Machine Learning on Graphs: A Model and Comprehensive Taxonomy

Ines Chami*†, Sami Abu-El-Haija‡, Bryan Perozzi††, Christopher Ré‡‡, and Kevin Murphy††

†Stanford University, Institute for Computational and Mathematical Engineering
‡University of Southern California, Information Sciences Institute
‡‡Stanford University, Department of Computer Science
††Google Research
{chami,chrismre}@cs.stanford.edu, sami@haija.org, bperozzi@acm.org, kpmurphy@google.com

January 19, 2021

Abstract

There has been a surge of recent interest in graph representation learning (GRL). GRL methods have generally fallen into three main categories, based on the availability of labeled data. The first, network embedding, focuses on learning unsupervised representations of relational structure. The second, graph regularized neural networks, leverages graphs to augment neural network losses with a regularization objective for semi-supervised learning. The third, graph neural networks, aims to learn differentiable functions over discrete topologies with arbitrary structure. However, despite the popularity of these areas there has been surprisingly little work on unifying the three paradigms. Here, we aim to bridge the gap between network embedding, graph regularization and graph neural networks. We propose a comprehensive taxonomy of GRL methods, aiming to unify several disparate bodies of work. Specifically, we propose the GRAPHEDM framework, which generalizes popular algorithms for semi-supervised learning (e.g. GraphSage, GCN, GAT), and unsupervised learning (e.g. DeepWalk, node2vec) of graph representations into a single consistent approach. To illustrate the generality of GRAPHEDM, we fit over thirty existing methods into this framework. We believe that this unifying view both provides a solid foundation for understanding the intuition behind these methods, and enables future research in the area.

*Work partially done during an internship at Google Research.

Contents

1 Introduction

Learning representations for complex structured data is a challenging task. In the last decade, many successful models have been developed for certain kinds of structured data, including data defined on a discretized Euclidean domain. For instance, sequential data, such as text or videos, can be modelled via recurrent neural networks, which can capture sequential information, yielding efficient representations as measured on machine translation and speech recognition tasks. Another example is convolutional neural networks (CNNs), which parameterize neural networks according to structural priors such as shift-invariance, and have achieved unprecedented performance in pattern recognition tasks such as image classification or speech recognition. These major successes have been restricted to particular types of data that have a simple relational structure (e.g. sequential data, or data following regular patterns).

In many settings, data is not nearly as regular: complex relational structures commonly arise, and extracting information from that structure is key to understanding how objects interact with each other. Graphs are a universal data structures that can represent complex relational data (composed of nodes and edges), and appear in multiple domains such as social networks, computational chemistry [55], biology [127], recommendation systems [78], semi-supervised learning [53], and others. For graph-structured data, it is challenging to define networks with strong structural priors, as structures can be arbitrary, and can vary significantly across different graphs and even different nodes within the same graph. In particular, operations like convolutions cannot be directly applied on irregular graph domains. For instance in images, each pixel has the same neighborhood structure, allowing to apply the same filter weights at multiple locations in the image. However in graphs, one can't define an ordering of node since each node might have a different neighborhood structure (Fig. 1). Furthermore, Euclidean convolutions strongly rely on geometric priors (e.g. shift invariance) which don't generalize to non-Euclidean domains (e.g. translations might not even be defined on non-Euclidean domains).

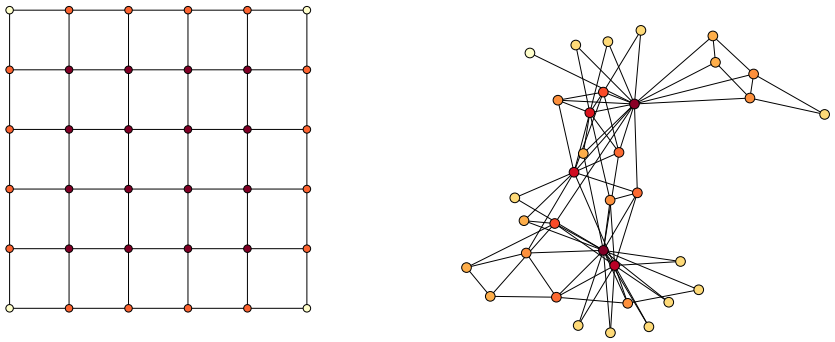
These challenges led to the development of Geometric Deep Learning (GDL) research which aims at applying deep learning techniques to non-Euclidean data. In particular, given the widespread prevalence of graphs in real-world applications, there has been a surge of interest in applying machine learning methods to graph-structured data. Among these, Graph Representation Learning (GRL) methods aim at learning low-dimensional continuous vector representations for graph-structured data, also called embeddings.

Broadly speaking, GRL can be divided into two classes of learning problems, **unsupervised** and **supervised** (or semi-supervised) GRL. The first family aims at learning low-dimensional Euclidean representations that preserve the structure of an input graph. The second family also learns low-dimensional Euclidean representations but for a specific downstream prediction task such as node or graph classification. Different from the unsupervised setting where inputs are usually graph structures, inputs in supervised settings are usually composed of different signals defined on graphs, commonly known as **node features**. Additionally, the underlying discrete graph domain can be fixed, which is the **transductive** learning setting (e.g. predicting user properties in a large social network), but can also vary in the **inductive** learning setting (e.g. predicting molecules attribute where each molecule is a graph). Finally, note that while most supervised and unsupervised methods learn representations in Euclidean vector spaces, there recently has been interest for **non-Euclidean representation learning**, which aims at learning non-Euclidean embedding spaces such as hyperbolic or spherical spaces. The main motivations for this body of work is to use a continuous embedding space that resembles the underlying discrete structure of the input data it tries to embed (e.g. the hyperbolic space is a continuous version of trees [119]).

Given the impressive pace at which the field of GRL is growing, we believe it is important to summarize and describe all methods in one unified and comprehensible framework. The goal of this survey is to provide a unified view of representation learning methods for graph-structured data, to better understand the different ways to leverage graph structure in deep learning models.

A number of graph representation learning surveys exist. First, there exist several surveys that cover shallow network embedding and auto-encoding techniques and we refer to [25, 32, 60, 65, 149] for a detailed overview of these methods. Second, Bronstein et al. [22] also gives an extensive overview of deep learning models for non-Euclidean data such as graphs or manifolds. Third, there have been several recent surveys [12, 141, 151, 153] covering methods applying deep learning to graphs, including graph neural networks. Most of these surveys focus on a specific sub-field of graph representation learning and do not draw connections between each sub-field.

In this work, we extend the encoder-decoder framework proposed by Hamilton et al. [65] and introduce a general framework, the Graph Encoder Decoder Model (GRAPHEDM), which allows us to group existing work into four major categories: (i) shallow embedding methods, (ii) auto-encoding methods, (iii) graph regularization methods, and (iv) graph neural networks (GNNs). Additionally, we introduce a Graph Convolution Framework (GCF), specifically



(a) Grid (Euclidean). (b) Arbitrary graph (Non-Euclidean).

Figure 1: An illustration of Euclidean vs. non-Euclidean graphs.

designed to describe convolution-based GNNs, which have achieved state-of-the art performance in a broad range of applications. This allows us to analyze and compare a variety of GNNs, ranging in construction from methods operating in the Graph Fourier[1] domain to methods applying self-attention as a neighborhood aggregation function [134]. We hope that this unified formalization of recent work would help the reader gain insights into the various learning methods on graphs to reason about similarities, differences, and point out potential extensions and limitations. That said, our contribution with regards to previous surveys are threefold:

- We introduce a general framework, GRAPHEDM, to describe a broad range of supervised and unsupervised methods that operate on graph-structured data, namely shallow embedding methods, graph regularization methods, graph auto-encoding methods and graph neural networks.
- Our survey is the first attempt to unify and view these different lines of work from the same perspective, and we provide a general taxonomy (Fig. 3) to understand differences and similarities between these methods. In particular, this taxonomy encapsulates over thirty existing GRL methods. Describing these methods within a comprehensive taxonomy gives insight to exactly how these methods differ.
- We release an open-source library for GRL which includes state-of-the-art GRL methods and important graph applications, including node classification and link prediction. Our implementation is publicly available at `https://github.com/google/gcnn-survey-paper`.

Organization of the survey We first review basic graph definitions and clearly state the problem setting for GRL (Section 2). In particular, we define and discuss the differences between important concepts in GRL, including the role of node features in GRL and how they relate to supervised GRL (Section 2.2.1), the distinctions between inductive and transductive learning (Section 2.2.2), positional and structural embeddings (Section 2.2.3) and the differences between supervised and unsupervised embeddings (Section 2.2.4). We then introduce GRAPHEDM (Section 3) a general framework to describe both supervised and unsupervised GRL methods, with or without the presence of node features, which can be applied in both inductive and transductive learning settings. Based on GRAPHEDM, we introduce a general taxonomy of GRL methods (Fig. 3) which encapsulates over thirty recent GRL models, and we describe both unsupervised (Section 4) and supervised (Section 5) methods using this taxonomy. Finally, we survey graph applications (Section 6).

[1] As defined by the eigenspace of the graph Laplacian.

2 Preliminaries

Here we introduce the notation used throughout this article (see Table 1 for a summary), and the generalized network embedding problem which graph representation learning methods aim to solve.

2.1 Definitions

Definition 2.1. (*Graph*). A graph G given as a pair: $G = (V, E)$, comprises a set of vertices (or nodes) $V = \{v_1, \ldots, v_{|V|}\}$ connected by edges $E = \{e_1, \ldots, e_{|E|}\}$, where each edge e_k is a pair (v_i, v_j) with $v_i, v_j \in V$. A graph is *weighted* if there exist a weight function: $w : (v_i, v_j) \to w_{ij}$ that assigns weight w_{ij} to edge connecting nodes $v_i, v_j \in V$. Otherwise, we say that the graph is *unweighted*. A graph is *undirected* if $(v_i, v_j) \in E$ implies $(v_j, v_i) \in E$, i.e. the relationships are symmetric, and *directed* if the existence of edge $(v_i, v_j) \in E$ does not necessarily imply $(v_j, v_i) \in E$. Finally, a graph can be *homogeneous* if nodes refer to one type of entity and edges to one relationship. It can be *heterogeneous* if it contains different types of nodes and edges.

For instance, social networks are homogeneous graphs that can be undirected (e.g. to encode symmetric relations like friendship) or directed (e.g. to encode the relation following); weighted (e.g. co-activities) or unweighted.

Definition 2.2. (*Path*). A path P is a sequence of edges $(u_{i_1}, u_{i_2}), (u_{i_2}, u_{i_3}), \ldots, (u_{i_k}, u_{i_{k+1}})$ of length k. A path is called simple if all u_{i_j} are distinct from each other. Otherwise, if a path visits a node more than once, it is said to contain a cycle.

Definition 2.3. (*Distance*). Given two nodes (u, v) in a graph G, we define the distance from u to v, denoted $d_G(u, v)$, to be the length of the shortest path from u to v, or ∞ if there exist no path from u to v.

The graph distance between two nodes is the analog of geodesic lengths on manifolds.

Definition 2.4. (*Vertex degree*). The *degree*, $\deg(v_i)$, of a vertex v_i in an unweighted graph is the number of edges incident to it. Similarly, the degree of a vertex v_i in a weighted graph is the sum of incident edges weights. The degree matrix D of a graph with vertex set V is the $|V| \times |V|$ diagonal matrix such that $D_{ii} = \deg(v_i)$.

Definition 2.5. (*Adjacency matrix*). A finite graph $G = (V, E)$ can be represented as a square $|V| \times |V|$ *adjacency matrix*, where the elements of the matrix indicate whether pairs of nodes are adjacent or not. The adjacency matrix is binary for unweighted graph, $A \in \{0, 1\}^{|V| \times |V|}$, and non-binary for weighted graphs $W \in \mathbb{R}^{|V| \times |V|}$. Undirected graphs have symmetric adjacency matrices, in which case, $\widetilde{W}$ denotes symmetrically-normalized adjacency matrix: $\widetilde{W} = D^{-1/2} W D^{-1/2}$, where D is the degree matrix.

Definition 2.6. (*Laplacian*). The *unnormalized Laplacian* of an undirected graph is the $|V| \times |V|$ matrix $L = D - W$. The *symmetric normalized Laplacian* is $\widetilde{L} = I - D^{-1/2} W D^{-1/2}$. The *random walk normalized Laplacian* is the matrix $L^{\text{rw}} = I - D^{-1} W$.

The name random walk comes from the fact that $D^{-1}W$ is a stochastic transition matrix that can be interpreted as the transition probability matrix of a random walk on the graph. The graph Laplacian is a key operator on graphs and can be interpreted as the analogue of the continuous Laplace-Beltrami operator on manifolds. Its eigenspace capture important properties about a graph (e.g. cut information often used for spectral graph clustering) but can also serve as a basis for smooth functions defined on the graph for semi-supervised learning [15]. The graph Laplacian is also closely related to the heat equation on graphs as it is the generator of diffusion processes on graphs and can be used to derive algorithms for semi-supervised learning on graphs [152].

Definition 2.7. (*First order proximity*). The *first order proximity* between two nodes v_i and v_j is a *local* similarity measure indicated by the edge weight w_{ij}. In other words, the first-order proximity captures the strength of an edge between node v_i and node v_j (should it exist).

Definition 2.8. (*Second-order proximity*). The *second order proximity* between two nodes v_i and v_j is measures the similarity of their neighborhood structures. Two nodes in a network will have a high second-order proximity if they tend to share many neighbors.

Note that there exist higher-order measures of proximity between nodes such as Katz Index, Adamic Adar or Rooted PageRank [89]. These notions of node proximity are particularly important in network embedding as many algorithms are optimized to preserve some order of node proximity in the graph.

	Notation	Meaning
Abbreviations	GRL	Graph Representation Learning
	GRAPHEDM	Graph Encoder Decoder Model
	GNN	Graph Neural Network
	GCF	Graph Convolution Framework
Graph notation	$G = (V, E)$	Graph with vertices (nodes) V and edges E
	$v_i \in V$	Graph vertex
	$d_G(\cdot, \cdot)$	Graph distance (length of shortest path)
	$\deg(\cdot)$	Node degree
	$D \in \mathbb{R}^{\lvert V\rvert \times \lvert V\rvert}$	Diagonal degree matrix
	$W \in \mathbb{R}^{\lvert V\rvert \times \lvert V\rvert}$	Graph weighted adjacency matrix
	$\widetilde{W} \in \mathbb{R}^{\lvert V\rvert \times \lvert V\rvert}$	Symmetric normalized adjacency matrix ($\widetilde{W} = D^{-1/2}WD^{-1/2}$)
	$A \in \{0, 1\}^{\lvert V\rvert \times \lvert V\rvert}$	Graph unweighted weighted adjacency matrix
	$L \in \mathbb{R}^{\lvert V\rvert \times \lvert V\rvert}$	Graph unnormalized Laplacian matrix ($L = D - W$)
	$\widetilde{L} \in \mathbb{R}^{\lvert V\rvert \times \lvert V\rvert}$	Graph normalized Laplacian matrix ($\widetilde{L} = I - D^{-1/2}WD^{-1/2}$)
	$L^{\mathrm{rw}} \in \mathbb{R}^{\lvert V\rvert \times \lvert V\rvert}$	Random walk normalized Laplacian ($L^{\mathrm{rw}} = I - D^{-1}W$)
GRAPHEDM notation	d_0	Input feature dimension
	$X \in \mathbb{R}^{\lvert V\rvert \times d_0}$	Node feature matrix
	d	Final embedding dimension
	$Z \in \mathbb{R}^{\lvert V\rvert \times d}$	Node embedding matrix
	d_ℓ	Intermediate hidden embedding dimension at layer ℓ
	$H^\ell \in \mathbb{R}^{\lvert V\rvert \times d_\ell}$	Hidden representation at layer ℓ
	$\mathcal{Y}$	Label space
	$y^S \in \mathbb{R}^{\lvert V\rvert \times \lvert \mathcal{Y}\rvert}$	Graph ($S = G$) or node ($S = N$) ground truth labels
	$\hat{y}^S \in \mathbb{R}^{\lvert V\rvert \times \lvert \mathcal{Y}\rvert}$	Predicted labels
	$s(W) \in \mathbb{R}^{\lvert V\rvert \times \lvert V\rvert}$	Target similarity or dissimilarity matrix in graph regularization
	$\widehat{W} \in \mathbb{R}^{\lvert V\rvert \times \lvert V\rvert}$	Predicted similarity or dissimilarity matrix
	$\mathrm{ENC}(\cdot; \Theta^E)$	Encoder network with parameters Θ^E
	$\mathrm{DEC}(\cdot; \Theta^D)$	Graph decoder network with parameters Θ^D
	$\mathrm{DEC}(\cdot; \Theta^S)$	Label decoder network with parameters Θ^S
	$\mathcal{L}^S_{\mathrm{SUP}}(y^S, \hat{y}^S; \Theta)$	Supervised loss
	$\mathcal{L}_{G,\mathrm{REG}}(W, \widehat{W}; \Theta)$	Graph regularization loss
	$\mathcal{L}_{\mathrm{REG}}(\Theta)$	Parameters' regularization loss
	$d_1(\cdot, \cdot)$	Matrix distance used for to compute the graph regularization loss
	$d_2(\cdot, \cdot)$	Embedding distance for distance-based decoders
	$\lvert\lvert \cdot \rvert\rvert_p$	$p-$norm
	$\lvert\lvert \cdot \rvert\rvert_F$	Frobenuis norm

Table 1: Summary of the notation used in the paper.

