhood. By introducing these hyperparameters, node2vec is able to smoothly interpolate between walks that are more akin to breadth-first or depth-first search (Figure 5.B). Grover et al. found that tuning these parameters allowed the model to trade

off between learning embeddings that emphasize community structures or embeddings that emphasize local structural roles [27] (see also Figure 1).

Large-scale information network embeddings (LINE). Another highly successful direct encoding approach, which is not based random walks but is contemporaneous and often compared with DeepWalk and node2vec, is the LINE method [53]. LINE combines two encoder-decoder objectives that optimize "first-order" and "second-order" graph proximity, respectively. The first-order objective uses a decoder based on the sigmoid function,

$$\text{DEC}(\mathbf{z}_i, \mathbf{z}_j) = \frac{1}{1 + e^{-\mathbf{z}_i^\top \mathbf{z}_j}}, \tag{12}$$

and an adjacency-based proximity measure (*i.e.*, $s_{\mathcal{G}}(v_i, v_j) = \mathbf{A}_{i,j}$). The second-order encoder-decoder objective is similar but considers two-hop adjacency neighborhoods and uses an encoder identical to Equation (10). Both the first-order and second-order objectives are optimized using loss functions derived from the KL-divergence metric [53]. Thus, LINE is conceptually related to node2vec and DeepWalk in that it uses a probabilistic decoder and loss, but it explicitly factorizes first- and second-order proximities, instead of combining them in fixed-length random walks.

HARP: Extending random-walk embeddings via graph pre-processing. Recently, Chen et al. [13] introduced a "meta-strategy", called HARP, for improving various random-walk approaches via a graph pre-processing step. In this approach, a graph coarsening procedure is used to collapse related nodes in $\mathcal{G}$ together into "supernodes", and then DeepWalk, node2vec, or LINE is run on this coarsened graph. After embedding the coarsened version of $\mathcal{G}$, the learned embedding of each supernode is used as an initial value for the random walk embeddings of the supernode's constituent nodes (in another round of non-convex optimization on a "finer-grained" version of the graph). This general process can be repeated in a hierarchical manner at varying levels of coarseness, and has been shown to consistently improve performance of DeepWalk, node2vec, and LINE [13].

Additional variants of the random-walk idea. There have also been a number of further extensions of the random walk idea. For example, Perozzi et al. [47] extend the DeepWalk algorithm to learn embeddings using random walks that "skip" or "hop" over multiple nodes at each step, resulting in a proximity measure similar to GraRep [9], while Chamberlan et al. [11] modify the inner-product decoder of node2vec to use a hyperbolic, rather than Euclidean, distance measure.

2.3 Generalized encoder-decoder architectures

So far all of the node embedding methods we have reviewed have been direct encoding methods, where the encoder is a simply an embedding lookup (Equation 5). However, these direct encoding approaches train unique embedding vectors for each node independently, which leads to a number of drawbacks:

1. No parameters are shared between nodes in the encoder (*i.e.*, the encoder is simply an embedding lookup based on arbitrary node ids). This can be statistically inefficient, since parameter sharing can act as a powerful form of regularization, and it is also computationally inefficient, since it means that the number of parameters in direct encoding methods necessarily grows as $O(|\mathcal{V}|)$.

2. Direct encoding also fails to leverage node attributes during encoding. In many large graphs nodes have attribute information (*e.g.*, user profiles on a social network) that is often highly informative with respect to the node's position and role in the graph.

3. Direct encoding methods are inherently *transductive* [28], *i.e.*, they can only generate embeddings for nodes that were present during the training phase, and they cannot generate embeddings for previously unseen nodes unless additional rounds of optimization are performed to optimize the embeddings for these nodes. This is highly problematic for evolving graphs, massive graphs that cannot be fully stored in memory, or domains that require generalizing to new graphs after training.

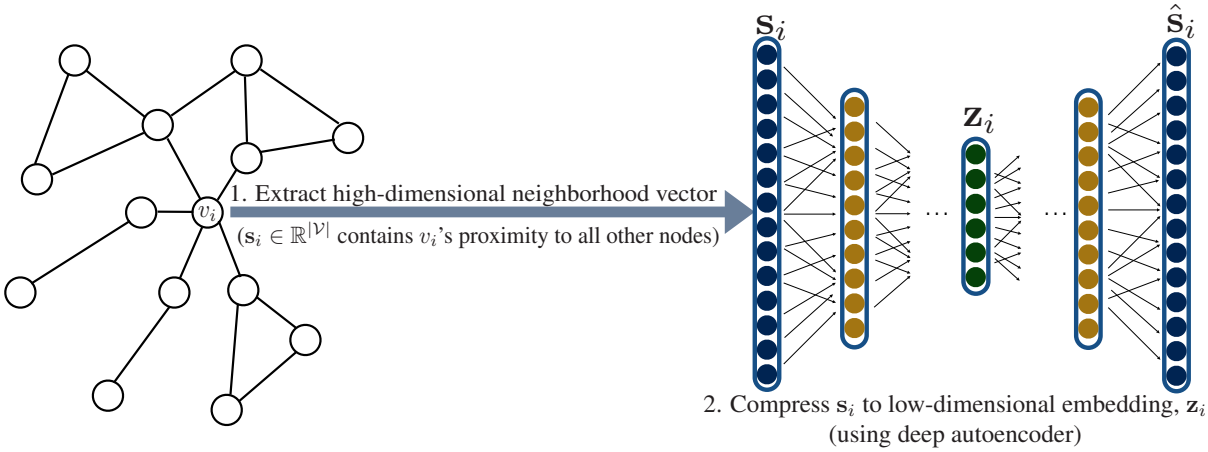


Figure 6: To generate an embedding for a node, v_i, the neighborhood autoencoder approaches first extract a high-dimensional neighborhood vector $\mathbf{s}_i \in \mathbb{R}^{|\mathcal{V}|}$, which summarizes v_i's proximity to all other nodes in the graph. The $\mathbf{s}_i$ vector is then fed through a deep autoencoder to reduce its dimensionality, producing the low-dimensional $\mathbf{z}_i$ embedding.

Recently, a number of approaches have been proposed to address some, or all, of these issues. These approaches still fall firmly within the encoder-decoder framework outlined in Section 2.1, but they differ from the direct encoding methods of Section 2.2 in that they use a more complex encoders, which depend more generally on the structure and attributes of the graph.

2.3.1 Neighborhood autoencoder methods

Deep Neural Graph Representations (DNGR) [10] and Structural Deep Network Embeddings (SDNE) [58] address the first problem outlined above: unlike the direct encoding methods, they directly incorporate graph structure into the encoder algorithm. The basic idea behind these approaches is that they use autoencoders—a well known approach for deep learning [30]—in order to compress information about a node's local neighborhood (Figure 6). DNGR and SDNE also differ from the previously reviewed approaches in that they use a *unary decoder* instead of a pairwise one.