2.2 The generalized network embedding problem

Network embedding is the task that aims at learning a mapping function from a discrete graph to a continuous domain. Formally, given a graph $G = (V, E)$ with weighted adjacency matrix $W \in \mathbb{R}^{|V|\times|V|}$, the goal is to learn low-dimensional vector representations $\{Z_i\}_{i\in V}$ (embeddings) for nodes in the graph $\{v_i\}_{i\in V}$, such that important graph properties (e.g. local or global structure) are preserved in the embedding space. For instance, if two nodes have similar connections in the original graph, their learned vector representations should be close. Let $Z \in \mathbb{R}^{|V|\times d}$ denote the node[2] embedding matrix. In practice, we often want low-dimensional embeddings ($d \ll |V|$) for scalability purposes. That is, network embedding can be viewed as a dimensionality reduction technique for graph structured data, where the input data is defined on a non-Euclidean, high-dimensional, discrete domain.

2.2.1 Node features in network embedding

Definition 2.9. (*Vertex and edge fields*). A *vertex field* is a function defined on vertices $f : V \to \mathbb{R}$ and similarly an *edge field* is a function defined on edges: $F : E \to \mathbb{R}$. Vertex fields and edge fields can be viewed as analogs of scalar fields and tensor fields on manifolds.

Graphs may have node attributes (e.g. gender or age in social networks; article contents for citation networks) which can be represented as multiple vertex fields, commonly referred to as *node features*. In this survey, we denote node features with $X \in \mathbb{R}^{|V|\times d_0}$, where d_0 is the input feature dimension. Node features might provide useful information about a graph. Some network embedding algorithms leverage this information by learning mappings:

$$W, X \to Z.$$

In other scenarios, node features might be unavailable or not useful for a given task: network embedding can be *featureless*. That is, the goal is to learn graph representations via mappings:

$$W \to Z.$$

Note that depending on whether node features are used or not in the embedding algorithm, the learned representation could capture different aspects about the graph. If nodes features are being used, embeddings could capture both *structural* and *semantic* graph information. On the other hand, if node features are not being used, embeddings will only preserve structural information of the graph.

Finally, note that edge features are less common than node features in practice, but can also be used by embedding algorithms. For instance, edge features can be used as regularization for node embeddings [34], or to compute messages from neighbors as in message passing networks [55].

2.2.2 Transductive and inductive network embedding

Historically, a popular way of categorizing a network embedding method has been by whether the model can generalize to unseen data instances – methods are referred to as operating in either a *transductive* or *inductive* setting [143]. While we do not use this concept for constructing our taxonomy, we include a brief discussion here for completeness.

In transductive settings, it assumed that all nodes in the graph are observed in training (typically the nodes all come from one fixed graph). These methods are used to infer information about or between observed nodes in the graph (e.g. predicting labels for all nodes, given a partial labeling). For instance, if a transductive method is used to embed the nodes of a social network, it can be used to suggest new edges (e.g. friendships) between the nodes of the graph. One major limitation of models learned in transductive settings is that they fail to generalize to new nodes (e.g. evolving graphs) or new graph instances.

On the other hand, in inductive settings, models are expected to generalize to new nodes, edges, or graphs that were not observed during training. Formally, given training graphs $(G_1, \ldots, G_k)$, the goal is to learn a mapping to continuous representations that can generalize to unseen test graphs $(G_{k+1}, \ldots, G_{k+l})$. For instance, inductive learning can be used to embed molecular graphs, each representing a molecule structure [55], generalizing to new

[2] Although we present the model taxonomy via embedding nodes yielding $Z \in \mathbb{R}^{|V|\times d}$, it can also be extended for models that embed an entire graph i.e. with $Z \in \mathbb{R}^d$ as a d-dimensional vector for the whole graph (e.g. [7, 46]), or embed graph edges $Z \in \mathbb{R}^{|V|\times|V|\times d}$ as a (potentially sparse) 3D matrix with $Z_{u,v} \in \mathbb{R}^d$ representing the embedding of edge (u, v).

graphs and showing error margins within chemical accuracy on many quantum properties. Embedding dynamic or temporally evolving graphs is also another inductive graph embedding problem.

There is a strong connection between inductive graph embedding and *node features* (Section 2.2.1) as the latter are usually necessary for most inductive graph representation learning algorithms. More concretely, node features can be leveraged to learn embeddings with parametric mappings and instead of directly optimizing the embeddings, one can optimize the mapping's parameters. The learned mapping can then be applied to any node (even those that were not present a training time). On the other hand, when node features are not available, the first mapping from nodes to embeddings is usually a one-hot encoding which fails to generalize to new graphs where the canonical node ordering is not available.

Finally, we note that this categorization of graph embedding methods is at best an incomplete lens for viewing the landscape. While some models are inherently better suited to different tasks in practice, recent theoretical results [126] show that models previously assumed to be capable of only one setting (e.g. only transductive) can be used in both.

2.2.3 Positional vs structural network embedding

An emerging categorization of graph embedding algorithms is about whether the learned embeddings are positional or structural. Position-aware embeddings capture global relative positions of nodes in a graph and it is common to refer to embeddings as positional if they can be used to approximately reconstruct the edges in the graph, preserving distances such as shortest paths in the original graph [147]. Examples of positional embedding algorithms include random walk or matrix factorization methods. On the other hand, structure-aware embeddings capture local structural information about nodes in a graph, i.e. nodes with similar node features or similar structural roles in a network should have similar embeddings, regardless of how far they are in the original graph. For instance, GNNs usually learn embeddings by incorporating information for each node's neighborhood, and the learned representations are thus structure-aware.

In the past, positional embeddings have commonly been used for unsupervised tasks where positional information is valuable (e.g. link prediction or clustering) while structural embeddings have been used for supervised tasks (e.g. node classification or whole graph classification). More recently, there has been attempts to bridge the gap between positional and structural representations, with positional GNNs [147] and theoretical frameworks showing the equivalence between the two classes of embeddings [126].

2.2.4 Unsupervised and supervised network embedding

Network embedding can be *unsupervised* in the sense that the only information available is the graph structure (and possibly node features) or *supervised*, if additional information such as node or graph labels is provided. In unsupervised network embedding, the goal is to learn embeddings that preserved the graph structure and this is usually achieved by optimizing some reconstruction loss, which measures how well the learned embeddings can approximate the original graph. In supervised network embedding, the goal is to learn embeddings for a specific purpose such as predicting node or graph attributes, and models are optimized for a specific task such as graph classification or node classification. We use the level of supervision to build our taxonomy and cover differences between supervised and unsupervised methods in more details in Section 3.

3 A Taxonomy of Graph Embedding Models

We first describe our proposed framework, GRAPHEDM, a general framework for GRL (Section 3.1). In particular, GRAPHEDM is general enough that it can be used to succinctly describe over thirty GRL methods (both unsupervised and supervised). We use GRAPHEDM to introduce a comprehensive taxonomy in Section 3.2 and Section 3.3, which summarizes exiting works with shared notations and simple block diagrams, making it easier to understand similarities and differences between GRL methods.

3.1 The GRAPHEDM framework

The GRAPHEDM framework builds on top of the work of Hamilton et al. [65], which describes unsupervised network embedding methods from an encoder-decoder perspective. Cruz et al. [41] also recently proposed a modular encoder-based framework to describe and compare unsupervised graph embedding methods. Different from these unsupervised frameworks, we provide a more general framework which additionally encapsulates supervised graph embedding

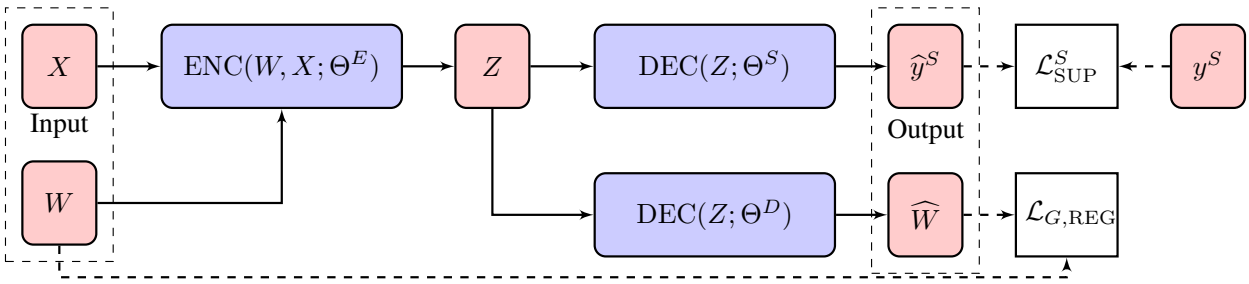


Figure 2: Illustration of the GRAPHEDM framework. Based on the supervision available, methods will use some or all of the branches. In particular, unsupervised methods do not leverage label decoding for training and only optimize the similarity or dissimilarity decoder (lower branch). On the other hand, semi-supervised and supervised methods leverage the additional supervision to learn models' parameters (upper branch).

methods, including ones utilizing the graph as a regularizer (e.g. [154]), and graph neural networks such as ones based on message passing [55, 120] or graph convolutions [23, 75].

Input The GRAPHEDM framework takes as input an undirected weighted graph $G = (V, E)$, with adjacency matrix $W \in \mathbb{R}^{|V|\times|V|}$, and optional node features $X \in \mathbb{R}^{|V|\times d_0}$. In (semi-)supervised settings, we assume that we are given training target labels for nodes (denoted N), edges (denoted E), and/or for the entire graph (denoted G). We denote the supervision signal as $S \in \{N, E, G\}$, as presented below.

Model The GRAPHEDM framework can be decomposed as follows:

- **Graph encoder network** $\text{ENC}_{\Theta^E} : \mathbb{R}^{|V|\times|V|} \times \mathbb{R}^{|V|\times d_0} \to \mathbb{R}^{|V|\times d}$, parameterized by Θ^E, which combines the graph structure with node features (or not) to produce node embedding matrix $Z \in \mathbb{R}^{|V|\times d}$ as:

$$Z = \text{ENC}(W, X; \Theta^E).$$

  As we shall see next, this node embedding matrix might capture different graph properties depending on the supervision used for training.

- **Graph decoder network** $\text{DEC}_{\Theta^D} : \mathbb{R}^{|V|\times d} \to \mathbb{R}^{|V|\times|V|}$, parameterized by Θ^D, which uses the node embeddings Z to compute similarity or dissimilarity scores for all node pairs, producing a matrix $\widehat{W} \in \mathbb{R}^{|V|\times|V|}$ as:

$$\widehat{W} = \text{DEC}(Z; \Theta^D).$$

- **Classification network** $\text{DEC}_{\Theta^S} : \mathbb{R}^{|V|\times d} \to \mathbb{R}^{|V|\times|\mathcal{Y}|}$, where $\mathcal{Y}$ is the label space. This network is used in (semi-)supervised settings and parameterized by Θ^S. The output is a distribution over the labels $\hat{y}^S$, using node embeddings, as:

$$\widehat{y}^S = \text{DEC}(Z; \Theta^S).$$

Our GRAPHEDM framework is general (see Fig. 2 for an illustration). Specific choices of the aforementioned (encoder and decoder) networks allows GRAPHEDM to realize specific graph embedding methods. Before presenting the taxonomy and showing realizations of various methods using our framework, we briefly discuss an application perspective.

Output The GRAPHEDM model can return a reconstructed graph similarity or dissimilarity matrix $\widehat{W}$ (often used to train *unsupervised* embedding algorithms), as well as a output labels $\widehat{y}^S$ for *supervised* applications. The label output space $\mathcal{Y}$ varies depending on the supervised application.

- **Node-level supervision,** with $\widehat{y}^N \in \mathcal{Y}^{|V|}$, where $\mathcal{Y}$ represents the node label space. If $\mathcal{Y}$ is categorical, then this is also known as (semi-)supervised node classification (Section 6.2.1), in which case the label decoder network produces labels for each node in the graph. If the embedding diemansions d is such that $d = |\mathcal{Y}|$,

then the label decoder network can be just a simple softmax activation across the rows of Z, producing a distribution over labels for each node. Additionally, the graph decoder network might also be used in supervised node-classification tasks, as it can be used to regularize embeddings (e.g. neighbor nodes should have nearby embeddings, regardless of node labels).

- **Edge-level supervision,** with $\widehat{y}^E \in \mathcal{Y}^{|V|\times|V|}$, where $\mathcal{Y}$ represents the edge label space. For example, $\mathcal{Y}$ can be multinomial in knowledge graphs (for describing the types of relationships between two entities), setting $\mathcal{Y} = \{0,1\}^{\#(\text{relation types})}$. It is common to have #(relation types) $= 1$, and this is is known as *link prediction*, where edge relations are binary. In this review, when $\widehat{y}^E = \{0,1\}^{|V|\times|V|}$ (i.e. $\mathcal{Y} = \{0,1\}$), then rather than naming the output of the decoder as $\widehat{y}^E$, we instead follow the nomenclature and position link prediction as an *unsupervised* task (Section 4). Then in lieu of $\widehat{y}^E$ we utilize $\widehat{W}$, the output of the graph decoder network (which is learned to reconstruct a target similarity or dissimilarity matrix) to rank potential edges.

- **Graph-level supervision,** with $\widehat{y}^G \in \mathcal{Y}$, where $\mathcal{Y}$ is the graph label space. In the graph classification task (Section 6.2.2), the label decoder network converts node embeddings into a single graph labels, using *graph pooling* via the graph edges captured by W. More concretely, the graph pooling operation is similar to pooling in standard CNNs, where the goal is to downsample local feature representations to capture higher-level information. However, unlike images, graphs don't have a regular grid structure and it is hard to define a pooling pattern which could be applied to every node in the graph. A possible way of doing so is via graph coarsening, which groups similar nodes into clusters to produce smaller graphs [44]. There exist other pooling methods on graphs such as DiffPool [145] or SortPooling [150] which creates an ordering of nodes based on their structural roles in the graph. Details about graph pooling operators is outside the scope of this work and we refer the reader to recent surveys [141] for a more in-depth treatment.

3.2 Taxonomy of objective functions

We now focus our attention on the optimization of models that can be described in the GRAPHEDM framework by describing the loss functions used for training. Let $\Theta = \{\Theta^E, \Theta^D, \Theta^S\}$ denote all model parameters. GRAPHEDM models can be optimized using a combination of the following loss terms:

- **Supervised loss** term, $\mathcal{L}^S_{\text{SUP}}$, which compares the predicted labels $\hat{y}^S$ to the ground truth labels y^S. This term depends on the task the model is being trained for. For instance, in semi-supervised node classification tasks ($S = N$), the graph vertices are split into labelled and unlabelled nodes ($V = V_L \cup V_U$), and the supervised loss is computed for each labelled node in the graph:

$$\mathcal{L}^N_{\text{SUP}}(y^N, \hat{y}^N; \Theta) = \sum_{i|v_i \in V_L} \ell(y^N_i, \hat{y}^N_i; \Theta),$$

  where $\ell(\cdot)$ is the loss function used for classification (e.g. cross-entropy). Similarly for graph classification tasks ($S = G$), the supervised loss is computed at the graph-level and can be summed across multiple training graphs:

$$\mathcal{L}^G_{\text{SUP}}(y^G, \hat{y}^G; \Theta) = \ell(y^G, \hat{y}^G; \Theta).$$

- **Graph regularization loss** term, $\mathcal{L}_{G,\text{REG}}$, which leverages the graph structure to impose regularization constraints on the model parameters. This loss term acts as a smoothing term and measures the distance between the decoded similarity or dissimilarity matrix $\widehat{W}$, and a target similarity or dissimilarity matrix $s(W)$, which might capture higher-order proximities than the adjacency matrix itself:

$$\mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta) = d_1(s(W), \widehat{W}), \tag{1}$$

  where $d_1(\cdot,\cdot)$ is a distance or dissimilarity function. Examples for such regularization are constraining neighboring nodes to share similar embeddings, in terms of their distance in L2 norm. We will cover more examples of regularization functions in Section 4 and Section 5.