In these approaches, each node, v_i, is associated with a neighborhood vector, $\mathbf{s}_i \in \mathbb{R}^{|\mathcal{V}|}$, which corresponds to v_i's row in the matrix $\mathbf{S}$ (recall that $\mathbf{S}$ contains pairwise node proximities, *i.e.*, $\mathbf{S}_{i,j} = s_{\mathcal{G}}(v_i, v_j)$). The $\mathbf{s}_i$ vector contains v_i's pairwise graph proximity with all other nodes and functions as a high-dimensional vector representation of v_i's neighborhood. The autoencoder objective for DNGR and SDNE is to embed nodes using the $\mathbf{s}_i$ vectors such that the $\mathbf{s}_i$ vectors can then be reconstructed from these embeddings:

$$\text{DEC}(\text{ENC}(\mathbf{s}_i)) = \text{DEC}(\mathbf{z}_i) \approx \mathbf{s}_i. \tag{13}$$

In other words, the loss for these methods takes the following form:

$$\mathcal{L} = \sum_{v_i \in \mathcal{V}} \|\text{DEC}(\mathbf{z}_i) - \mathbf{s}_i\|_2^2. \tag{14}$$

As with the pairwise decoder, we have that the dimension of the $\mathbf{z}_i$ embeddings is much smaller than $|\mathcal{V}|$ (the dimension of the $\mathbf{s}_i$ vectors), so the goal is to compress the node's neighborhood information into a low-dimensional vector. For both SDNE and DNGR, the encoder and decoder functions consist of multiple stacked neural network layers: each layer of the encoder reduces the dimensionality of its input, and each layer of the decoder increases the dimensionality of its input (Figure 6; see [30] for an overview of deep autoencoders).

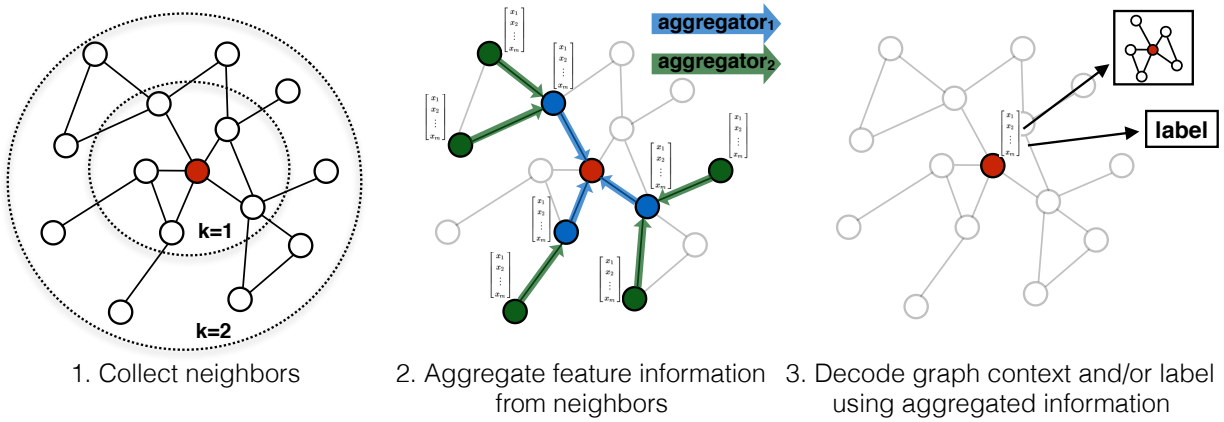


Figure 7: Overview of the neighborhood aggregation methods. To generate an embedding for a node, these methods first collect the node's k-hop neighborhood (occasionally sub-sampling the full neighborhood for efficiency). In the next step, these methods aggregate the attributes of node's neighbors, using neural network aggregators. This aggregated neighborhood information is used to generate an embedding, which is then fed to the decoder. Adapted from [28].

SDNE and DNGR differ in the similarity functions they use to construct the neighborhood vectors $\mathbf{s}_i$ and also in the exact details of how the autoencoder is optimized. DNGR defines $\mathbf{s}_i$ according to the pointwise mutual information of two nodes co-occurring on random walks, similar to DeepWalk and node2vec. SDNE simply sets $\mathbf{s}_i \triangleq \mathbf{A}_i$, *i.e.*, equal to v_i's adjacency vector. SDNE also combines the autoencoder objective (Equation 13) with the Laplacian eigenmaps objective (Equation 6) [58].

Note that the encoder in Equation (13) depends on the input $\mathbf{s}_i$ vector, which contains information about v_i's local graph neighborhood. This dependency allows SDNE and DNGR to incorporate structural information about a node's local neighborhood directly into the encoder as a form of regularization, which is not possible for the direct encoding approaches (since their encoder depends only on the node id). However, despite this improvement, the autoencoder approaches still suffer from some serious limitations. Most prominently, the input dimension to the autoencoder is fixed at $|\mathcal{V}|$, which can be extremely costly and even intractable for graphs with millions of nodes. In addition, the structure and size of the autoencoder is fixed, so SDNE and DNGR are strictly transductive and cannot cope with evolving graphs, nor can they generalize across graphs.

2.3.2 Neighborhood aggregation and convolutional encoders

A number of recent node embedding approaches aim to solve the main limitations of the direct encoding and autoencoder methods by designing encoders that rely on a node's local neighborhood, but not necessarily the entire graph. The intuition behind these approaches is that they generate embeddings for a node by aggregating information from its local neighborhood (Figure 7).

Unlike the previously discussed methods, these *neighborhood aggregation* algorithms rely on node features or attributes (denoted $\mathbf{x}_i \in \mathbb{R}^m$) to generate embeddings. For example, a social network might have text data (*e.g.*, profile information), or a protein-protein interaction network might have molecular markers associated with each node. The neighborhood aggregation methods leverage this attribute information to inform their embeddings. In cases where attribute data is not given, these methods can use simple graph statistics as attributes (*e.g.*, node degrees) [28], or assign each node a one-hot indicator vector as an attribute [35, 52]. These methods are often called *convolutional* because they represent a node as a function of its surrounding neighborhood, in a manner similar to the receptive field of a center-surround convolutional kernel in computer vision [34].[5]

[5]These methods also have theoretical connections to approximate spectral kernels on graphs [18]; see [34] for a further discussion.

Algorithm 1: Neighborhood-aggregation encoder algorithm. Adapted from [28].