- **Weight regularization loss** term, $\mathcal{L}_{\text{REG}}$, e.g. for representing prior, on trainable model parameters for reducing overfitting. The most common regularization is L2 regularization (assumes a standard Gaussian prior):

$$\mathcal{L}_{\text{REG}}(\Theta) = \sum_{\theta \in \Theta} ||\theta||_2^2.$$

Finally, models realizable by GRAPHEDM framework are trained by minimizing the total loss $\mathcal{L}$ defined as:

$$\mathcal{L} = \alpha \mathcal{L}^S_{\text{SUP}}(y^S, \hat{y}^S; \Theta) + \beta \mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta) + \gamma \mathcal{L}_{\text{REG}}(\Theta), \tag{2}$$

where α, β and γ are hyper-parameters, that can be tuned or set to zero. Note that graph embedding methods can be trained in a *supervised* ($\alpha \neq 0$) or *unsupervised* ($\alpha = 0$) fashion. Supervised graph embedding approaches leverage an additional source of information to learn embeddings such as node or graph labels. On the other hand, unsupervised network embedding approaches rely on the graph structure only to learn node embeddings.

A common approach to solve supervised embedding problems is to first learn embeddings with an unsupervised method (Section 4) and then train a supervised model on the learned embeddings. However, as pointed by Weston et al. [140] and others, using a two-step learning algorithm might lead to sub-optimal performances for the supervised task, and in general, supervised methods (Section 5) outperform two-step approaches.

3.3 Taxonomy of encoders

Having introduced all the building blocks of the GRAPHEDM framework, we now introduce our graph embedding taxonomy. While most methods we describe next fall under the GRAPHEDM framework, they will significantly differ based on the encoder used to produce the node embeddings, and the loss function used to learn model parameters. We divide graph embedding models into four main categories:

- **Shallow embedding methods**, where the encoder function is a simple embedding lookup. That is, the parameters of the model Θ^E are directly used as node embeddings:

$$\begin{aligned} Z &= \text{ENC}(\Theta^E) \\ &= \Theta^E \in \mathbb{R}^{|V| \times d}. \end{aligned}$$

  Note that shallow embedding methods rely on an embedding lookup and are therefore *transductive*, i.e. they generally cannot be directly applied in *inductive* settings where the graph structure is not fixed.

- **Graph regularization methods**, where the encoder network ignores the graph structure and only uses node features as input:

$$Z = \text{ENC}(X; \Theta^E).$$

  As its name suggests, graph regularization methods leverage the graph structure through the graph regularization loss term in Eq. (2) ($\beta \neq 0$) to regularize node embeddings.

- **Graph auto-encoding methods**, where the encoder is a function of the graph structure only:

$$Z = \text{ENC}(W; \Theta^E).$$

- **Neighborhood aggregation methods**, including graph convolutional methods, where both the node features and the graph structure are used in the encoder network. Neighborhood aggregation methods use the graph structure to propagate information across nodes and learn embeddings that encode structural properties about the graph:

$$Z = \text{ENC}(W, X; \Theta^E).$$

In what follows, we review recent methods for supervised and unsupervised graph embedding techniques using GRAPHEDM and summarize the proposed taxonomy in Fig. 3.

4 Unsupervised Graph Embedding

We now give an overview of recent unsupervised graph embedding approaches using the taxonomy described in the previous section. These methods map a graph, its nodes, and/or its edges, onto a continuous vector space, without using task-specific labels for the graph or its nodes. Some of these methods optimize an objective to learn an embedding that preserves the graph structure e.g. by learning to reconstruct some node-to-node similarity or dissimilarity matrix, such as the adjacency matrix. Some of these methods apply a contrastive objective, e.g. contrasting close-by node-pairs versus distant node-pairs [110]: nodes co-visited in short random walks should have a similarty score higher than distant ones, or contrasting real graphs versus fake ones [135]: the mutual information between a graph and all of its nodes, should be higher in real graphs than in fake graphs.

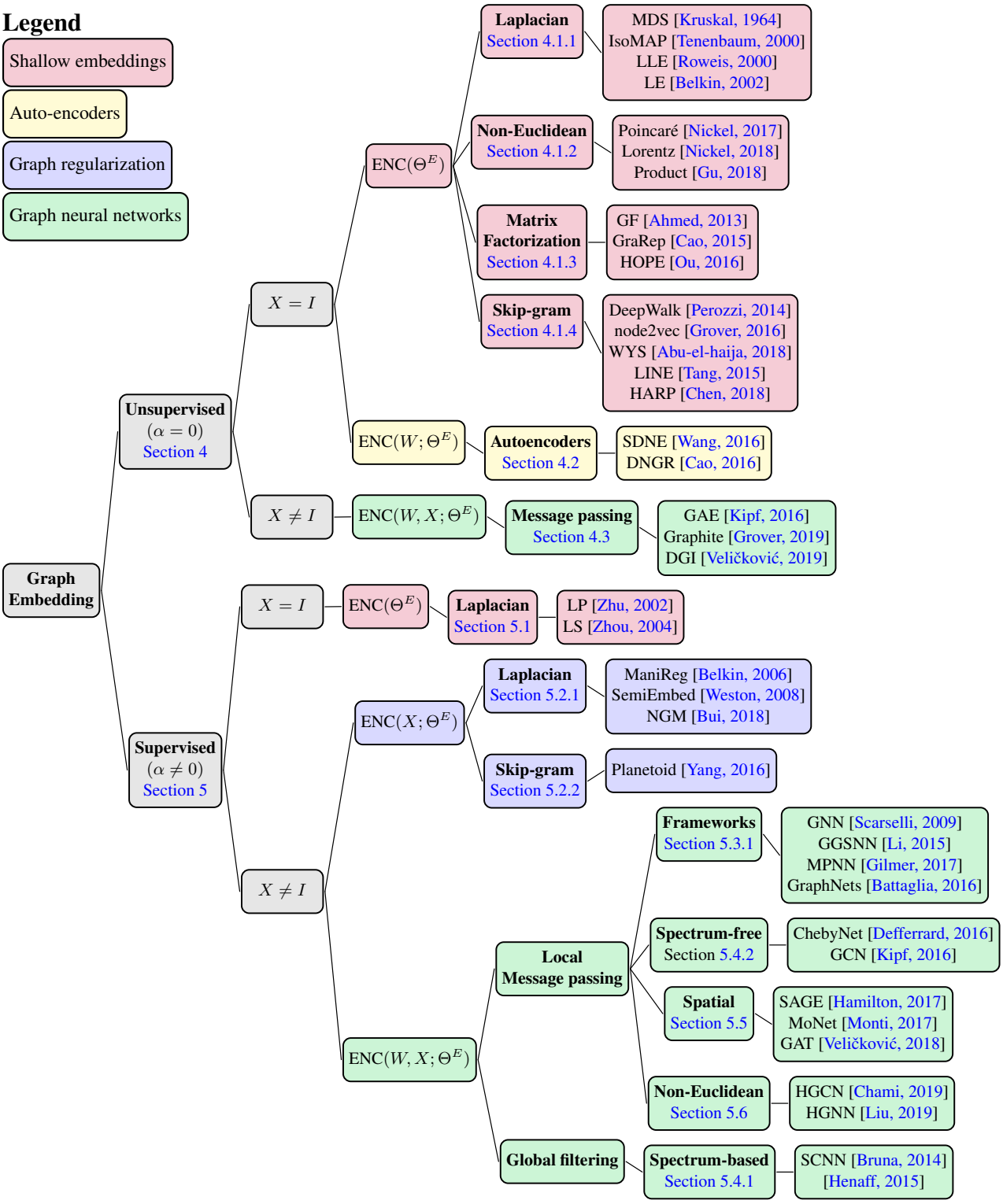


Figure 3: Taxonomy of graph representation learning methods. Based on what information is used in the encoder network, we categorize graph embedding approaches into four categories: shallow embeddings, graph auto-encoders, graph-based regularization and graph neural networks. Note that message passing methods can also be viewed as spatial convolution, since messages are computed over local neighborhood in the graph domain. Reciprocally, spatial convolutions can also be described using message passing frameworks.

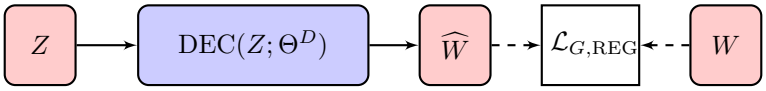


Figure 4: Shallow embedding methods. The encoder is a simple embedding look-up and the graph structure is only used in the loss function.

4.1 Shallow embedding methods

Shallow embedding methods are transductive graph embedding methods where the encoder function is a simple embedding lookup. More concretely, each node $v_i \in V$ has a corresponding low-dimensional learnable embedding vector $Z_i \in \mathbb{R}^d$ and the shallow encoder function is simply:

$$\begin{aligned} Z &= \mathrm{ENC}(\Theta^E) \\ &= \Theta^E \in \mathbb{R}^{|V| \times d}. \end{aligned}$$

Embeddings of nodes can be learned such that the structure of the data in the embedding space corresponds to the underlying graph structure. At a high level, this is similar to dimensionality reduction methods such as PCA, except that the input data might not have a linear structure. In particular, methods used for non-linear dimensionality reduction often start by building a discrete graph from the data (to approximate the manifold) and can be applied to graph embedding problems. Here, we analyze two major types of shallow graph embedding methods, namely *distance-based* and *outer product-based* methods.

Distance-based methods These methods optimize embeddings such that points that are close in the graph (as measured by their graph distances for instance) stay as close as possible in the embedding space using a predefined distance function. Formally, the decoder network computes pairwise distance for some distance function $d_2(\cdot,\cdot)$, which can lead to Euclidean (Section 4.1.1) or non-Euclidean (Section 4.1.2) embeddings:

$$\begin{aligned} \widehat{W} &= \mathrm{DEC}(Z;\Theta^D) \\ \text{with } \widehat{W}_{ij} &= d_2(Z_i, Z_j) \end{aligned}$$

Outer product-based methods These methods on the other hand rely on pairwise dot-products to compute node similarities and the decoder network can be written as:

$$\begin{aligned} \widehat{W} &= \mathrm{DEC}(Z;\Theta^D) \\ &= ZZ^\top. \end{aligned}$$

Embeddings are then learned by minimizing the graph regularization loss: $\mathcal{L}_{G,\mathrm{REG}}(W, \widehat{W}; \Theta) = d_1(s(W), \widehat{W})$. Note that for distance-based methods, the function $s(\cdot)$ measures dissimilarity or distances between nodes (higher values mean less similar pairs of nodes), while in outer-product methods, it measures some notion of similarity in the graph (higher values mean more similar pairs).

4.1.1 Distance-based: Euclidean methods

Most distance-based methods optimize Euclidean embeddings by minimizing Euclidean distances between similar nodes. Among these, we find linear embedding methods such as PCA or MDS, which learn low-dimensional linear projection subspaces, or nonlinear methods such as Laplacian eigenmaps, IsoMAP and Local linear embedding. Note that all these methods have originally been introduced for dimensionality reduction or visualization purposes, but can easily be extended to the context of graph embedding.

Multi-Dimensional Scaling (MDS) [80] refers to a set of embedding techniques used to map objects to positions while preserving the distances between these objects. In particular, metric MDS (mMDS) [40] minimizes the regularization loss in Eq. (1) with $s(W)$ set to some distance matrix measuring the dissimilarity between objects (e.g.

Euclidean distance between points in a high-dimensional space):

$$d_1(s(W), \widehat{W}) = \left(\frac{\sum_{ij} (s(W)_{ij} - \widehat{W}_{ij})^2}{\sum_{ij} s(W)_{ij}^2} \right)^{1/2}$$
$$\widehat{W}_{ij} = d_2(Z_i, Z_j) = ||Z_i - Z_j||_2.$$

That is, mMDS finds an embedding configuration where distances in the low-dimensional embedding space are preserved by minimizing a residual sum of squares called the *stress* cost function. Note that if the dissimilarities are computed from Euclidean distances of a higher-dimensional representation, then mMDS is equivalent to the PCA dimensionality reduction method. Finally, there exist variants of this algorithm such as non-metric MDS, when the dissimilarity matrix $s(W)$ is not a distance matrix, or classical MDS (cMDS) which can be solved in closed form using a low-rank decomposition of the gram matrix.

Isometric Mapping (IsoMap) [129] is an algorithm for non-linear dimensionality reduction which estimates the intrinsic geometry of a data lying on a manifold. This method is similar to MDS, except for a different choice of the distance matrix. IsoMap approximates manifold distances (in contrast with straight-line Euclidean geodesics) by first constructing a discrete neighborhood graph G, and then using the graph distances (length of shortest paths computed using Dijkstra's algorithm for example) to approximate the manifold geodesic distances:

$$s(W)_{ij} = d_G(v_i, v_j).$$

IsoMAP then uses the cMDS algorithm to compute representations that preserve these graph geodesic distances. Different from cMDS, IsoMAP works for distances that do not necessarily come from a Euclidean metric space (e.g. data defined on a Riemannian manifold). It is however computationally expensive due to the computation of all pairs of shortest path lengths in the neighborhood graph.

Locally Linear Embedding (LLE) [116] is another non-linear dimensionality reduction technique which was introduced around the same time as IsoMap and improves over its computational complexity via sparse matrix operations. Different from IsoMAP which preserves the global geometry of manifolds via geodesics, LLE is based on the local geometry of manifolds and relies on the assumptions that when locally viewed, manifolds are approximately linear. The main idea behind LLE is to approximate each point using a linear combination of embeddings in its local neighborhood (linear patches). These local neighborhoods are then compared globally to find the best non-linear embedding.

Laplacian Eigenmaps (LE) [14] is a non-linear dimensionality reduction methods that seeks to preserve *local* distances. Spectral properties of the graph Laplacian matrix capture important structural information about graphs. In particular, eigenvectors of the graph Laplacian provide a basis for smooth functions defined on the graph vertices (the "smoothest" function being the constant eigenvector corresponding to eigenvalue zero). LE is a non-linear dimensionality reduction technique which builds on this intuition. LE first constructs a graph from datapoints (e.g. k-NN graph or ε-neighborhood graph) and then represents nodes in the graphs via the Laplacian's eigenvectors corresponding to smaller eigenvalues. The high-level intuition for LE is that points that are close on the manifold (or graph) will have similar representations, due to the "smoothness" of Laplacian's eigenvectors with small eigenvalues. Formally, LE learns embeddings by solving the generalized eigenvector problem:

$$\min_{Z \in \mathbb{R}^{|V| \times d}} Z^\top L\, Z$$
$$\text{subject to } Z^\top D Z = I \text{ and } Z^\top D 1 = 0,$$

where the first constraint removes an arbitrary scaling factor in the embedding and the second one removes trivial solutions corresponding to the constant eigenvector (with eigenvalue zero for connected graphs). Further, note that $Z^\top L\, Z = \frac{1}{2} \sum_{ij} W_{ij} ||Z_i - Z_j||_2^2$ and therefore the minimization objective can be equivalently written as a graph regularization term using our notations:

$$d_1(W, \widehat{W}) = \sum_{ij} W_{ij} \widehat{W}_{ij}$$
$$\widehat{W}_{ij} = d_2(Z_i, Z_j) = ||Z_i - Z_j||_2^2.$$

Therefore, LE learns embeddings such that the Euclidean distance in the embedding space is small for points that are close on the manifold.

4.1.2 Distance-based: Non-Euclidean methods

The distance-based methods described so far assumed embeddings are learned in a Euclidean space. Graphs are non-Euclidean discrete data structures, and several works proposed to learn graph embeddings into non-Euclidean spaces instead of conventional Euclidean space. Examples of such spaces include the hyperbolic space, which has a non-Euclidean geometry with a constant negative curvature and is well-suited to represent hierarchical data.