Input : Graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$; input features $\{\mathbf{x}_v, \forall v \in \mathcal{V}\}$; depth K; weight matrices $\{\mathbf{W}^k, \forall k \in [1, K]\}$; non-linearity σ; differentiable aggregator functions $\{\text{AGGREGATE}_k, \forall k \in [1, K]\}$; neighborhood function $\mathcal{N} : v \rightarrow 2^{\mathcal{V}}$

Output: Vector representations $\mathbf{z}_v$ for all $v \in \mathcal{V}$

1 $\mathbf{h}_v^0 \leftarrow \mathbf{x}_v, \forall v \in \mathcal{V}$;
2 **for** $k = 1...K$ **do**
3 **for** $v \in \mathcal{V}$ **do**
4 $\mathbf{h}_{\mathcal{N}(v)}^k \leftarrow \text{AGGREGATE}_k(\{\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v)\})$;
5 $\mathbf{h}_v^k \leftarrow \sigma\left(\mathbf{W}^k \cdot \text{COMBINE}(\mathbf{h}_v^{k-1}, \mathbf{h}_{\mathcal{N}(v)}^k)\right)$
6 **end**
7 $\mathbf{h}_v^k \leftarrow \text{NORMALIZE}(\mathbf{h}_v^k), \forall v \in \mathcal{V}$
8 **end**
9 $\mathbf{z}_v \leftarrow \mathbf{h}_v^K, \forall v \in \mathcal{V}$

In the encoding phase, the neighborhood aggregation methods build up the representation for a node in an iterative, or recursive, fashion (see Algorithm 1 for pseudocode). First, the node embeddings are initialized to be equal to the input node attributes. Then at each iteration of the encoder algorithm, nodes aggregate the embeddings of their neighbors, using an aggregation function that operates over sets of vectors. After this aggregation, every node is assigned a new embedding, equal to its aggregated neighborhood vector combined with its previous embedding from the last iteration. Finally, this combined embedding is fed through a dense neural network layer and the process repeats. As the process iterates, the node embeddings contain information aggregated from further and further reaches of the graph. However, the dimensionality of the embeddings remains constrained as the process iterates, so the encoder is forced to compress all the neighborhood information into a low dimensional vector. After K iterations the process terminates and the final embedding vectors are output as the node representations.

There are a number of recent approaches that follow the basic procedure outlined in Algorithm 1, including graph convolutional networks (GCN) [34, 35, 52, 55], column networks [49], and the GraphSAGE algorithm [28]. The trainable parameters in Algorithm 1—a set of aggregation functions and a set weight matrices $\{\mathbf{W}^k, \forall k \in [1, K]\}$—specify how to aggregate information from a node's local neighborhood and, unlike the direct encoding approaches (Section 2.2), these parameters are shared across nodes. The same aggregation function and weight matrices are used to generate embeddings for all nodes, and only the input node attributes and neighborhood structure change depending on which node is being embedded. This parameter sharing increases efficiency (*i.e.*, the parameter dimensions are independent of the size of the graph), provides regularization, and allows this approach to be used to generate embeddings for nodes that were not observed during training [28].

GraphSAGE, column networks, and the various GCN approaches all follow Algorithm 1 but differ primarily in how the aggregation (line 4) and vector combination (line 5) are performed. GraphSAGE uses concatenation in line 5 and permits general aggregation functions; the authors experiment with using the element-wise mean, a max-pooling neural network and LSTMs [31] as aggregators, and they found the the more complex aggregators, especially the max-pooling neural network, gave significant gains. GCNs and column networks use a weighted sum in line 5 and a (weighted) element-wise mean in line 4.

Column networks also add an additional "interpolation" term before line 7, setting

$$\mathbf{h}_v^{k'} = \alpha \mathbf{h}_v^k + (1 - \alpha)\mathbf{h}_v^{k-1}, \tag{15}$$

where α is an interpolation weight computed as a non-linear function of $\mathbf{h}_v^{k-1}$ and $\mathbf{h}_{\mathcal{N}(v)}^{k-1}$. This interpolation term

allows the model to retain local information as the process iterates (*i.e.*, as k increases and the model integrates information from further reaches of the graph).

In principle, the GraphSAGE, column network, and GCN encoders can be combined with any of the previously discussed decoders and loss functions, and the entire system can be optimized using SGD. For example, Hamilton et al. [28] use an identical decoder and loss as node2vec, while Kipf et al. [35] use a decoder and loss function similar to the Graph Factorization approach.

Neighborhood aggregation encoders following Algorithm 1 have been found to provide consistent gains compared to their direct encoding counterparts, on both node classification [28, 34] and link prediction [55, 35, 52] benchmarks. At a high level, these approaches solve the four main limitations of direct encoding, noted at the beginning of Section 2.3: they incorporate graph structure into the encoder; they leverage node attributes; their parameter dimension can be made sub-linear in $|\mathcal{V}|$; and they can generate embeddings for nodes that were not present during training.

2.4 Incorporating task-specific supervision

The basic encoder-decoder framework described thus far is by default unsupervised, *i.e.*, the model is optimized, or trained, over set of node pairs to reconstruct pairwise proximity values, $s_{\mathcal{G}}(v_i, v_j)$, which depend only on the graph, $\mathcal{G}$. However, many node embedding algorithms—especially the neighborhood aggregation approaches presented in Section 2.3.2—can also incorporate task-specific supervision [28, 34, 52, 59]. In particular, it is common for methods incorporate supervision from node classification tasks in order to learn the embeddings.[6] For simplicity, we discuss the case where nodes have an associated binary classification label, but the approach we describe is easily extended to more complex classification settings.

Assume that we have a binary classification label, $y_i \in \mathbb{Z}$, associated with each node. To learn to map nodes to their labels, we can feed our embedding vectors, $\mathbf{z}_i$, through a logistic, or sigmoid, function $\hat{y}_i = \sigma(\mathbf{z}_i^\top \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a trainable parameter vector. We can then compute the cross-entropy loss between these predicted class probabilities and the true labels:

$$\mathcal{L} = \sum_{v_i \in \mathcal{V}} y_i \log(\sigma(\textsc{enc}(v_i)^\top \boldsymbol{\theta})) + (1 - y_i) \log(1 - \sigma(\textsc{enc}(v_i)^\top \boldsymbol{\theta})). \tag{16}$$

The gradient computed according to Equation (16) can then be backpropagated through the encoder to optimize its parameters. This task-specific supervision can completely replace the reconstruction loss computed using the decoder (*i.e.*, Equation 3) [28, 34], or it can be included along with the decoder loss [59].

2.5 Extensions to multi-modal graphs

While we have focused on simple, undirected graphs, many real-world graphs have complex multi-modal, or multi-layer, structures (*e.g.*, heterogeneous node and edge types), and a number of works have introduced strategies to cope with this heterogeneity.

2.5.1 Dealing with different node and edge types

Many graphs contain different types of nodes and edges. For example, recommender system graphs consist of two distinct layers—users and content—while many biological networks have a variety of layers, with distinct interactions between them (*e.g.*, diseases, genes, and drugs).

A general strategy for dealing with this issue is to (i) use different encoders for nodes of different types [12] and (ii) extend pairwise decoders with type-specific parameters [42, 52]. For example, in graphs with varying

[6] The unsupervised pairwise decoder is already naturally aligned with the link prediction task.