To give more intuition, the hyperbolic space can be thought of as continuous versions of trees, where geodesics (generalization of shortest paths on manifolds) resemble shortest paths in discrete trees. Further, the volume of balls grows exponentially with radius in hyperbolic space, similar to trees where the number of nodes within some distance to the root grows exponentially. In contrast, this volume growth is only polynomial in Euclidean space and therefore, the hyperbolic space has more "room" to fit complex hierarchies and compress representations. In particular, hyperbolic embeddings can embed trees with arbitrary low distortion in just two-dimensions [119] whereas this is not possible in Euclidean space. This makes hyperbolic space a natural candidate to embed tree-like data and more generally, hyperbolic geometry offers an exciting alternative to Euclidean geometry for graphs that exhibit hierarchical structures, as it enables embeddings with much smaller distortion.

Before its use in machine learning applications, hyperbolic geometry has been extensively studied and used in network science research. Kleinberg [77] proposed a greedy algorithm for geometric rooting, which maps nodes in sensor networks to coordinates on a hyperbolic plane via spanning trees, and then performs greedy geographic routing. Hyperbolic geometry has also been used to study the structural properties of complex networks (networks with non-trivial topological features used to model real-world systems). Krioukov et al. [79] develop a geometric framework to construct scale-free networks (a family of complex networks with power-law degree distributions), and conversely show that any scale-free graph with some metric structure has an underlying hyperbolic geometry. Papadopoulos et al. [106] introduce the Popularity-Similarity (PS) framework to model the evolution and growth of complex networks. In this model, new nodes are likely to be connected to popular nodes (modelled by their radial coordinates in hyperbolic space) as well as similar nodes (modelled by the angular coordinates). This framework has further been used to map nodes in graphs to hyperbolic coordinates, by maximising the likelihood that the network is produced by the PS model [107]. Further works extend non-linear dimensionality reduction techniques such as LLE [14] to efficiently map graphs to hyperbolic coordinates [8, 100].

More recently, there has been interest in learning hyperbolic representations of hierarchical graphs or trees, via gradient-based optimization. We review some of these machine learning-based algorithms next.

Poincaré embeddings Nickel and Kiela [101] learn embeddings of hierarchical graphs such as lexical databases (e.g. WordNet) in the Poincaré model hyperbolic space. Using our notations, this approach learns hyperbolic embeddings via the Poincaré distance function:

$$\begin{aligned} d_2(Z_i, Z_j) &= d_{\text{Poincaré}}(Z_i, Z_j) \\ &= \operatorname{arcosh}\left(1 + 2\frac{||Z_i - Z_j||_2^2}{(1 - ||Z_i||_2^2)(1 - ||Z_j||_2^2)}\right). \end{aligned}$$

Embeddings are then learned by minimizing distances between connected nodes while maximizing distances between disconnected nodes:

$$d_1(W, \widehat{W}) = -\sum_{ij} W_{ij} \log \frac{e^{-\widehat{W}_{ij}}}{\sum_{k|W_{ik}=0} e^{-\widehat{W}_{ik}}} = -\sum_{ij} W_{ij} \log \operatorname{Softmax}_{k|W_{ik}=0}(-\widehat{W}_{ij}),$$

where the denominator is approximated using negative sampling. Note that since the hyperbolic space has a manifold structure, embeddings need to be optimized using Riemannian optimization techniques [19] to ensure that they remain on the manifold.

Other variants of these methods have been proposed. In particular, Nickel and Kiela [102] explore a different model of hyperbolic space, namely the Lorentz model (also known as the hyperboloid model), and show that it provides better numerical stability than the Poincaré model. Another line of work extends non-Euclidean embeddings to mixed-curvature product spaces [63], which provide more flexibility for other types of graphs (e.g. ring of trees). Finally, Chamberlain et al. [28] extend Poincaré embeddings to incorporate skip-gram losses using hyperbolic inner products.

4.1.3 Outer product-based: Matrix factorization methods

Matrix factorization approaches learn embeddings that lead to a low rank representation of some similarity matrix $s(W)$, where $s : \mathbb{R}^{|V|\times|V|} \to \mathbb{R}^{|V|\times|V|}$ is a transformation of the weighted adjacency matrix, and many methods set it to the identity, i.e. $s(W) = W$. Other transformations include the Laplacian matrix or more complex similarities derived from proximity measures such as the Katz Index, Common Neighbours or Adamic Adar. The decoder function in matrix factorization methods is a simple outer product:

$$\widehat{W} = \text{DEC}(Z; \Theta^D) = ZZ^\top. \tag{3}$$

Matrix factorization methods learn embeddings by minimizing the regularization loss in Eq. (1) with:

$$\mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta) = ||s(W) - \widehat{W}||_F^2. \tag{4}$$

That is, $d_1(\cdot, \cdot)$ in Eq. (1) is the Frobenius norm between the reconstructed matrix and the target similarity matrix. By minimizing the regularization loss, graph factorization methods learn low-rank representations that preserve structural information as defined by the similarity matrix $s(W)$ and we now review important matrix factorization methods.

Graph factorization (GF) [6] learns a low-rank factorization for the adjacency matrix by minimizing graph regularization loss in Eq. (1) using:

$$d_1(W, \widehat{W}) = \sum_{(v_i,v_j)\in E} (W_{ij} - \widehat{W}_{ij})^2.$$

Note that if A is the binary adjacency matrix, that is $A_{ij} = 1$ *iif* $(v_i, v_j) \in E$ and $A_{ij} = 0$ otherwise, then we can express the graph regularization loss in terms of Frobenius norm:

$$\mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta) = ||A \cdot (W - \widehat{W})||_F^2,$$

where $\cdot$ is the element-wise matrix multiplication operator. Therefore, GF also learns a low-rank factorization of the adjacency matrix W measured in Frobenuis norm. Note that the sum is only over existing edges in the graph, which reduces the computational complexity of this method from $O(|V|^2)$ to $O(|E|)$.

Graph representation with global structure information (GraRep) [26] The methods described so far are all symmetric, that is, the similarity score between two nodes (v_i, v_j) is the same a the score of (v_j, v_i). This might be a limiting assumption when working with directed graphs as some nodes can be strongly connected in one direction and disconnected in the other direction. overcomes this limitation by learning two embeddings per node, a source embedding Z^s and a target embedding Z^t, which capture asymmetric proximity in directed networks. GraRep learns embeddings that preserve k-hop neighborhoods via powers of the adjacency and minimizes the graph regularization loss with:

$$\widehat{W}^{(k)} = Z^{(k),s} {Z^{(k),t}}^\top$$
$$\mathcal{L}_{G,\text{REG}}(W, \widehat{W}^{(k)}; \Theta) = ||D^{-k}W^k - \widehat{W}^{(k)}||_F^2,$$

for each $1 \le k \le K$. GraRep concatenates all representations to get source embeddings $Z^s = [Z^{(1),s}| \dots |Z^{(K),s}]$ and target embeddings $Z^t = [Z^{(1),t}| \dots |Z^{(K),t}]$. Finally, note that GraRep is not very scalable as the powers of $D^{-1}W$ might be dense matrices.

HOPE [104] Similar to GraRep, HOPE learns asymmetric embeddings but uses a different similarity measure. The distance function in HOPE is simply the Frobenius norm and the similarity matrix is a high-order proximity matrix (e.g. Adamic-Adar):

$$\widehat{W} = Z^s {Z^t}^\top$$
$$\mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta) = ||s(W) - \widehat{W}||_F^2.$$

The similarity matrix in HOPE is computed with sparse matrices, making this method more efficient and scalable than GraRep.

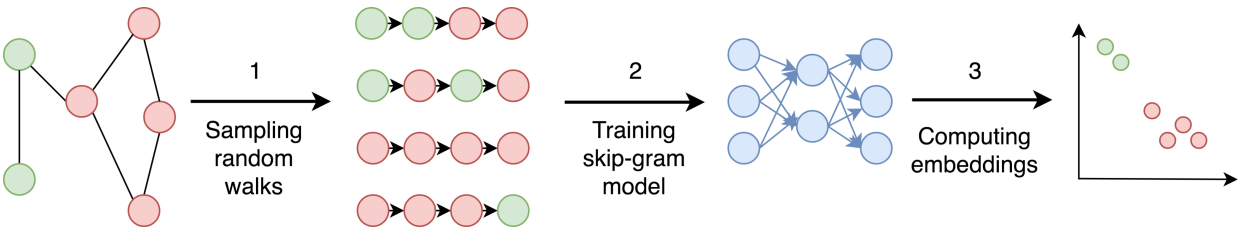


Figure 5: An overview of the pipeline for random-walk graph embedding methods. Reprinted with permission from [56].

4.1.4 Outer product-based: Skip-gram methods

Skip-gram graph embedding models were inspired by efficient NLP methods modeling probability distributions over words for learning word embeddings [97, 109]. Skip-gram word embeddings are optimized to predict context words, or surrounding words, for each target word in a sentence. Given a sequence of words $(w_1, \ldots, w_T)$, skip-gram will minimize the objective:

$$\mathcal{L} = - \sum_{-K \leq i \leq K, i \neq 0} \log \mathbb{P}(w_{k-i} | w_k),$$

for each target words w_k. In practice, the conditional probabilities can be estimated using neural networks, and skip-gram methods can be trained efficiently using negative sampling.

Perozzi et al. [110] empirically show the frequency statistics induced by random walks also follow Zipf's law, thus motivating the development of skip-gram graph embedding methods. These methods exploit random walks on graphs and produce node sequences that are similar in positional distribution, as to words in sentences. In skip-gram graph embedding methods, the decoder function is also an outer product (Eq. (3)) and the graph regularization term is computed over random walks on the graph.

DeepWalk [110] was the first attempt to generalize skip-gram models to graph-structured data. DeepWalk draws analogies between graphs and language. Specifically, writing a sentence is analogous to performing a random walk, where the sequence of nodes visited during the walk, is treated as the words of the sentence. DeepWalk trains neural networks by maximizing the probability of predicting context nodes for each target node in a graph, namely nodes that are close to the target node in terms of hops and graph proximity. For this purpose, node embeddings are decoded into probability distributions over nodes using row-normalization of the decoded matrix with softmax.

To train embeddings, DeepWalk generates sequences of nodes using truncated unbiased random walks on the graph—which can be compared to sentences in natural language models—and then maximize their log-likelihood. Each random walk starts with a node $v_{i_1} \in V$ and repeatedly sample next node at uniform: $v_{i_{j+1}} \in \{v \in V \mid (v_{i_j}, v) \in E\}$. The walk length is a hyperparameter. All generated random-walk can then be passed to an NLP-embedding algorithm e.g. word2vec's Skipgram model. This two-step paradigm introduced by Perozzi et al. [110] is followed by many subsequent works, such as node2vec [61].

We note that is common for underlying implementations to use two distinct representations for each node (one for when a node is center of a truncated random walk, and one when it is in the context). The implications of this modeling choice is studied further in [2].

Abu-El-Haija et al. [3] show that training DeepWalk, in expectation, is equivalent to first sampling integer $q \sim [1, 2, \ldots, T_{\max}]$ with mass $\propto [1, \frac{T_{\max}-1}{T_{\max}}, \ldots, \frac{1}{T_{\max}}]$. Specifically, if $s(W) = \mathbb{E}_q \left[\left(D^{-1} W \right)^q \right]$, then training DeepWalk is equivalent to minimizing:

$$\mathcal{L}_{G,\mathrm{REG}}(W, \widehat{W}; \Theta) = \log C - \sum_{v_i \in V, v_j \in V} s(W)_{ij} \widehat{W}_{ij}, \tag{5}$$

where $C = \prod_i \sum_j \exp(\widehat{W}_{ij})$ is a normalizing constant. Note that computing C requires summing over all nodes in the graph which is computationally expensive. DeepWalk overcomes this issue by using a technique called hierarchical softmax, which computes C efficiently using binary trees. Finally, note that by computing truncated random walks on the graph, DeepWalk embeddings capture high-order node proximity.

	Method	Model	$s(W)_{ij}$	$\mathrm{DEC}(Z;\Theta^D)_{ij}$	$d_1(W \leftarrow s(W), \widehat{W} \leftarrow \mathrm{DEC}(Z;\Theta^D))$	order of proximity
Distance $\mathrm{DEC}(Z;\Theta^D)_{ij}$ $\parallel$ $d_2(Z_i,Z_j)$	Euclidean	mMDS	distance matrix	$\|\|Z_i - Z_j\|\|_2$	$\left(\frac{\sum_{ij}(W_{ij}-\widehat{W}_{ij})^2}{\sum_{ij} W_{ij}^2}\right)^{1/2}$	high
		LE	W_{ij}	$\|\|Z_i - Z_j\|\|_2^2$	$\sum_{ij} W_{ij}\widehat{W}_{ij}$	1st
	Non-Euclidean	Poincaré	W_{ij}	$d_{\text{Poincaré}}(Z_i, Z_j)$	$-\sum_{ij} W_{ij} \log \mathrm{Softmax}_{k\|W_{ik}=0}(-\widehat{W}_{ij})$	1st
Outer product $\mathrm{DEC}(Z;\Theta^D)_{ij}$ $\parallel$ $Z_i^\top Z_j$	Matrix Factorization	GF	W_{ij}	$Z_i^\top Z_j$	$\sum_{ij\|W_{ij}>0}(W_{ij} - \widehat{W}_{ij})^2$	1st
		GraRep	$(D^{-k}W^k)_{ij}$	$Z_i^{(k),s^\top} Z_j^{(k),t}$	$\|\|W - \widehat{W}\|\|_F^2$	k^{th}
		HOPE	$s(W)_{ij}$	$Z_i^{s\top} Z_j^\top$	$\|\|W - \widehat{W}\|\|_F^2$	high
	Skip-gram	DeepWalk	$\propto \mathbb{E}_q\left[(D^{-1}W)^q\right]_{ij}$	$Z_i^\top Z_j$	$-\sum_{ij} W_{ij} \log \mathrm{Softmax}_j(\widehat{W}_{ij})$	high
		node2vec	$\mathrm{n2vWalk}(W;p,q)_{ij}$	$Z_i^\top Z_j$	$-\sum_{ij} W_{ij} \log \mathrm{Softmax}_j(\widehat{W}_{ij})$	high
		WYS	$\propto \mathbb{E}_q\left[(D^{-1}W)^q\right]_{ij}$	$Z_i^\top Z_j$	$\mathrm{BCE}(W, \widehat{W})$	high

Table 2: An overview of unsupervised shallow embedding methods, where the encoding function is a simple embedding look-up $Z = \mathrm{ENC}(\Theta^E)$. Softmax represents sampled/hierarchical softmax; $\propto$ for approximating random walks; n2vWalk is a traversal algorithm with (back) teleportation (approximates combination of BFS & DFS). BCE is the sigmoid cross entropy loss for binary classification.

node2vec [61] is a random-walk based approach for unsupervised network embedding, that extends DeepWalk's sampling strategy. The authors introduce a technique to generate biased random walks on the graph, by combining graph exploration through breadth first search (BFS) and through depth first search (DFS). Intuitively, node2vec also preserves high order proximities in the graph but the balance between BFS and DFS allows node2vec embeddings to capture local structures in the graph, as well as global community structures, which can lead to more informative embeddings. Finally, note that negative sampling [97] is used to approximate the normalization factor C in Eq. (5).

Watch Your Step (WYS) [3] Random walk methods are very sensitive to the sampling strategy used to generate random walks. For instance, some graphs may require shorter walks if local information is more informative that global graph structure, while in other graphs, global structure might be more important. Both DeepWalk and node2vec sampling strategies use hyper-parameters to control this, such as the length of the walk or ratio between breadth and depth exploration. Optimizing over these hyper-parameters through grid search can be computationally expensive and can lead to sub-optimal embeddings. WYS learns such random walk hyper-parameters to minimize the overall objective (in analogy: each graph gets to choose its own preferred "context size", such that the probability of predicting random walks is maximized). WYS shows that, when viewed in expectation, these hyperparameters only correspond in the objective to coefficients to the powers of the adjacency matrix $(W^k)_{1\leq k\leq K}$. These coefficients are denoted $q = (q_k)_{1\leq k\leq K}$ and are learned through back-propagation. Should q's learn a left-skewed distribution, then the embedding would prioritize local information and right-skewed distribution will enhance high-order relationships and graph global structure. This concept has been extended to other forms of attention to the 'graph context', such using a personalized context distributions for each node [70].