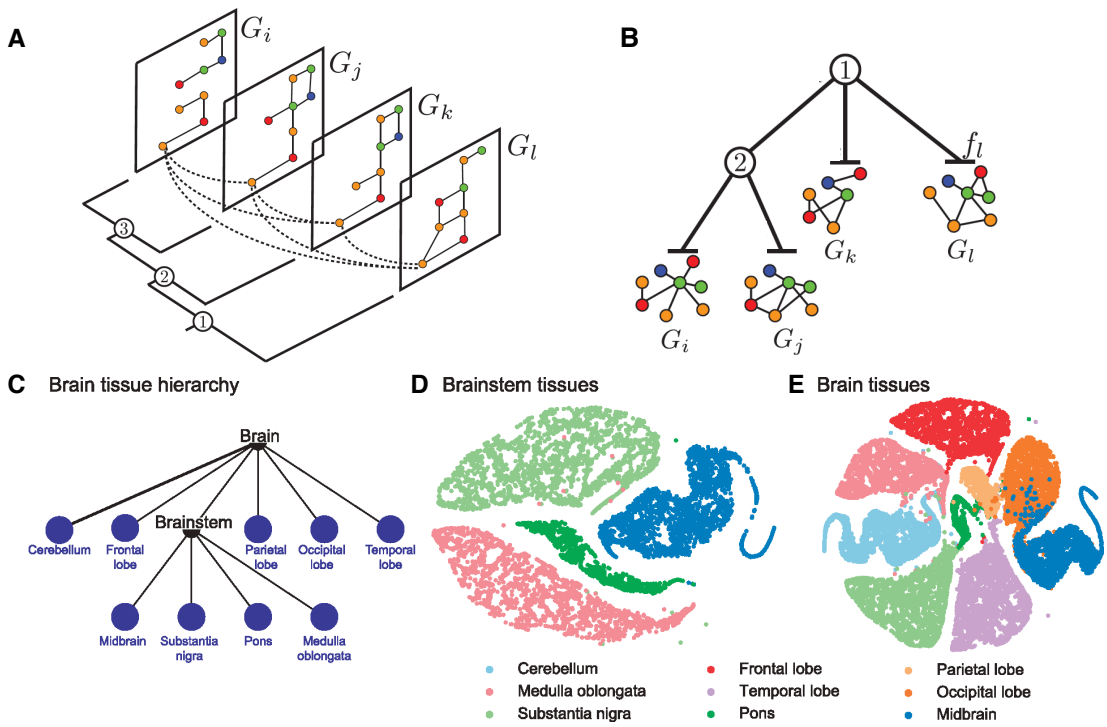


Figure 8: **A**, Example of a 4-layer graph, where the same nodes occur in multiple different layers. This multi-layer structure can be exploited to regularize learning at the different layers by requiring that the embeddings for the same node in different layers are similar to each other. **B**, Multi-layer graphs can exhibit hierarchical structure, where non-root layers in the hierarchy contain the union of the edges present in their child layers—*e.g.*, a biological interaction graph derived from the entire human brain contains the union of the interactions in the frontal and temporal lobes. This structure can be exploited by learning embeddings at various levels of the hierarchy, and only applying the regularization between layers that are in a parent-child relationship. **C-E**, Example application of multi-layer graph embedding to protein-protein interaction graphs derived from different brain tissues; **C** shows the hierarchy between the different tissue regions, while **D** and **E** visualize the protein embeddings generated at the brainstem and whole-brain layers. The embeddings were generated using the multi-layer OhmNet method and projected to two dimensions using t-SNE. Adapted from [60].

edge types, the standard inner-product edge decoder (*i.e.*, $\mathbf{z}_i^\top \mathbf{z}_j \approx \mathbf{A}_{i,j}$) can be replaced with a bilinear form [12, 42, 52]:

$$\text{DEC}_\tau(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}^\top \mathbf{A}_\tau \mathbf{z}, \tag{17}$$

where τ indexes a particular edge type and $\mathbf{A}_\tau$ is a learned parameter specific to edges of type τ. The matrix, $\mathbf{A}_\tau$, in Equation (17) can be regularized in various ways (*e.g.*, constrained to be diagonal) [52], which can be especially useful when there are a large number of edge types, as in the case for embedding knowledge graphs. Indeed, the literature on knowledge-graph completion—where the goal is predict missing relations in knowledge graphs—contains many related techniques for decoding a large number of edge types (*i.e.*, relations) [42].[7]

Recently, Dong et al. [19] also proposed a strategy for sampling random walks from heterogeneous graphs, where the random walks are restricted to only transition between particular types of nodes. This approach allows many of the methods in Section 2.2.2 to be applied on heterogeneous graphs and is complementary to the idea of including type-specific encoders and decoders.

2.5.2 Tying node embeddings across layers

In some cases graphs have multiple "layers" that contain copies of the same nodes (Figure 8.A). For example, in protein-protein interaction networks derived from different tissues (*e.g.*, brain or liver tissue), some proteins occur across multiple tissues. In these cases it can be beneficial to share information across layers, so that a node's embedding in one layer can be informed by its embedding in other layers. Zitnik et al. [60] offer one solution to this problem, called OhmNet, that combines node2vec with a regularization penalty that ties the embeddings across layers. In particular, assuming that we have a node v_i, which belongs to two distinct layers $\mathcal{G}_1$ and $\mathcal{G}_2$, we can augment the standard embedding loss on this node as follows:

$$\mathcal{L}(v_i)' = \mathcal{L}(v_i) + \lambda \|\mathbf{z}_i^{\mathcal{G}_1} - \mathbf{z}_i^{\mathcal{G}_2}\| \tag{18}$$

where $\mathcal{L}$ denotes the usual embedding loss for that node (*e.g.*, from Equation 8 or 11), λ denotes the regularization strength, and $\mathbf{z}_i^{\mathcal{G}_1}$ and $\mathbf{z}_i^{\mathcal{G}_2}$ denote v_i's embeddings in the two different layers, respectively.

Zitnik et al. further extend this idea by exploiting hierarchies between graph layers (Figure 8.B). For example, in protein-protein interaction graphs derived from various tissues, some layers correspond to interactions throughout large regions (*e.g.*, interactions that occur in any brain tissue) while other interaction graphs are more fine-grained (*e.g.*, only interactions that occur in the frontal lobe). To exploit this structure, embeddings can be learned at the various levels of the hierarchy, and the regularization in Equation (18) can recursively applied between layers that have a parent-child relationship in the hierarchy.

2.6 Embedding structural roles

So far, all the approaches we have reviewed optimize node embeddings so that nearby nodes in the graph have similar embeddings. However, in many tasks it is more important to learn representations that correspond to the structural roles of the nodes, independent of their global graph positions (*e.g.*, in communication or transportation networks) [29]. The node2vec approach introduced in Section 2.2.2 offers one solution to this problem, as Grover et al. found that biasing the random walks allows their model to better capture structural roles (Figure 5). However, more recently, Ribeiro et al. [50] and Donnat et al. [20] have developed node embedding approaches that are specifically designed to capture structural roles.

Ribeiro et al. propose struc2vec, which involves generating a a series of weighted auxiliary graphs $\mathcal{G}'_k, k = \{1, 2, ...\}$ from the original graph $\mathcal{G}$, where the auxiliary graph $\mathcal{G}'_k$ captures structural similarities between nodes' k-hop neighborhoods. In particular, letting $R_k(v_i)$ denote the ordered sequence of degrees of the nodes that are exactly k-hops away from v_i, the edge-weights, $w_k(v_i, v_j)$, in auxiliary graph G'_k are recursively defined as

$$w_k(v_i, v_j) = w_{k-1}(v_i, v_j) + d(R_k(v_i), R_k(v_j)), \tag{19}$$

where $w_0(v_i, v_j) = 0$ and $d(R_k(v_i), R_k(v_j))$ measures the "distance" between the ordered degree sequences $R_k(v_i)$ and $R_k(v_j)$ (*e.g.*, computed via dynamic time warping [50]). After computing these weighted auxillary graphs, struc2vec runs biased random walks over them and uses these walks as input to the node2vec optimization algorithm.

Donnat et al. take a very different approach to capturing structural roles, called GraphWave, which relies on spectral graph wavelets and heat kernels [20]. In brief, we let $\mathbf{L}$ denote the graph Laplacian—*i.e.*, $\mathbf{L} = \mathbf{D} - \mathbf{A}$ where $\mathbf{D}$ contains node degrees on the diagonal and $\mathbf{A}$ is the adjacency matrix—and we let $\mathbf{U}$ and $\lambda_i, i = 1...|\mathcal{V}|$ denote the eigenvector matrix and eigenvalues of $\mathbf{L}$, respectively. Finally, we assume that we have a heat kernel, $g(\lambda) = e^{-s\lambda}$, with pre-defined scale s. Using $\mathbf{U}$ and $g(\lambda)$, GraphWave computes a vector, $\boldsymbol{\psi}_{v_i}$, corresponding to the structural role of node, $v_i \in \mathcal{V}$, as

$$\boldsymbol{\psi}_{v_i} = \mathbf{U}\mathbf{G}\mathbf{U}^\top \mathbf{v}_i \tag{20}$$

[7] We do not review this literature in detail here, and refer the reader to Nickel et al. [42] for a recent review.

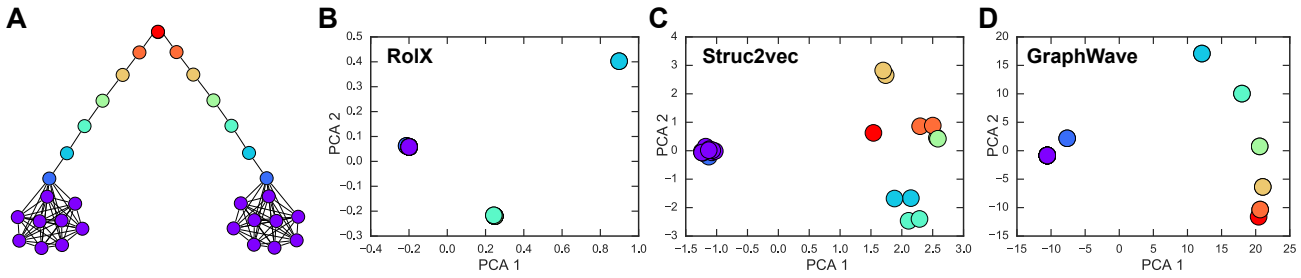


Figure 9: A, Synthetic barbell graph used as a test dataset for detecting structural roles, where nodes are colored according to their structural roles. In this case, the structural roles (*i.e.*, colors) are computed by examining the degrees of each node's immediate neighbors, and their 2-hop neighbors, and so on (up to $|\mathcal{V}|$-hop neighborhoods). **B-D**, Visualization of the output of three role-detection algorithms on the barbell graph, where the model outputs are projected using principal components analysis. RolX (**B**) [29] is a baseline approach based upon hand-designed features, while struc2vec (**C**) and GraphWave (**D**) use different representation learning approaches. Note that all methods correctly differentiate the ends of the barbells from the rest of the graph, but only GraphWave is able to correctly differentiate all the various roles. Note also that there are fewer visible nodes in part **D** compared to **A** because GraphWave maps identically colored (*i.e.*, structurally equivalent) nodes to the exact same position in the embedding space. Reprinted from [20].

where $\mathbf{G} = \text{diag}([g(\lambda_1), ..., g(\lambda_{|\mathcal{V}|})])$ and $\mathbf{v}_i$ is a one-hot indicator vector corresponding to v_i's row/column in the Laplacian.[8] Donnat et al. show that these $\boldsymbol{\psi}_{v_i}$ vectors implicitly relate to topological quantities, such as v_i's degree and the number of k-cycles v_i is involved in. They find that—with a proper choice of scale, s—WaveGraph is able to effectively capture structural information about a nodes role in a graph (Figure 9).

2.7 Applications of node embeddings

The most common use cases for node embeddings are for visualization, clustering, node classification, and link prediction, and each of these use cases is relevant to a number of application domains, ranging from computational social science to computational biology.

Visualization and pattern discovery. The problem of visualizing graphs in a 2D interface has a long history, with applications throughout data mining, the social sciences, and biology [17]. Node embeddings offer a powerful new paradigm for graph visualization: because nodes are mapped to real-valued vectors, researchers can easily leverage existing, generic techniques for visualization high-dimensional datasets [56, 54]. For example, node embeddings can be combined with well-known techniques such as t-SNE [56] or principal components analysis (PCA) in order to generate 2D visualizations of graphs [46, 53], which can be useful for discovering communities and other hidden structures (Figures 2 and 8).

Clustering and community detection. In a similar vein as visualization, node embeddings are a powerful tool for clustering related nodes, a task that has countless applications from computational biology (*e.g.*, discovering related drugs) to marketing (*e.g.*, discovering related products) [23]. Again, because each node is associated with real-valued vector embedding, it is possible to apply any generic clustering algorithm to the set of learned node embeddings (*e.g.*, k-means or DB-scan [22]). This offers an open-ended and powerful alternative to traditional community detection techniques, and it also opens up new methodological opportunities, since node embeddings can capture the functional or structural roles played by different nodes, rather than just community structure.