Large scale Information Network Embedding (LINE) [128] learns embeddings that preserve first and second order proximity. To learn first order proximity preserving embeddings, LINE minimizes the graph regularization loss:

$$\widehat{W}_{ij}^{(1)} = {Z_i^{(1)}}^\top Z_j^{(1)}$$
$$\mathcal{L}_{G,\mathrm{REG}}(W, \widehat{W}^{(1)}; \Theta) = -\sum_{(i,j)|(v_i,v_j)\in E} W_{ij} \log \sigma(\widehat{W}_{ij}^{(1)}).$$

LINE also assumes that nodes with multiple edges in common should have similar embeddings and learns second-order proximity preserving embeddings by minimizing:

$$\widehat{W}_{ij}^{(2)} = {Z_i^{(2)}}^\top Z_j^{(2)}$$
$$\mathcal{L}_{G,\mathrm{REG}}(W, \widehat{W}^{(2)}; \Theta) = -\sum_{(i,j)|(v_i,v_j)\in E} W_{ij} \log \frac{\exp(\widehat{W}_{ij}^{(2)})}{\sum_k \exp(\widehat{W}_{ik}^{(2)})}.$$

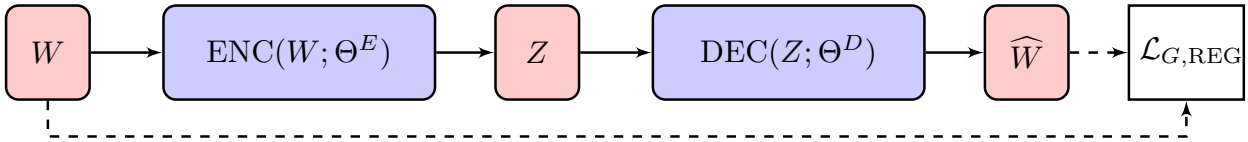


Figure 6: Auto-encoder methods. The graph structure (stored as the graph adjacency matrix) is encoded and reconstructed using encoder-decoder networks. Models are trained by optimizing the graph regularization loss computed on the reconstructed adjacency matrix.

Intuitively, LINE with second-order proximity decodes embeddings into context conditional distributions for each node $p_2(\cdot|v_i)$. Note that optimizing the second-order objective is computationally expensive as it requires a sum over the entire set of edges. LINE uses negative sampling to sample negative edges according to some noisy distribution over edges. Finally, as in GraRep, LINE combines first and second order embeddings with concatenation $Z = [Z^{(1)}|Z^{(1)}]$.

Hierarchical representation learning for networks (HARP) [33] Both node2vec and DeepWalk learn node embeddings by minimizing non-convex functions, which can lead to local minimas. HARP introduces a strategy that computes initial embeddings, leading to more stable training and convergence. More precisely, HARP hierarchically reduces the number of nodes in the graph via graph coarsening. Nodes are iteratively grouped into super nodes that form a graph with similar properties as the original graph, leading to multiple graphs with decreasing size $(G_1, \ldots, G_T)$. Node embeddings are then learned for each coarsened graph using existing methods such as LINE or DeepWalk, and at time-step t, embeddings learned for G_t are used as initialized embedding for the random walk algorithm on G_{t-1}. This process is repeated until each node is embedded in the original graph. The authors show that this hierarchical embedding strategy produces stable embeddings that capture macroscopic graph information.

Splitter [49] What if a node is not the correct 'base unit' of analysis for a graph? Unlike HARP, which coarsens a graph to preserve high-level topological features, Splitter is a graph embedding approach designed to better model nodes which have membership in multiple communities. It uses the Persona decomposition [50], to create a derived graph, G_P which may have multiple *persona* nodes for each original node in G (the edges of each original node are divided among its personas). G_P can then be embedded (with some constraints) using any of the embedding methods discussed so far. The resulting representations allow persona nodes to be separated in the embedding space, and the authors show benefits to this on link prediction tasks.

Matrix view of Skip-gram methods As noted by [86], Skip-gram methods can be viewed as implicit matrix factorization, and the methods discussed here are related to those of Matrix Factorization (Section 4.1.3). This relationship is discussed in depth by [113], who propose a general matrix factorization framework, NetMF, which uses the same underlying graph proximity information as DeepWalk, LINE, and node2vec. Casting the node embedding problem as matrix factorization can offer benefits like easier algorithmic analysis, and can also allow for efficient sparse matrix operations [114].

4.2 Auto-encoders

Shallow embedding methods hardly capture non-linear complex structures that might arise in graphs. Graph auto-encoders were originally introduced to overcome this issue by using deep neural network encoder and decoder functions, due to their ability model non-linearities. Instead of exploiting the graph structure through the graph regularization term, auto-encoders directly incorporate the graph adjacency matrix in the encoder function. Auto-encoders generally have an encoding and decoding network which are multiple layers of non-linear layers. For graph auto-encoders, the encoder function has the form:

$$Z = \text{ENC}(W;\Theta^E).$$

That is, the encoder is a function of the adjacency matrix W only. These models are trained by minimizing a reconstruction error objective and we review examples of such objectives next.

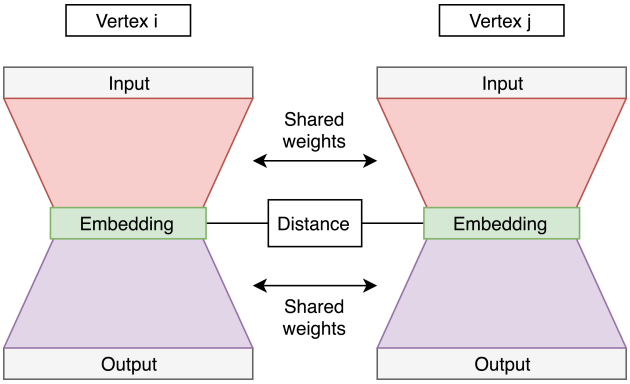


Figure 7: Illustration of the SDNE model. The embedding layer (denoted Z) is shown in green. Reprinted with permission from [56].

Structural Deep Network Embedding (SDNE) [138] learns auto-encoders that preserve first and second-order node proximity (Section 2.1). The SDNE encoder takes as input a node vector: a row of the adjacency matrix as they explicitly set $s(W) = W$, and produces node embeddings Z. The SDNE decoder return a reconstruction $\widehat{W}$, which is trained to recover the original graph adjacency matrix (Fig. 7). SDNE preserves second order node proximity by minimizing the graph regularization loss:

$$||(s(W) - \widehat{W}) \cdot B||_F^2 + \alpha_{\text{SDNE}} \sum_{ij} s(W)_{ij} ||Z_i - Z_j||_2^2,$$

where B is the indicator matrix for $s(W)$ with $B = 1[s(W) > 0]$. Note that the second term is the regularization loss used by distance-based shallow embedding methods. The first term is similar to the matrix factorization regularization objective, except that $\widehat{W}$ is not computed using outer products. Instead, SDNE computes a unique embedding for each node in the graph using a decoder network.

Deep neural Networks for learning Graph Representations (DNGR) [27] Similar to SDNE, DNGR uses deep auto-encoders to encode and decode a node similarity matrix, $s(W)$. The similarity matrix is computed using a probabilistic method called random surfing, that returns a probabilistic similarity matrix through graph exploration with random walks. Therefore, DNGR captures higher-order dependencies in the graph. The similarity matrix $s(W)$ is then encoded and decoded with stacked denoising auto-encoders [137], which allows to reduce the noise in $s(W)$. DNGR is optimized by minimizing the reconstruction error:

$$\mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta) = ||s(W) - \widehat{W}||_F^2.$$

4.3 Graph neural networks

In graph neural networks, both the graph structure and node features are used in the encoder function to learn structural representations of nodes:

$$Z = \text{ENC}(\text{X}, \text{W}; \Theta^{\text{E}}).$$

We first review unsupervised graph neural networks, and will cover supervised graph neural networks in more details in Section 5.

Variational Graph Auto-Encoders (VGAE) [76] use graph convolutions [75] to learn node embeddings $Z = \text{GCN}(W, X; \Theta^E)$ (see Section 5.3.1 for more details about graph convolutions). The decoder is an outer product: $\text{DEC}(Z; \Theta^D) = ZZ^\top$. The graph regularization term is the sigmoid cross entropy between the true adjacency and

the predicted edge similarity scores:

$$\mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta) = -\Big(\sum_{ij} (1 - W_{ij}) \log(1 - \sigma(\widehat{W}_{ij})) + W_{ij} \log \sigma(\widehat{W}_{ij}) \Big).$$

Computing the regularization term over all possible nodes pairs is computationally challenging in practice, and the Graph Auto Encoders (GAE) model uses negative sampling to overcome this challenge.

Note that GAE is a deterministic model but the authors also introduce variational graph auto-encoders (VGAE), where they use variational auto-encoders to encode and decode the graph structure. In VGAE, the embedding Z is modelled as a latent variable with a standard multivariate normal prior $p(Z) = \mathcal{N}(Z|0, I)$ and the amortized inference network $q_\Phi(Z|W, X)$ is also a graph convolution network. VGAE is optimized by minimizing the corresponding negative evidence lower bound:

$$\begin{aligned} \text{NELBO}(W, X; \Theta) &= -\mathbb{E}_{q_\Phi(Z|W,X)}[\log p(W|Z)] + \text{KL}(q_\Phi(Z|W, X)||p(Z)) \\ &= \mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta) + \text{KL}(q_\Phi(Z|W, X)||p(Z)). \end{aligned}$$

Iterative generative modelling of graphs (Graphite) [62] extends GAE and VGAE by introducing a more complex decoder, which iterates between pairwise decoding functions and graph convolutions. Formally, the graphite decoder repeats the following iteration:

$$\begin{aligned} \widehat{W}^{(k)} &= \frac{Z^{(k)} {Z^{(k)}}^\top}{||Z^{(k)}||_2^2} + \frac{11^\top}{|V|} \\ Z^{(k+1)} &= \text{GCN}(\widehat{W}^{(k)}, Z^{(k)}) \end{aligned}$$

where $Z^{(0)}$ are initialized using the output of the encoder network. By using this parametric iterative decoding process, Graphite learns more expressive decoders than other methods based on non-parametric pairwise decoding. Finally, similar to GAE, Graphite can be deterministic or variational.

Deep Graph Infomax (DGI) [135] is an unsupervised graph-level embedding method. Given one or more *real* (positive) graphs, each with its adjacency matrix $W \in \mathbb{R}^{|V|\times|V|}$ and node features $X \in \mathbb{R}^{|V|\times d_0}$, this method creates *fake* (negative) adjacency matrices $W^- \in \mathbb{R}^{|V^-|\times|V^-|}$ and their features $X^- \in \mathbb{R}^{|V^-|\times d_0}$. It trains (i) an encoder that processes real and fake samples, respectively giving $Z = \text{ENC}(X, W; \Theta^E) \in \mathbb{R}^{|V|\times d}$ and $Z^- = \text{ENC}(X^-, W^-; \Theta^E) \in \mathbb{R}^{|V|\times d}$, (ii) a (readout) graph pooling function $\mathcal{R} : \mathbb{R}^{|V|\times d} \to \mathbb{R}^d$, and (iii) a descriminator function $\mathcal{D} : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ which is trained to output $\mathcal{D}(Z_i, \mathcal{R}(Z)) \approx 1$ and $\mathcal{D}(Z_j^-, \mathcal{R}(Z^-)) \approx 0$, respectively, for nodes corresponding to given graph $i \in V$ and fake graph $j \in V^-$. Specifically, DGI optimizes:

$$\min_\Theta - \mathop{\mathbb{E}}_{X,W} \sum_{i=1}^{|V|} \log \mathcal{D}(Z_i, \mathcal{R}(Z)) - \mathop{\mathbb{E}}_{X^-,W^-} \sum_{j=1}^{|V^-|} \log \left(1 - \mathcal{D}(Z_j^-, \mathcal{R}(Z^-))\right), \tag{6}$$

where Θ contains Θ^E and the parameters of $\mathcal{R}, \mathcal{D}$. In the first expectation, DGI samples from the real (positive) graphs. If only one graph is given, it could sample some subgraphs from it (e.g. connected components). The second expectation samples fake (negative) graphs. In DGI, fake samples exhibit the real adjacency $W^- := W$ but fake features X^- are a row-wise random permutation of real X, though other negative sampling strategies are plausible. The ENC used in DGI is a graph convolutional network, though any GNN can be used. The readout $\mathcal{R}$ summarizes an entire (variable-size) graph to a single (fixed-dimension) vector. Veličković et al. [135] use $\mathcal{R}$ as a row-wise mean, though other graph pooling might be used e.g. ones aware of the adjacency, $\mathcal{R} : \mathbb{R}^{|V|\times d} \times \mathbb{R}^{|V|\times|V|} \to \mathbb{R}^d$.

The optimization (Eq. (6)) is shown by [135] to maximize a lower-bound on the Mutual Information (MI) between the outputs of the encoder and the graph pooling function. In other words, it maximizes the MI between individual node representations and the graph representation.

Graphical Mutual Information [GMI, 108] presents another MI alternative: rather than maximizing MI of node information and an entire graph, GMI maximizes the MI between the representation of a node and its neighbors.

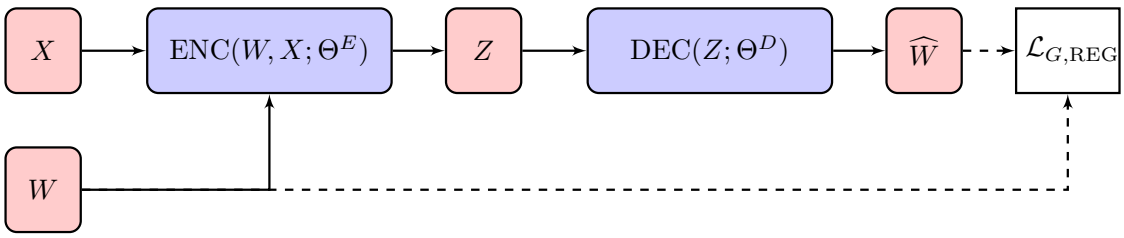


Figure 8: Unsupervised graph neural networks. Graph structure and input features are mapped to low-dimensional embeddings using a graph neural network encoder. Embeddings are then decoded to compute a graph regularization loss (unsupervised).

5 Supervised Graph Embedding

A common approach for supervised network embedding is to use an unsupervised network embedding method, like the ones described in Section 4 to first map nodes to an embedding vector space, and then use the learned embeddings as input for another neural network. However, an important limitation with this two-step approach is that the unsupervised node embeddings might not preserve important properties of graphs (e.g. node labels or attributes), that could have been useful for a downstream supervised task.

Recently, methods combining these two steps, namely learning embeddings and predicting node or graph labels, have been proposed. We describe these methods next.

5.1 Shallow embedding methods

Similar to unsupervised shallow embedding methods, supervised shallow embedding methods use embedding look-ups to map nodes to embeddings. However, while the goal in unsupervised shallow embeddings is to learn a good graph representation, supervised shallow embedding methods aim at doing well on some downstream prediction task such as node or graph classification.

Label propagation (LP) [154] is a very popular algorithm for graph-based semi-supervised node classification. It directly learns embeddings in the label space, i.e. the supervised decoder function in LP is simply the identity function:

$$\hat{y}^N = \text{DEC}(Z; \Theta^C) = Z.$$

In particular, LP uses the graph structure to smooth the label distribution over the graph by adding a regularization term to the loss function, where the underlying assumption is that neighbor nodes should have similar labels (i.e. there exist some label consistency between connected nodes). The regularization in LP is computed with Laplacian eigenmaps:

$$\mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta) = \sum_{ij} W_{ij} \widehat{W}_{ij} \tag{7}$$

$$\text{where } \widehat{W}_{ij} = ||\hat{y}_i^N - \hat{y}_j^N||_2^2. \tag{8}$$

LP minimizes this energy function over the space of functions that take fixed values on labelled nodes (i.e. $\hat{y}_i^N = y_i^N \ \forall i | v_i \in V_L$) using an iterative algorithm that updates a unlabelled node's label distribution via the weighted average of its neighbors' labels.

There exists variants of this algorithm such as Label Spreading (LS) [152], which minimizes the energy function:

$$\mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta) = \sum_{ij} W_{ij} \left|\left| \frac{\hat{y}_i^N}{\sqrt{D_i}} - \frac{\hat{y}_j^N}{\sqrt{D_j}} \right|\right|_2^2, \tag{9}$$

where $D_i = \sum_j W_{ij}$ is the degree of node v_i. The supervised loss in label spreading is simply the sum of distances between predicted labels and ground truth labels (one-hot vectors):

$$\mathcal{L}_{\text{SUP}}^N(y^N, \hat{y}^N; \Theta) = \sum_{i | v_i \in V_L} ||y_i^N - \hat{y}_i^N||_2^2.$$

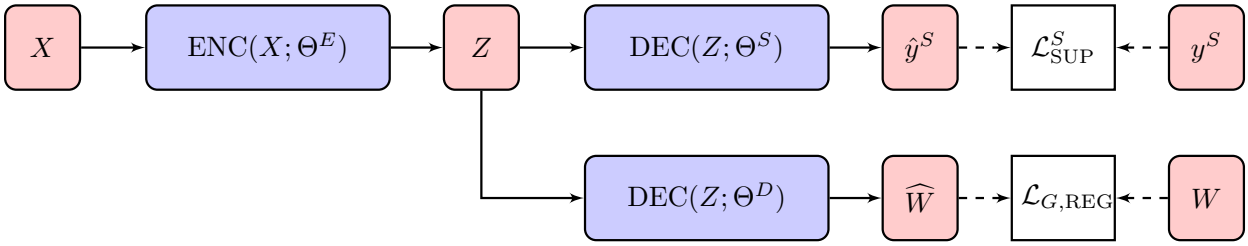


Figure 9: Supervised graph regularization methods. The graph structure is not used in the encoder nor the decoder networks. It instead acts as a regularizer in the loss function.

Note that the supervised loss is computed over labelled nodes only, while the regularization term is computed over all nodes in the graph. These methods are expected to work well with *consistent* graphs, that is graphs where node proximity in the graph is positively correlated with label similarity.

5.2 Graph regularization methods

Supervised graph regularization methods also aim at learning to predict graph properties such as node labels. Similar to shallow embeddings, these methods compute a graph regularization loss defined over the graph structure, and a supervised loss for the downstream task (Fig. 9). However, the main difference with shallow embeddings lies in the encoder network: rather than using embedding look-ups, graph regularization methods learn embeddings as parametric function defined over node features, which might capture valuable information for downstream applications. That is, encoder functions in these methods can be written as:

$$Z = \text{ENC}(\text{X};\Theta^{\text{E}}).$$

We review two types of semi-supervised [31] graph regularization approaches: Laplacian-based regularization methods and methods that use random walks to regularize embeddings.

5.2.1 Laplacian

Manifold Regularization (ManiReg) [16] builds on the LP model and uses Laplacian Eigenmaps to smoothen the label distribution via the regularization loss in Eq. (7). However, instead of using shallow embeddings to predict labels, ManiReg uses support vector machines to predict labels from node features. The supervised loss in ManiReg is computed as:

$$\mathcal{L}^N_{\text{SUP}}(y^N,\hat{y}^N;\Theta) = \sum_{i|v_i\in V_L}\sum_{1\le k\le C} H(y^N_{ik}\hat{y}^N_{ik}), \tag{10}$$

where $H(x) = \max(0, 1-x)$ is the hinge loss, C is the number of classes, and $\hat{y}^N_i = f(X_i;\Theta^E)$ are computed using Reproducing Kernel Hilbert Space (RKHS) functions that act on input features.

Semi-supervised Embeddings (SemiEmb) [140] further extend ManiReg and instead of using simple SVMs, this method uses feed-forward neural networks (FF-NN) to learn embeddings $Z = \text{ENC}(X;\Theta^E)$ and distance-based graph decoders:

$$\begin{aligned}\widehat{W}_{ij} &= \text{DEC}(Z;\Theta^D)_{ij}\\ &= ||Z_i - Z_j||^2\end{aligned}$$

where $||\cdot||$ can be the L2 or L1 norm. SemiEmb regularizes intermediate or auxiliary layers in the network using the same regularizer as the LP loss in Eq. (7). SemiEmb uses FF-NN to predict labels from intermediate embeddings, which are then compared to ground truth labels via the Hinge loss in Eq. (10).

Note that SemiEmb leverages multi-layer neural networks and regularizes intermediate hidden representations, while LP does not learn intermediate representations, and ManiReg only regularizes the last layer.

Model	$Z = \text{ENC}(X; \Theta^E)$	$\widehat{y}^N = \text{DEC}(Z; \Theta^N)$	$\widehat{W}_{ij} = \text{DEC}(Z; \Theta^D)_{ij}$	$\mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta)$	$\mathcal{L}^N_{\text{SUP}}(y^N, \widehat{y}^N; \Theta)$
LP	Shallow	$\widehat{y}^N = Z$	$\|\|Z_i - Z_j\|\|_2^2$	$\sum_{ij} W_{ij}\widehat{W}_{ij}$	$\widehat{y}^N_i = y^N_i$ is fixed $\forall v_i \in V_L$
LS	Shallow	$\widehat{y}^N = D^{-1/2}Z$	$\|\|Z_i - Z_j\|\|_2^2$	$\sum_{ij} W_{ij}\widehat{W}_{ij}$	$\sum_{i\|v_i\in V_L} \|\|y^N_i - \widehat{y}^N_i\|\|_2^2$
ManiReg	RKHS	$\widehat{y}^N = Z$	$\|\|Z_i - Z_j\|\|_2^2$	$\sum_{ij} W_{ij}\widehat{W}_{ij}$	$\sum_{i\|v_i\in V_L} \sum_{1\le k\le C} H(y^N_{ik}\widehat{y}^N_{ik})$
SemiEmb	FF-NN	FF-NN	$\|\|Z_i - Z_j\|\|^2$	$\sum_{ij} W_{ij}\widehat{W}_{ij}$	$\sum_{i\|v_i\in V_L} \sum_{1\le k\le C} H(y^N_{ik}\widehat{y}^N_{ik})$
NGM	CNN, LSTM, ...	CNN, LSTM, ...	$\|\|Z_i - Z_j\|\|^2$	$\sum_{ij} W_{ij}\widehat{W}_{ij}$	$-\frac{1}{\|V_L\|}\sum_{i\|v_i\in V_L} \sum_{1\le k\le C} y^N_{ik}\log \widehat{y}^N_{ik}$
Planetoid	FF-NN	FF-NN	$Z_i^\top Z_j$	$-\mathbb{E}_{(i,j,\gamma)} \log \sigma\left(\gamma\widehat{W}_{ij}\right)$	$-\frac{1}{\|V_L\|}\sum_{i\|v_i\in V_L} \sum_{1\le k\le C} y^N_{ik}\log \widehat{y}^N_{ik}$

Table 3: An overview of supervised shallow and graph regularization methods, where the graph structure is leveraged through the graph regularization term $\mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta)$.

Neural Graph Machines (NGM) [24] More recently, NGM generalize the regularization objective in Eq. (7) to more complex neural architectures than feed-forward neural networks (FF-NN), such as Long short-term memory (LSTM) networks [68] or CNNs [84]. in contrast with previous methods, NGM use the cross entropy loss for classification.

5.2.2 Skip-gram

The Laplacian-based regularization methods covered so far only capture first order proximities in the graphs. Skip-gram graph regularization methods further extend these methods to incorporate random walks, which are effective at capturing higher-order proximities.

Planetoid [143] Unsupervised skip-gram methods like node2vec and DeepWalk learn embeddings in a multi-step pipeline where random walks are first generated from the graph and then used to learn embeddings. These embeddings are not learned for a downstream classification task which might be suboptimal. Planetoid extends random walk methods to leverage node label information during the embedding algorithm.

Planetoid first maps nodes to embeddings $Z = [Z^c || Z^F] = \text{ENC}(X; \Theta^E)$ with neural networks, where Z^c are node embeddings that capture structural information while Z^F capture node feature information. The authors propose two variants, a transductive variant that directly learns embedding Z^c (as an embedding lookup), and an inductive variant where Z^c are computed with parametric mappings that act on input features X. Embeddings are then learned by minimizing a supervised loss and a graph regularization loss, where the regularization loss measures the ability to predict context using nodes embeddings, while the supervised loss measures the ability to predict the correct label. More specifically, the regularization loss in Planetoid is given by:

$$\mathcal{L}_{G,\text{REG}}(W, \widehat{W}; \Theta) = -\mathbb{E}_{(i,j,\gamma)} \log \sigma\left(\gamma\widehat{W}_{ij}\right),$$

with $\widehat{W}_{ij} = Z_i^\top Z_j$ and $\gamma \in \{-1, 1\}$ with $\gamma = 1$ if $(v_i, v_j) \in E$ is a positive pair and $\gamma = -1$ if (v_i, v_j) is a negative pair. The distribution under the expectation is directly defined through a sampling process[3]. The supervised loss in Planetoid is the negative log-likelihood of predicting the correct labels:

$$\mathcal{L}^N_{\text{SUP}}(y^N, \widehat{y}^N; \Theta) = -\frac{1}{|V_L|} \sum_{i|v_i\in V_L} \sum_{1\le k\le C} y^N_{ik}\log \widehat{y}^N_{ik}, \tag{11}$$

where i is a node's index while k indicates label classes, and $\widehat{y}^N_i$ are computed using a neural network followed by a softmax activation, mapping Z_i to predicted labels.

5.3 Graph convolution framework

We now focus on (semi-)supervised neighborhood aggregation methods, where the encoder uses input features and the graph structure to compute embeddings:

$$Z = \text{ENC}(X, W; \Theta^E).$$

[3]There are two kinds of sampling strategies to sample positive pairs of nodes i, j: (i.) samples drawn by conducting random walks, similar to DeepWalk and (ii.) samples drawn from the same class i.e. $y_i = y_j$. These samples are positive i.e. with $\gamma = 1$. The negative samples simply replace one of the nodes with another randomly-sampled (negative) node yielding $\gamma = -1$. The ratio of these kinds of samples are determined by hyperparameters.

We first review the graph neural network model—which was the first attempt to use deep learning techniques on graph-structured data—and other related frameworks such as message passing networks [55]. We then introduce a new Graph Convolution Framework (GCF), which is designed specifically for convolution-based graph neural networks. While GCF and other frameworks overlap on some methods, GCF emphasizes the geometric aspects of convolution and propagation, allowing to easily understand similarities and differences between existing convolution-based approaches.

5.3.1 The Graph Neural Network model and related frameworks

The Graph Neural Network model (GNN) [58, 120] The first formulation of deep learning methods for graph-structured data dates back to the graph neural network GNN) model of Gori et al. [58]. This formulation views the supervised graph embedding problem as an information diffusion mechanism, where nodes send information to their neighbors until some stable equilibrium state is reached. More concretely, given randomly initialized node embeddings Z^0, the following recursion is applied until convergence:

$$Z^{t+1} = \mathrm{ENC}(X, W, Z^t; \Theta^E), \tag{12}$$

where parameters Θ^E are reused at every iteration. After convergence ($t = T$), the node embeddings Z^T are used to predict the final output such as node or graph labels:

$$\hat{y}^S = \mathrm{DEC}(X, Z^T; \Theta^S).$$

This process is repeated several times and the GNN parameters Θ^E and Θ^D are learned with backpropagation via the Almeda-Pineda algorithm [9, 111]. Note that by Banach's fixed point theorem, the iteration in Eq. (12) is guaranteed to converge to a unique solution when the iteration mapping is a contraction mapping. In particular, Scarselli et al. [120] explore maps that can be expressed using message passing networks:

$$Z_i^{t+1} = \sum_{j|(v_i,v_j)\in E} f(X_i, X_j, Z_j^t; \Theta^E),$$

where $f(\cdot)$ is a multi-layer perception (MLP) constrained to be a contraction mapping. On the other hand, the decoder function in GNNs does not need to fulfill any constraint and can be any MLP.

Gated Graph Neural Networks (GGNN) [87] Gated Graph Sequence Neural Networks (GGSNN) or their simpler version GGNN are similar to GNNs but remove the contraction mapping requirement. In GGSNNs, the recursive algorithm in Eq. (12) is relaxed and approximated by applying mapping functions for a fixed number of steps, where each mapping function is a gated recurrent unit [39] with parameters shared for every iteration. The GGSNN model is particularly useful for machine learning tasks with sequential structure (such as temporal graphs) as it outputs predictions at every step.

Message Passing Neural Networks (MPNN) Gilmer et al. [55] In the same vein, MPNN provide a framework for graph neural networks, encapsulating many recent graph neural networks. In contrast with the GNN model which runs for an indefinite number of iterations, MPNN provide an abstraction for modern graph neural networks, which consist of multi-layer neural networks with a *fixed* number of layers. At every layer ℓ, message functions $f^\ell(.)$ compute messages using neighbors' hidden representations, which are then passed to aggregation functions $h^\ell(.)$:

$$m_i^{\ell+1} = \sum_{j|(v_i,v_j)\in E} f^\ell(H_i^\ell, H_j^\ell) \tag{13}$$

$$H_i^{\ell+1} = h^\ell(H_i^\ell, m_i^{\ell+1}). \tag{14}$$

After ℓ layers of message passing, nodes' hidden representations encode structural information within ℓ-hop neighborhoods. Gilmer et al. [55] explore additional variations of message functions within the MPNN framework, and achieve state-of-the-art results for prediction tasks defined on molecular graphs.

GraphNet [11] This framework further extends the MPNN framework to learn representations for edges, nodes and the entire graph using message passing functions. This framework is more general than the MPNN framework as it incorporates edge and graph representations.

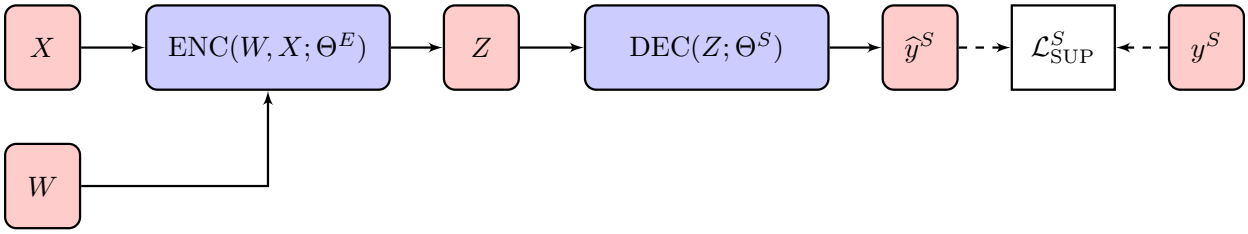


Figure 10: Supervised graph neural networks (GNNs). Rather than leveraging the graph structure to act as a regularizer, GNNs leverage the graph structure in the encoder to propagate information across neighbouring nodes and learn structural representations. Labels are then decoded and compared to ground truth labels (e.g. via the cross-entropy loss).

5.3.2 Graph Convolution Framework

We now introduce our Graph Convolution Framework (GCF); and as we shall see, many recent graph neural networks can be described using this framework. Different from the MPNN and GraphNet frameworks, our framework focuses on convolution-based methods, and draws direct connections between convolutions on grids and graph convolutions. While GCF does not include sophisticated message passing networks (e.g. messages computed with edge features), it emphasizes geometric properties of convolution operators, and provides a simple way to understand similarities and differences between state-of-the-art graph convolution methods.