Node classification and semi-supervised learning. Node classification is perhaps the most common benchmark task used for evaluating node embeddings. In most cases, the node classification task is a form of semi-supervised learning, where labels are only available for a small proportion of nodes, with the goal being to label the full graph based only on this small initial seed set. Common applications of semi-supervised node classification include classifying proteins according to their biological function [27] and classifying documents,

[8] Note that Equation (20) can be efficiently approximated via Chebyshev polynomials [20].

videos, web pages, or individuals into different categories/communities [27, 34, 46, 53]. Recently, Hamilton et al. [28] introduced the task of inductive node classification, where the goal is to classify nodes that were not seen during training, *e.g.* classifying new documents in evolving information graphs or generalizing to unseen protein-protein interaction networks.

Link prediction. Node embeddings are also extremely useful as features for link prediction, where the goal is to predict missing edges, or edges that are likely to form in the future [3]. Link prediction is at the core of recommender systems and common applications of node embeddings reflect this deep connection, including predicting missing friendship links in social networks [53] and affinities between users and movies [55]. Link prediction also has important applications in computational biology. Many biological interaction graphs (*e.g.*, between proteins and other proteins, or drugs and diseases) are incomplete, since they rely on data obtained from costly lab experiments. Predicting links in these noisy graphs is an important method for automatically expanding biological datasets and for recommending new directions for wet-lab experimentation [40]. More generally, link prediction is closely related to statistical relational learning [24], where a common task is to predict missing relations between entities in a knowledge graph [42].

3 Embedding subgraphs

We now turn to the task of representation learning on (sub)graphs, where the goal is to encode a set of nodes and edges into a low-dimensional vector embedding. More formally, the goal is to learn a continuous vector representation, $\mathbf{z}_{\mathcal{S}} \in \mathbb{R}^d$, of an induced subgraph $\mathcal{G}[\mathcal{S}]$ of the full graph $\mathcal{G}$, where $\mathcal{S} \subseteq \mathcal{V}$. (Note that these methods can embed both subgraphs ($\mathcal{S} \subset \mathcal{V}$) as well as entire graphs ($\mathcal{S} = \mathcal{V}$).) The embedding, $\mathbf{z}_{\mathcal{S}}$, can then be used to make predictions about the entire subgraph; for example, one might embed graphs corresponding to different molecules to predict their therapeutic properties [21].

Representation learning on subgraphs is closely related to the design of graph kernels, which define a distance measure between subgraphs [57]. That said, we omit a detailed discussion of graph kernels, which is a large and rich research area of its own, and refer the reader to [57] for a detailed discussion. The methods we review differ from the traditional graph kernel literature primarily in that we seek to learn useful representations from data, rather than pre-specifying feature representations through a kernel function.

Many of the methods in this section build upon the techniques used to embed individual nodes, introduced in Section 2. However, unlike the node embedding setting, most subgraph embedding approaches are fully-supervised, being used for subgraph classification, where the goal is to predict a label associated with a particular subgraph. Thus, in this section we will focus on the various different approaches for generating the $\mathbf{z}_{\mathcal{S}}$ embeddings, with the assumption that these embeddings are being fed through a cross-entropy loss function, analogous to Equation (16).

3.1 Sets of node embeddings and convolutional approaches

There are several subgraph embedding techniques that can be viewed as direct extensions of the convolutional node embedding algorithms (described in Section 2.3.2). The basic intuition behind these approaches is that they equate subgraphs with sets of node embeddings. They use the convolutional neighborhood aggregation idea (*i.e.*, Algorithm 1) to generate embeddings for nodes and then use additional modules to aggregate sets of node embeddings corresponding to subgraphs. The primary distinction between the different approaches in this section is how they aggregate the set of node embeddings corresponding to a subgraph.

3.1.1 Sum-based approaches

For example, "convolutional molecular fingerprints" introduced by Duvenaud et al. [21] represent subgraphs in molecular graph representations by summing all the individual node embeddings in the subgraph:

$$\mathbf{z}_{\mathcal{S}} = \sum_{v_i \in \mathcal{S}} \mathbf{z}_i, \tag{21}$$

where the embeddings, $\{\mathbf{z}_i, \forall v_i \in \mathcal{S}\}$, are generated using Algorithm 1.

Dai et al. [16] employ an analogous sum-based approach but note that it has conceptual connections to mean-field inference: if the nodes in the graph are viewed as latent variables in a graphical model, then Algorithm 1 can be viewed as a form of mean-field inference where the message-passing operations have been replaced with differentiable neural network alternatives. Motivated by this connection, Dai et al. [16] also propose a modified encoder based on Loopy Belief Propagation [41]. Using the placeholders and notation from Algorithm 1, the basic idea behind this alternative is to construct intermediate embeddings, $\boldsymbol{\eta}_{i,j}$, corresponding to edges, $(i, j) \in \mathcal{E}$:

$$\boldsymbol{\eta}_{i,j}^k = \sigma(\mathbf{W}_{\mathcal{E}}^k \cdot \text{COMBINE}(\mathbf{x}_i, \text{AGGREGATE}(\boldsymbol{\eta}_{l,i}^{k-1}, \forall v_l \in \mathcal{N}(v_i) \setminus v_j\}))). \tag{22}$$

These edge embeddings are then aggregated to form the node embeddings:

$$\mathbf{z}^i = \sigma(\mathbf{W}_{\mathcal{V}}^k \cdot \text{COMBINE}(\mathbf{x}_i, \text{AGGREGATE}(\{\boldsymbol{\eta}_{i,l}^K, \forall v_l \in \mathcal{N}(v_i)\}))). \tag{23}$$

Once the embeddings are computed, Dai et al. [16], use a simple element-wise sum to combine the node embeddings for a subgraph, as in Equation (21).

3.1.2 Graph-coarsening approaches

Defferrard et al. [18] and Bruna et al. [8] also employ convolutional approaches, but instead of summing the node embeddings for the whole graph, they stack convolutional and "graph coarsening" layers (similar to the HARP approach in Section 2.2.2). In the graph coarsening layers, nodes are clustered together (using any graph clustering approach), and the clustered node embeddings are combined using element-wise max-pooling. After clustering, the new coarser graph is again fed through a convolutional encoder and the process repeats.

Unlike the convolutional approaches discussed in 2.3.2, Defferrard et al. [18] and Bruna et al. [8] also place considerable emphasis on designing convolutional encoders based upon the graph Fourier transform [15]. However, because the graph Fourier transform requires identifying and manipulating the eigenvectors of the graph Laplacian, naive versions of these approaches are necessarily $O(|\mathcal{V}|^3)$. State-of-the-art approximations to these spectral approaches (e.g., using Chebyshev polynomials) are conceptually similar to Algorithm 1, with some minor variations, and we refer the reader to Bronstein et al. [7] for a thorough discussion of these techniques.

3.1.3 Further variations

Other variants of the convolutional idea are proposed by Neipert et al. [43] and Kearnes et al. [33]. Both advocate alternative methods for aggregating sets of node embeddings corresponding to subgraphs: Kearnes et al. aggregate sets of nodes using "fuzzy" histograms instead of a sum, and they also employ edge embedding layers similar to [16]. Neipart et al. define an ordering on the nodes—*e.g.* using a problem specific ordering or by employing an off-the-shelf vertex coloring algorithm—and using this ordering, they concatenate the embeddings for all nodes and feed this concatenated vector through a standard convolutional neural network architecture.