GCF In GCF, node embeddings are initialized using input features $H^0 = X \in \mathbb{R}^{|V| \times d_0}$, and then updated with multiple layers of graph convolutions. Graph convolution layers provide a generalization of standard convolutions on grids to graph-structured data and are composed of four main components:

- **Patch functions**, which define the shape of convolutional filters (specifies which nodes interact with each other at every step of convolution), that is matrices of size $|V| \times |V|$:

  $$(f_1(W, H^\ell), \ldots, f_K(W, H^\ell)),$$

  where H^ℓ are node features at layer ℓ and K is the total number of patches. Note that the number of patches K might be defined in the spectral domain (e.g. rank of a matrix) or in the spatial domain (e.g. different neighborhood sizes). In standard CNNs (which are defined in the spatial pixel domain), these patches usually have rectangular shapes, where nodes (pixels in images) communicate with their top, left, bottom, and right neighbors. However, since graphs do not have a grid-like structure, the shape of convolutional filters does not follow regular patterns and is instead defined by the graph structure itself. While most methods use non-parametric patches at every layer, some methods such as attention-based methods (Section 5.5.2) learn patches using parametric functions.
- **Convolution filters' weights** at every layer, which are $d_\ell \times d_{\ell+1}$ matrices, representing the filter weights for each patch:

  $$(\Theta^\ell_1, \ldots, \Theta^\ell_K).$$

  Each column can be interpreted as a single convolution filter's weight, and we stack $d_{\ell+1}$ filters filters to compute features in the next layer. Similarly, d_ℓ and $d_{\ell+1}$ are analogous to the number of channels in CNNs for layer ℓ and $\ell + 1$ respectively. At every layer in the GCF, hidden representations H^ℓ are convolved with every patch using the convolution filter weights:

  $$m^{\ell+1}_k = f_k(W, H^\ell) H^\ell \Theta^\ell_k \quad \text{for} \quad 1 \leq k \leq K.$$

- **Merging functions**, which combine outputs from multiple convolution steps into one representation:

  $$H^{\ell+1} = h(m^{\ell+1}_1, \ldots, m^{\ell+1}_K).$$

  For instance, $h(\cdot)$ can be averaging or concatenation along the feature dimension followed by some non-linearity. Alternatively, $h(\cdot)$ can also be a more complicated operation parameterized by a neural network.

Method	Model	$g_k(.)$	$h(m_1, \ldots, m_k)$
Spectrum-based: $\widetilde{L} = U\Lambda U^\top,\ f_k(W, H) = g_k(U)$	SCNN	$g_k(U) = u_k u_k^\top$	$\sigma(\sum_k m_k)$
Spectrum-free: $f_k(W, H) = g_k(W, D)$	ChebNet	$g_k(W, D) = T_k(\frac{2(I - D^{-1/2}WD^{-1/2})}{\lambda_{max}(I - D^{-1/2}WD^{-1/2})} - I)$	$\sigma(\sum_k m_k)$
	GCN	$g_1(W, D) = (D + I)^{-1/2}(W + I)(D + I)^{-1/2}$	$\sigma(m_1)$
Spatial: $f_k(W, H) = g_k(W, D)$	SAGE-mean	$g_1(W, D) = I, g_2(W, D) \sim \mathcal{U}_{\text{norm}}(D^{-1}W, q)$	$\sigma(m_1 + m_2)$
	GGNN	$g_1(W, D) = I, g_2(W, D) = W$	$\text{GRU}(m_1, m2)$
Attention: $f_k(W, H) = \alpha(W \cdot g_k(H))$	MoNet	$g_k(U^s) = \exp(-\frac{1}{2}(U^s - \mu_k)^\top \Sigma_k^{-1}(U^s - \mu_k))$	$\sigma(\sum_k m_k)$
	GAT	$g_k(H) = \text{LeakyReLU}(HB^\top b_0 \oplus b_1^\top BH^\top)$	$\sigma([m_1 \| \ldots \| m_k])$

Table 4: An overview of graph convolution methods described using GCF.

After L convolution layers, nodes' embeddings $Z = H^L$ can be used to decode node or graph labels. Next, we review state-of-the-art GNNs, including spectral and spatial graph convolution methods using the proposed GCF framework.

5.4 Spectral Graph Convolutions

Spectral methods apply convolutions in the the spectral domain of the graph Laplacian matrix. These methods broadly fall into two categories: *spectrum-based methods*, which explicitly compute the Laplacian's eigendecomposition, and *spectrum-free* methods, which are motivated by spectral graph theory but do not explicitly compute the Laplacian's eigenvectors. One disadvantage of spectrum-based methods is that they rely on the spectrum of the graph Laplacian and are therefore domain-dependent (i.e. cannot generalize to new graphs). Moreover, computing the Laplacian's spectral decomposition is computationally expensive. Spectrum-free methods overcome these limitations by providing approximations for spectral filters.

5.4.1 Spectrum-based methods

Spectrum-based graph convolutions were the first attempt to generalize convolutions to non-Euclidean graph domains. Given a signal $x \in \mathbb{R}^{|V|}$ defined on a Euclidean discrete domain (e.g. grid), applying any linear translation-equivariant operator (i.e. with a Toeplitz structure) Θ in the discrete domain is equivalent to elementwise multiplication in the Fourier domain:

$$\mathcal{F}(\Theta x) = \mathcal{F}x \cdot \mathcal{F}\theta. \tag{15}$$

In non-Euclidean domains, the notion of translation (shift) is not defined and it is not trivial to generalize spatial convolutions operators (Θ) to non-Euclidean domains. Note that Eq. (15) can be equivalently written as:

$$\Theta x = \mathcal{F}^{-1}(\mathcal{F}x \cdot \mathcal{F}\theta).$$

While the left hand side is the Euclidean spatial convolution which is not defined for general graphs, the right hand side is a convolution in the Fourier domain which is defined for non-Euclidean domains. In particular, if $\widetilde{L} = I - D^{-1/2}WD^{-1/2}$ is the normalized Laplacian of a non-Euclidean graph, it is a real symmetric positive definite matrix and admits an orthonormal eigendecomposition: $\widetilde{L} = U\Lambda U^\top$. If $x \in \mathbb{R}^{|V|}$ is a signal defined on nodes in the graph, the discrete graph Fourier transform and its inverse can be written as:

$$\mathcal{F}x = \hat{x} = U^\top x \ \text{ and } \ \mathcal{F}^{-1}\hat{x} = U\hat{x}.$$

Spectral graph convolutions build on this observation to generalize convolutions to graphs, by learning convolution filters in the spectral domain of the normalized Laplacian matrix:

$$\begin{aligned} x * \theta &= U(U^\top x \cdot U^\top \theta) \\ &= U\text{diag}(U^\top \theta)U^\top x \end{aligned}$$

Using GCF, patch functions in spectrum-based methods can be expressed in terms of eigenvectors of the graph normalized Laplacian:

$$f_k(W, H^\ell) = g_k(U)$$

for some function $g_k(.)$. Note that this dependence on the spectrum of the Laplacian makes spectrum-based methods domain-dependent (i.e. they can only be used in transductive settings).

Spectral Convolutional Neural Networks (SCNN) [23] learn convolution filters as multipliers on the eigenvalues of the normalized Laplacian. SCNN layers compute feature maps at layer $\ell + 1$ with:

$$H^{\ell+1}_{:,j} = \sigma\Big(\sum_{i=1}^{d_\ell} U_K F^\ell_{i,j} U_K^\top H^\ell_{:,i}\Big), 1 \leq j \leq d_{\ell+1} \text{ and } 1 \leq i \leq d_\ell \tag{16}$$

where $\sigma(\cdot)$ is a non-linear transformation, U_K is a $|V| \times K$ matrix containing the top K eigenvectors of $\widetilde{L}$ and $F^\ell_{i,j}$ are $K \times K$ trainable diagonal matrices representing filters' weights in the spectral domain. We note that this spectral convolution operation can equivalently be written as:

$$H^{\ell+1}_{:,j} = \sigma\Big(\sum_{k=1}^{K} u_k u_k^\top \sum_{i=1}^{d_\ell} F^\ell_{i,j,k} H^\ell_{:,i}\Big), \tag{17}$$

where $(u_k)_{k=1,\dots,K}$ are the top K eigenvectors of $\widetilde{L}$ and $F^\ell_{i,j,k}$ is the k^{th} diagonal element of $F^\ell_{i,j}$. We can also write Eq. (17) using matrix notation as:

$$H^{\ell+1} = \sigma\Big(\sum_{k=1}^{K} u_k u_k^\top H^\ell \Theta^\ell_k\Big),$$

where Θ^ℓ_k are trainable matrices of shape $d_\ell \times d_{\ell+1}$ containing the filter weights. Using notation from GCF, SCNNs use patch functions expressed in terms of eigenvectors of the graph Laplacian $g_k(U) = u_k u_k^\top$, and the merging function $h(.)$ is the sum operator followed by a non-linearity $\sigma(.)$.

Euclidean grids have a natural ordering of nodes (top, left, bottom, right) allowing the use of spatially localized convolution filters with fixed size, independent of the input size. In contrast, SCNN layers require $\mathcal{O}(d_\ell d_{\ell+1} K)$ parameters, which is not scalable if K is $\mathcal{O}(|V|)$. Bruna et al. [23], Henaff et al. [67] note that spatial localization in the graph domain is equivalent to smoothness in the spectral domain, and propose smooth spectral multipliers in order to reduce the number of parameters in the model and avoid overfitting. Instead of learning K free parameters for each filter F^ℓ_{ij}, the idea behind smooth spectral multipliers is to parameterize F^ℓ_{ij} with polynomial interpolators such as cubic splines and learn a fixed number of interpolation coefficients. This modeling assumption leads to a constant number of parameters, independent of the graph size $|V|$.

In practice, SCNNs can be used for node classification or graph classification with graph pooling. However, SCNNs have two major limitations: (1) computing the eigendecomposition of the graph Laplacian is computationally expensive and (2) this method is domain-dependent, as its filters are eigen-basis dependent and cannot be shared across graphs.

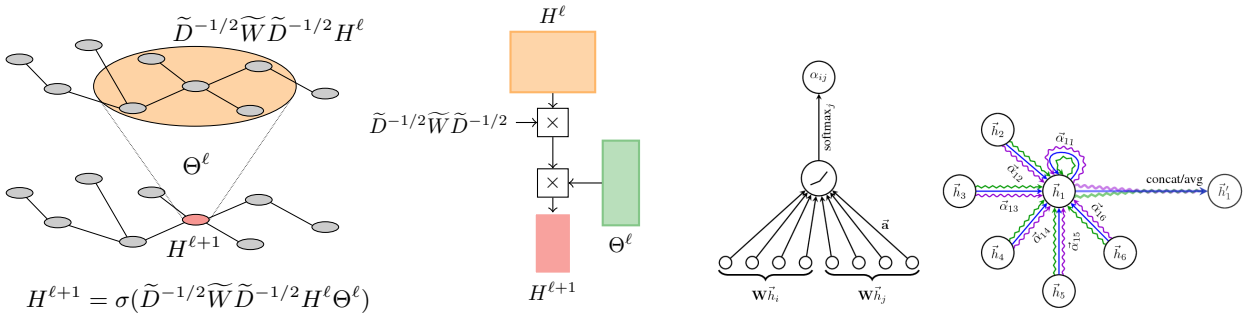
5.4.2 Spectrum-free methods

We now cover spectrum-free methods, which approximate convolutions in the spectral domain overcoming computational limitations of SCNNs by avoiding explicit computation of the Laplacian's eigendecomposition. SCNNs filters are neither localized nor parametric, in the sense that the parameters in F^ℓ_{ij} in Eq. (17) are all free. To overcome this issue, sprectrum-free methods use polynomial expansions to approximate spectral filters in Eq. (16) via:

$$F^\ell_{i,j} = P^\ell_{ij}(\Lambda)$$

where $P^\ell_{ij}(\cdot)$ is a finite degree polynomial. Therefore, the total number of free parameters per filter depends on the polynomial's degree, which is independent of the graph size. Assuming all eigenvectors are kept in Eq. (16), it can be rewritten as:

$$H^{\ell+1}_{:,j} = \sigma\Big(\sum_{i=1}^{d_\ell} P^\ell_{ij}(\Lambda) H^\ell_{:,i}\Big).$$



(a) GCN aggregation. (b) GCN layer. (c) GAT aggregation.

Figure 11: An illustration of neighborhood aggregation methods. Reprinted with permission from [4, 134].

If we write $P_{ij}^{\ell}(\lambda) = \sum_{k=1}^{K} \theta_{i,j,k}^{\ell} \left(\lambda^k\right)$, this yields in matrix notation:

$$H^{\ell+1} = \sigma\Big(\sum_{k=1}^{K} \left(\widetilde{L}^k\right) H^{\ell}\Theta_k^{\ell}\Big),$$

where Θ_k^{ℓ} is the matrix containing the polynomials' coefficients. These filters are k-localized, in the sense that the receptive field of each filter is k, and only nodes at a distance less than k will interact in the convolution operation. Since the normalized Laplacian is expressed in terms of the graph adjacency and degree matrices, we can write patch functions in spectrum-free method using notation from GCF:

$$f_k(W, H^{\ell}) = g_k(W, D).$$

Chebyshev Networks (ChebNets) [44] use the Chebyshev expansion [66] to approximate spectral filters. Chebyshev polynomials form an orthonormal basis in $[-1, 1]$ and can be computed efficiently with the recurrence:

$$T_0(x) = 1, \;\; T_1(x) = x, \;\; \text{and} \;\; T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x) \;\; \text{for } k \geq 2. \tag{18}$$

In order to use Chebyshev polynomials, ChebNets rescale the normalized adjacency martrix $\widetilde{L}$ to ensure that its eigenvalues are in $[-1, 1]$. The convolution step in ChebNet can be written as:

$$H^{\ell+1} = \sigma\Big(\sum_{k=1}^{K} T_k\Big(\frac{2}{\lambda_{max}(\widetilde{L})}\widetilde{L} - I\Big) H^{\ell}\Theta_k^{\ell}\Big),$$

where $\lambda_{max}(\widetilde{L})$ is the largest eigenvalue of $\widetilde{L}$.

Graph Convolution Networks (GCN) [75] further simplify ChebNet by letting $K = 2$, adding a weight sharing constraint for the first and second convolutions $\Theta_1^{\ell} = -\Theta_2^{\ell} := \Theta^{\ell}$, and assuming $\lambda_{max}(\widetilde{L}) \simeq 2$. This yields:

$$H^{\ell+1} = \sigma((2I - \widetilde{L})H^{\ell}\Theta^{\ell}) \tag{19}$$

$$= \sigma((I + D^{-1/2}WD^{-1/2})H^{\ell}\Theta^{\ell}), \tag{20}$$

Furthermore, since $I + D^{-1/2}WD^{-1/2}$ has eigenvalues in $[0, 2]$, applying Eq. (20) multiple times might lead to numerical instabilities or exploding gradients. To overcome this issue, GCNs use a re-normalization trick, which maps the eigenvalues of $I + D^{-1/2}WD^{-1/2}$ to $[0, 1]$:

$$I + D^{-1/2}WD^{-1/2} \to (D + I)^{-1/2}(W + I)(D + I)^{-1/2}.$$

Using GCF notation, GCN patch functions can be written as:

$$g_1(W, D) = (D + I)^{-1/2}(W + I)(D + I)^{-1/2},$$

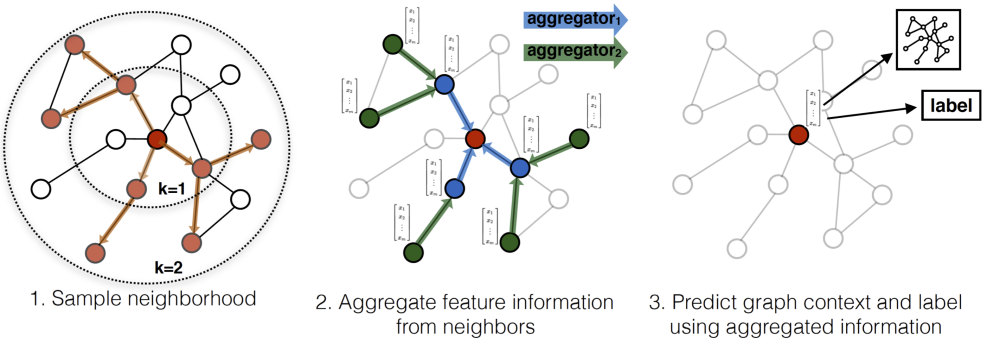


Figure 12: Illustration of the GraphSAGE model. Reprinted with permission from [64].

and the graph convolution layer (see 11 for an illustration) is:

$$H^{\ell+1} = \sigma(g_1(W, D)H^{\ell}\Theta^{\ell}). \tag{21}$$

This model has been applied to many problems including matrix completion [18], link prediction in knowledge graphs [121], and unsupervised graph embedding with variational inference [76].

Note that in contrast with spectrum-based methods covered in the previous section, both ChebyNet and GCN do not rely on computations of the Laplacian's eigenvectors. The convolution step is only defined over the local neighborhood of each node (as defined by the adjacency matrix W), and therefore we can view these methods as local message passing algorithms (see the Taxonomy in Fig. 3), even though these are motivated by spectral graph theory.