3.2 Graph neural networks

In addition to the convolution-inspired subgraph embedding approaches discussed above, there is a related—and chronologically prior—line of work on "graph neural networks" (GNNs) [51]. Conceptually, the GNN idea is closely related to Algorithm 1. However, instead of aggregating information from neighbors, the intuition behind GNNs is that subgraphs can be viewed as specifying a "compute graph", *i.e.*, a recipe for accumulating and passing information between nodes.

In the original GNN framework [25, 51] every node, v_i, is initialized with a random embedding, $\mathbf{h}_i^0$ (node attributes are ignored), and at each iteration of the GNN algorithm nodes accumulate inputs from their neighbors using simple neural network layers:[9]

$$\mathbf{h}_i^k = \sum_{v_j \in \mathcal{N}(v_i)} \sigma(\mathbf{W}\mathbf{h}_j^{k-1} + \mathbf{b}), \tag{24}$$

where $\mathbf{W} \in \mathbb{R}^{d \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ are trainable parameters and σ is a non-linearity (*e.g.*, $\tanh$ or a rectified linear unit). Equation (24) is repeatedly applied in a recursive fashion until the embeddings converge, and special care must be taken during initialization to ensure convergence [51]. Once the embeddings have converged, they are aggregated for the entire (sub)graph and this aggregated embedding is used for subgraph classification. Any of the aggregation procedures described in Section 3.1 could be employed, but Scarselli et al. [51] also suggest that the aggregation can be done by introducing a "dummy" super-node that is connected to all nodes in the target subgraph.

Li et al. [38] extend and modify the GNN framework to use Gated Recurrent Units and back propagation through time [14], which removes the need to run the recursion in Equation (24) to convergence. Adapting the GNN framework to use modern recurrent units also allows Li et al. to leverage node attributes and to use the output of intermediate embeddings of subgraphs.

The GNN framework is highly expressive, but it is also computationally intensive compared to the convolutional approaches, due to the complexities of ensuring convergence [51] or running back propagation through time [38]. Thus, unlike the convolutional approaches, which are most commonly used to classify molecular graphs in large datasets, the GNN approach has been used for more complex, but smaller scale, tasks, *e.g.*, for approximate formal verification using graph-based representations of programs [38].

3.3 Applications of subgraph embeddings

The primary use case for subgraph embeddings is for subgraph classification, which has important applications in a number of areas. The most prominent application domain is for classifying the properties of graphs corresponding to different molecules [16, 21, 43, 33]. Subgraph embeddings can be used to classify or predict various properties of molecular graphs, including predicting the efficacy of potential solar cell materials [16], or predicting the therapeutic effect of candidate drugs [33]. More generally, subgraph embeddings have been used to classify images (after converting the image to a graph representation) [8], to predict whether a computer program satisfies certain formal properties [38], and to perform logical reasoning tasks [38].

4 Conclusion and future directions

Representation learning approaches for machine learning on graphs offer a power alternative to traditional feature engineering. In recent years, these approaches have consistently pushed the state of the art on tasks such as node classification and link prediction. However, much work remains to be done, both in improving the performance of these methods, and—perhaps more importantly—in developing consistent theoretical frameworks that future innovations can build upon.

[9] Other parameterizations and variations are discussed in [51].

4.1 Challenges to future progress

In this review, we attempted to unify a number of previous works, but the field as a whole still lacks a consistent theoretical framework—or set of frameworks—that precisely delineate the goals of representation learning on graphs. At the moment, the implicit goal of most works is to generate representations that perform well on a particular set of classification or link prediction benchmarks (and perhaps also generate qualitatively pleasing visualizations). However, the unchecked proliferation of disparate benchmarks and conceptual models presents a real risk to future progress, and this problem is only exacerbated by the popularity of node and graph embedding techniques across distinct, and somewhat disconnected, subfields within the machine learning and data mining communities. Moving forward as a field will require new theoretical work that more precisely describes the kinds of graph structures that we expect the learned representations to encode, how we expect the models to encode this information, and what constraints (if any) should be imposed upon on these learned latent spaces.

More developed theoretical foundations would not only benefit researchers in the field—*e.g.*, by informing consistent and meaningful benchmark tasks—these foundations would also allow application domain-experts to more effectively choose and differentiate between the various approaches. Current methods are often evaluated on a variety of distinct benchmarks that emphasize various different graph properties (*e.g.*, community structures, relationship strengths between nodes, or structural roles). However, many real-world applications are more focused, and it is not necessary to have representations that are generically useful for a wide variety of tasks. As a field, we need to make it clear what method should be used when, and prescribing such use-cases requires a more precise theoretical understanding of what exactly our learned representations are encoding.

4.2 Important open problems

In addition to the general challenges outlined above, there are a number of concrete open problems that remain to be addressed within the area of representation learning on graphs.

Scalability. While most of the works we reviewed are highly scalable in theory (*i.e.*, $O(|\mathcal{E}|)$ training time), there is still significant work to be done in scaling node and graph embedding approaches to truly massive datasets (*e.g.*, billions of nodes and edges). For example, most methods rely on training and storing a unique embedding for each individual node. Moreover, most evaluation setups assume that the attributes, embeddings, and edge lists of all nodes used for both training and testing can fit in main memory—an assumption that is at odds with the reality of most application domains, where graphs are massive, evolving, and often stored in a distributed fashion. Developing representation learning frameworks that are truly scalable to realistic production settings is necessary to prevent widening the disconnect between the academic research community and the application consumers of these approaches.

Decoding higher-order motifs. While much work in recent years has been dedicated to refining and improving the encoder algorithm used to generate node embeddings, most methods still rely on basic pairwise decoders, which predict pairwise relations between nodes and ignore higher-order graph structures involving more than two nodes. It is well-known that higher-order structural motifs are essential to the structure and function of complex networks [5], and developing decoding algorithms that are capable of decoding complex motifs is an important direction for future work.

Modeling dynamic, temporal graphs. Many application domains involve highly dynamic graphs where timing information is critical—*e.g.*, instant messaging networks or financial transaction graphs. However, we lack embedding approaches that can cope with the unique challenges presented by temporal graphs, such as the task of incorporating timing information about edges. Temporal graphs are becoming an increasingly important object of study [45], and extending graph embedding techniques to operate over them will open up a wide range of exciting application domains.

Reasoning about large sets of candidate subgraphs. A major technical limitation of current subgraph embedding approaches is that they require the target subgraphs to be pre-specified before the learning process.

However, many applications seek to *discover* subgraphs with certain properties, and these applications require models that can reason over the combinatorially large space of *possible* candidate subgraphs. For example, one might want to discover central subgraphs in a gene regulatory network, or uncover nefarious sub-communities in a social network. We need improved subgraph embedding approaches that can efficiently reason over large sets of candidate subgraphs, as such improvements are critical to expand the usefulness of subgraph embeddings beyond the task of basic subgraph classification.

Improving interpretability. Representation learning is attractive because it relieves much of the burden of hand designing features, but it also comes at a well-known cost of interpretability. We know that embedding-based approaches give state-of-the-art performance, but the fundamental limitations—and possible underlying biases—of these algorithms are relatively unknown. In order to move forward, care must be taken to develop new techniques to improve the interpretability of the learned representations, beyond visualization and benchmark evaluation. Given the complexities and representational capacities of these approaches, researchers must be ever vigilant to ensure that their methods are truly learning to represent relevant graph information, and not just exploiting statistical tendencies of benchmarks.

Acknowledgments

The authors thank Marinka Zitnik, Zoubin Ghahramani, Richard Turner, Stephen Bach, and Manan Ajay Shah for their helpful discussions and comments on early drafts. This research has been supported in part by NSF IIS-1149837, DARPA SIMPLEX, Stanford Data Science Initiative, and Chan Zuckerberg Biohub. W.L.H. was also supported by the SAP Stanford Graduate Fellowship and an NSERC PGS-D grant. The views and conclusions expressed in this material are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the above funding agencies, corporations, or the U.S. and Canadian governments.

References

[1] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A.J. Smola. Distributed large-scale natural graph factorization. In *WWW*, 2013.

[2] R. Angles and C. Gutierrez. Survey of graph database models. *ACM Computing Surveys*, 40(1):1, 2008.

[3] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *WSDM*, 2011.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 2002.

[5] A.R. Benson, D.F. Gleich, and J. Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[6] S. Bhagat, G. Cormode, and S. Muthukrishnan. Node classification in social networks. In *Social Network Data Analytics*, pages 115–148. 2011.

[7] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[8] J. Bruna, W. Zaremba, and Y. Szlam, A.and LeCun. Spectral networks and locally connected networks on graphs. In *ICLR*, 2014.

[9] S. Cao, W. Lu, and Q. Xu. Grarep: Learning graph representations with global structural information. In *KDD*, 2015.

[10] S. Cao, W. Lu, and Q. Xu. Deep neural networks for learning graph representations. In *AAAI*, 2016.

[11] B.P. Chamberlain, J. Clough, and M.P. Deisenroth. Neural embeddings of graphs in hyperbolic space. *arXiv preprint arXiv:1705.10359*, 2017.

[12] S. Chang, W. Han, J. Tang, G. Qi, C.C. Aggarwal, and T.S. Huang. Heterogeneous network embedding via deep architectures. In *KDD*, 2015.

[13] H. Chen, B. Perozzi, Y. Hu, and S. Skiena. Harp: Hierarchical representation learning for networks. *arXiv preprint arXiv:1706.07845*, 2017.

[14] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

[15] Fan RK Chung. *Spectral Graph Theory*. Number 92. American Mathematical Soc., 1997.

[16] H. Dai, B. Dai, and L. Song. Discriminative embeddings of latent variable models for structured data. In *ICML*, 2016.

[17] M.C.F. De Oliveira and H. Levkowitz. From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):378–394, 2003.

[18] M. Defferrard and P. Bresson, X.and Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, 2016.

[19] Y. Dong, N.V. Chawla, and A. Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD*, 2017.

[20] C. Donnat, M. Zitnik, D. Hallac, and J. Leskovec. Graph wavelets for structural role similarity in complex networks. *Under review*, 2017.

[21] D. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R.P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*, 2015.

[22] M. Ester, H. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

[23] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[24] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning*. MIT press, 2007.

[25] M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *IEEE International Joint Conference on Neural Networks*, 2005.

[26] P. Goyal and E. Ferrara. Graph embedding techniques, applications, and performance: A survey. *arXiv preprint arXiv:1605.09096*, 2017.

[27] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *KDD*, 2016.

[28] W.L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. *arXiv preprint, arXiv:1603.04467*, 2017.

[29] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li. Rolx: structural role extraction & mining in large graphs. In *KDD*, 2012.

[30] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[31] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[32] P. Hoff, A.E. Raftery, and M.S. Handcock. Latent space approaches to social network analysis. *JASA*, 97(460):1090–1098, 2002.

[33] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, 2016.

[34] T.N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2016.

[35] T.N. Kipf and M. Welling. Variational graph auto-encoders. In *NIPS Workshop on Bayesian Deep Learning*, 2016.

[36] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[37] J.A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[38] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel. Gated graph sequence neural networks. In *ICLR*, 2015.

[39] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology*, 58(7):1019–1031, 2007.

[40] Q. Lu and L. Getoor. Link-based classification. In *ICML*, volume 3, pages 496–503, 2003.

[41] K. Murphy, Y. Weiss, and M. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *UAI*, 1999.

[42] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

[43] M. Niepert, M. Ahmed, and K. Kutzkov. Learning convolutional neural networks for graphs. In *ICML*, 2016.

[44] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu. Asymmetric transitivity preserving graph embedding. In *KDD*, 2016.

[45] A. Paranjape, A. R. Benson, and J. Leskovec. Motifs in temporal networks. In *WSDM*, 2017.

[46] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *KDD*, 2014.

[47] B. Perozzi, V. Kulkarni, and S. Skiena. Walklets: Multiscale graph embeddings for interpretable network classification. *arXiv preprint arXiv:1605.02115*, 2016.

[48] Bryan Perozzi. *Local Modeling of Attributed Graphs: Algorithms and Applications*. PhD thesis, Stony Brook University, 2016.

[49] T. Pham, T. Tran, D.Q. Phung, and S. Venkatesh. Column networks for collective classification. In *AAAI*, 2017.

[50] L.F.R. Ribeiro, P.H.P. Saverese, and D.R. Figueiredo. struc2vec: Learning node representations from structural identity. In *KDD*, 2017.

[51] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[52] M. Schlichtkrull, T.N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling. Modeling relational data with graph convolutional networks. *arXiv preprint arXiv:1703.06103*, 2017.

[53] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In *WWW*, 2015.

[54] J. Tenenbaum, V. De Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[55] R. van den Berg, T.N. Kipf, and M. Welling. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263*, 2017.

[56] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 9:2579–2605, 2008.

[57] S.V.N. Vishwanathan, N.N. Schraudolph, R. Kondor, and K.M. Borgwardt. Graph kernels. *JMLR*, 11:1201–1242, 2010.

[58] D. Wang, P. Cui, and W. Zhu. Structural deep network embedding. In *KDD*, 2016.

[59] Z. Yang, W. Cohen, and R. Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *ICML*, 2016.

[60] M. Zitnik and J. Leskovec. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 2017.