5.5 Spatial Graph Convolutions

Spectrum-based methods are limited by their domain dependency and cannot be applied in inductive settings. Furthermore, *spectrum-free* methods such as GCNs require storing the entire graph adjacency matrix, which can be computationally expensive for large graphs.

To overcome these limitations, *spatial* methods borrow ideas from standard CNNs, where convolutions are applied in the spatial domain as defined by the graph topology. For instance, in computer vision, convolutional filters are spatially localized by using fixed rectangular patches around each pixel. Additionally, since pixels in images have a natural ordering (top, left, bottom, right), it is possible to reuse filters' weights at every location, significantly reducing the total number of parameters. While such spatial convolutions cannot directly be applied in graph domains, spatial graph convolutions use ideas such as *neighborhood sampling* and *attention mechanisms* to overcome challenges posed by graphs' irregularities.

5.5.1 Sampling-based spatial methods

Inductive representation learning on large graphs (SAGE) [65] While GCNs can be used in inductive settings, they were originally introduced for semi-supervised transductive settings, and the learned filters might strongly rely on the Laplacian used for training. Furthermore, GCNs require storing the entire graph in memory which can be computationally expensive for large graphs.

To overcome these limitations, Hamilton et al. [64] propose SAGE, a general framework to learn inductive node embeddings while reducing the computational complexity of GCNs. Instead of averaging signals from all one-hop neighbors using multiplications with the Laplacian matrix, SAGE samples fixed neighborhoods (of size q) to remove the strong dependency on a fixed graph structure and generalize to new graphs. At every SAGE layer, nodes aggregate information from nodes sampled from their neighborhood, and the propagation rule can be written as:

$$H^{\ell+1}_{:,i} = \sigma(\Theta^{\ell}_1 H^{\ell}_{:,i} + \Theta^{\ell}_2 \mathrm{AGG}(\{H^{\ell}_{:,j} : j | v_j \in \mathrm{Sample}(\mathcal{N}(v_i), q)\})), \tag{22}$$

where $\mathrm{AGG}(\cdot)$ is an aggregation function, which can be any permutation invariant operator such as averaging (SAGE-mean) or max-pooling (SAGE-pool).

Note that SAGE can also be described using GCF. For simplicity, we describe SAGE-mean using GCF notation, and refer to [64] for details regarding other aggregation schemes. In GCF notation, SAGE-mean uses two patch learning functions with $g_1(W, D) = I$ being the identity, and $g_2(W, D) \sim \mathcal{U}_{\text{norm}}(D^{-1}W, q)$, where $\mathcal{U}_{\text{norm}}(\cdot, q)$ indicates uniformly sampling q nonzero entries per row, followed by row normalization. That is, the second patch propagates information using neighborhood sampling, and the SAGE-mean layer is:

$$H^{\ell+1} = \sigma(g_1(W, D)H^{\ell}\Theta_1^{\ell} + g_2(W, D)H^{\ell}\Theta_2^{\ell}).$$

5.5.2 Attention-based spatial methods

Attention mechanisms [133] have been successfully used in language models, and are particularly useful when operating on long sequence inputs, they allow models to identify relevant parts of the inputs. Similar ideas have been applied to graph convolution networks. Graph attention-based models learn to pay attention to important neighbors during the message passing step. This provides more flexibility in inductive settings, compared to methods that rely on fixed weights such as GCNs.

Broadly speaking, attention methods learn neighbors' importance using parametric functions whose inputs are node features at the previous layer. Using GCF, we can abstract patch functions in attention-based methods as functions of the form:

$$f_k(W, H^{\ell}) = \alpha(W \cdot g_k(H^{\ell})),$$

where $\cdot$ indicates element-wise multiplication and $\alpha(\cdot)$ is an activation function such as softmax or ReLU.

Graph Attention Networks (GAT) [134] is an attention-based version of GCNs, which incorporate self-attention mechanisms when computing patches. At every layer, GAT attends over the neighborhood of each node and learns to selectively pick nodes which lead to the best performance for some downstream task. The high-level intuition is similar to SAGE [64] and makes GAT suitable for inductive and transductive problems. However, instead of limiting the convolution step to fixed size-neighborhoods as in SAGE, GAT allows each node to attend over the entirety of its neighbors and uses attention to assign different weights to different nodes in a neighborhood. The attention parameters are trained through backpropagation, and the GAT self-attention mechanism is:

$$g_k(H^{\ell}) = \text{LeakyReLU}(H^{\ell}B^{\top}b_0 \oplus b_1^{\top}BH^{\ell^{\top}})$$

where $\oplus$ indicates summation of row and column vectors with broadcasting, and (b_0, b_1) and B are trainable attention weight vectors and weight matrix respectively. The edge scores are then row normalized with softmax. In practice, the authors propose to use multi-headed attention and combine the propagated signals with a concatenation of the average operator followed by some activation function. GAT can be implemented efficiently by computing the self-attention scores in parallel across edges, as well as computing the output representations in parallel across nodes.

Mixture Model Networks (MoNet) Monti et al. [98] provide a general framework that works particularly well when the node features lie in multiple domains such as 3D point clouds or meshes. MoNet can be interpreted as an attention method as it learns patches using parametric functions in a pre-defined spatial domain (e.g. spatial coordinates), and then applies convolution filters in the graph domain.

Note that MoNet is a generalization of previous spatial approaches such as Geodesic CNN (GCNN) [95] and Anisotropic CNN (ACNN) [20], which both introduced constructions for convolution layers on manifolds. However, both GCNN and ACNN use fixed patches that are defined on a specific coordinate system and therefore cannot generalize to graph-structured data. The MoNet framework is more general; any pseudo-coordinates such as local graph features (e.g. vertex degree) or manifold features (e.g. 3D spatial coordinates) can be used to compute the patches. More specifically, if U^s are pseudo-coordinates and H^{ℓ} are features from another domain, then using GCF, the MoNet layer can be expressed as:

$$H^{\ell+1} = \sigma\left(\sum_{k=1}^{K}(W \cdot g_k(U^s))H^{\ell}\Theta_k^{\ell}\right), \tag{23}$$

(a) GCN layers. (b) HGCN layers.

Figure 13: Euclidean (left) and hyperbolic (right) embeddings of a tree graph. Hyperbolic embeddings learn natural hierarchies in the embedding space (depth indicated by color). Reprinted with permission from [29].

where $\cdot$ is element-wise multiplication and $g_k(U^s)$ are the learned parametric patches, which are $|V| \times |V|$ matrices. In practice, MoNet uses Gaussian kernels to learn patches, such that:

$$g_k(U^s) = \exp\Big(-\frac{1}{2}(U^s - \mu_k)^\top \Sigma_k^{-1}(U^s - \mu_k)\Big),$$

where μ_k and Σ_k are learned parameters, and Monti et al. [98] restrict Σ_k to be a diagonal matrix.

5.6 Non-Euclidean Graph Convolutions

Hyperbolic shallow embeddings enable embeddings of hierarchical graphs with smaller distortion than Euclidean embeddings. However, one major downside of shallow embeddings is that they are inherently transductive and cannot generalize to new graphs. On the other hand, Graph Neural Networks, which leverage node features, have achieved state-of-the-art performance on inductive graph embedding tasks.

Recently, there has been interest in extending Graph Neural Networks to learn non-Euclidean embeddings and thus benefit from both the expressiveness of Graph Neural Networks and hyperbolic geometry. One major challenge in doing so is how to perform convolutions in a non-Euclidean space, where standard operations such as inner products and matrix multiplications are not defined.

Hyperbolic Graph Convolutional Neural Networks (HGCN) [29] and Hyperbolic Graph Neural Networks (HGNN) [91] apply graph convolutions in hyperbolic space by leveraging the Euclidean tangent space, which provides a first-order approximation of the hyperbolic manifold at a point. For every graph convolution step, node embeddings are mapped to the Euclidean tangent space at the origin, where convolutions are applied, and then mapped back to the hyperbolic space. These approaches yield significant improvements on graphs that exhibit hierarchical structure (Fig. 13).

6 Applications

Graph representation learning methods can be applied to a wide range of applications, which can be unsupervised or supervised. In unsupervised applications, the goal is to learn embeddings that preserve the graph structure and unsupervised supervised embedding methods (Section 4, upper branch of the Taxonomy in Fig. 3) can be applied. On the other hand, in supervised applications, node embeddings are optimized for some specific task, such as classifying nodes or graphs. In this setting, supervised embedding methods (Section 5, lower branch of the Taxonomy in Fig. 3) can be applied. We review common unsupervised and supervised graph applications next.

6.1 Unsupervised applications

6.1.1 Graph reconstruction

The most standard unsupervised graph application is graph reconstruction. In this setting, the goal is to learn mapping functions (which can be parametric or not) that map nodes to dense distributed representations which preserve graph properties such as node similarity. Graph reconstruction doesn't require any supervision and models can be trained by minimizing a reconstruction error, which is the error in recovering the original graph from learned embeddings. Several algorithms were designed specifically for this task, and we refer to Section 4 for some examples of reconstruction objectives. At a high level, graph reconstruction is similar to dimensionality reduction in the sense that the main goal is summarize some input data into a low-dimensional embedding. Instead of compressing high dimensional vectors into low-dimensional ones as standard dimensionality reduction methods (e.g. PCA) do, the goal of graph reconstruction models is to compress data defined on graphs into low-dimensional vectors.

6.1.2 Link prediction

Link prediction is the task of predicting links in a graph. In other words, the goal in link prediction tasks is to predict missing or unobserved links (e.g. links that may appear in the future for dynamic and temporal networks). Link prediction can also help identifying spurious link and remove them. It is a major application of graph learning models in industry, and common example of applications include predicting friendships in social networks or predicting user-product interactions in recommendation systems.

A common approach for training link prediction models is to mask some edges in the graph (positive and negative edges), train a model with the remaining edges and then test it on the masked set of edges. Note that link prediction is different from graph reconstruction. In link prediction, we aim at predicting links that are not observed in the original graph while in graph reconstruction, we only want to compute embeddings that preserve the graph structure through reconstruction error minimization.

Finally, while link prediction has similarities with supervised tasks in the sense that we have labels for edges (positive, negative, unobserved), we group it under the unsupervised class of applications since edge labels are usually not used during training, but only used to measure the predictive quality of embeddings. That is, models described in Section 4 can be applied to the link prediction problem.

6.1.3 Clustering

Clustering is particularly useful for discovering communities and has many real-world applications. For instance, clusters exist in biological networks (e.g. as groups of proteins with similar properties), or in social networks (e.g. as groups of people with similar interests).

Note that unsupervised methods introduced in this survey can be used to solve clustering problems: one can run a clustering algorithm (e.g. k-means) on embeddings that are output by an encoder. Further, clustering can be joined with the learning algorithm while learning a shallow [117] or Graph Convolution [36, 38] embedding model.

6.1.4 Visualization

There are many off-the-shelf tools for mapping graph nodes onto two-dimensional manifolds for the purpose of visualization. Visualizations allow network scientists to qualitatively understand graph properties, understand relationships between nodes or visualize node clusters. Among the popular tools are methods based on *Force-Directed Layouts*, with various web-app Javascript implementations.

Unsupervised graph embedding methods are also used for visualization purposes: by first training an encoder-decoder model (corresponding to a shallow embedding or graph convolution network), and then mapping every node representation onto a two-dimensional space using, t-distributed stochastic neighbor embeddings (t-SNE) [93] or PCA [72]. Such a process (embedding $\rightarrow$ dimensionality reduction) is commonly used to qualitatively evaluate the performance of graph learning algorithms. If nodes have attributes, one can use these attributes to color the nodes on 2D visualization plots. Good embedding algorithms embed nodes that have similar attributes nearby in the embedding space, as demonstrated in visualizations of various methods [3, 75, 110]. Finally, beyond mapping every node to a 2D coordinate, methods which map every graph to a representation [7] can similarly be projected into two dimensions to visualize and qualitatively analyze graph-level properties.

6.2 Supervised applications

6.2.1 Node classification

Node classification is an important supervised graph application, where the goal is to learn node representations that can accurately predict node labels. For instance, node labels could be scientific topics in citation networks, or gender and other attributes in social networks.

Since labelling large graphs can be time-consuming and expensive, semi-supervised node classification is a particularly common application. In semi-supervised settings, only a small fraction of nodes is labelled and the goal is to leverage links between nodes to predict attributes of unlabelled nodes. This setting is transductive since there is only one partially labelled fixed graph. It is also possible to do inductive node classification, which corresponds to the task of classifying nodes in multiple graphs.

Note that node features can significantly boost the performance on node classification tasks if these are descriptive for the target label. Indeed, recent methods such as GCN [75] or GraphSAGE [64] have achieved state-of-the-art performance on multiple node classification benchmarks due to their ability to combine structural information and semantics coming from features. On the other hand, other methods such as random walks on graphs fail to leverage feature information and therefore achieve lower performance on these tasks.

6.2.2 Graph classification

Graph classification is a supervised application where the goal is to predict graph labels. Graph classification problems are inductive and a common example is classifying chemical compounds (e.g. predicting toxicity).

Graph classification is a particularly challenging task because it requires some notion of pooling, in order to aggregate node-level information into graph-level information. As discussed earlier, generalizing this notion of pooling to arbitrary graphs is non trivial because of the lack of regularity in the graph structure making graph pooling an active research area. In addition to the supervised methods discussed above, a number of unsupervised methods for learning graph-level representations have been proposed [7, 131, 132].

7 Conclusion and Open Research Directions

In this survey, we introduced a unified framework to compare machine learning models for graph-structured data. We presented a generalized GRAPHEDM framework, previously applied to unsupervised network embedding, that encapsulates shallow graph embedding methods, graph auto-encoders, graph regularization methods and graph neural networks. We also introduced a graph convolution framework (GCF), which is used to describe and compare convolution-based graph neural networks, including spatial and spectral graph convolutions. Using this framework, we introduced a comprehensive taxonomy of GRL methods, encapsulating over thirty methods for graph embedding (both supervised and unsupervised).

We hope that this survey will help and encourage future research in GRL, to hopefully solve the challenges that these models are currently facing. In particular, practitioners can reference the taxonomy to better understand the available tools and applications, and easily identify the best method for a given problem. Additionally, researchers with new research questions can use the taxonomy to better classify their research questions, reference the existing work, identify the right baselines to compare to, and find the appropriate tools to answer their questions.

While GRL methods have achieved state-of-the-art performance on node classification or link prediction tasks, many challenges remain unsolved. Next, we discuss ongoing research directions and challenges that graph embedding models are facing.

Evaluation and benchmarks The methods covered in this survey are typically evaluated using standard node classification or link prediction benchmarks. For instance, citation networks are very often used as benchmarks to evaluate graph embedding methods. However, these small citation benchmarks have drawbacks since results might significantly vary based on datasets' splits, or training procedures (e.g. early stopping), as shown in recent work [123].

To better advance GRL methods, it is important to use robust and unified evaluation protocols, and evaluate these methods beyond small node classification and link prediction benchmarks. Recently, there has been progress in this direction, including new graph benchmarks with leaderboards [47, 69] and graph embedding libraries [52, 59, 139].

On a similar vein, Sinha et al. [125] recently proposed a set of tasks grounded in first-order logic, to evaluate the reasoning capabilities of GNNs.

Fairness in Graph Learning The emerging field of Fairness in Machine Learning seeks to ensure that models avoid correlation between 'sensitive' features and the model's predicted output [96]. These concerns can be especially relevant for graph learning problems, where we must also consider the correlation of the graph structure (the edges) in addition to the feature vectors of the nodes with the final output.

The most popular technique for adding fairness constraints to models relies on using adversarial learning to debias the model's predictions relative to the sensitive feature(s), and can be extended to GRL [21]. However, adversarial methods do not offer strong guarantees about the actual amount of bias removed. In addition, many debiasing methods may not be effective at the debiasing task in practice [57]. Recent work in the area aims to provide provable guarantees for debiasing GRL [105].

Application to large and realistic graphs Most learning methods on graphs are applied only on smaller datasets, with sizes of up to hundred of thousands of nodes. However, many real-world graphs are much larger, containing up to billions of nodes. Methods that scale for large graphs [85, 144] require a Distributed Systems setup with many machines, such as MapReduce [43]. Given a large graph that fits on a single hard disk (e.g. with one terabyte size) but does not fit on RAM, how can a researcher apply a learning method on such a large graphs, using just a personal computer? Contrast this with a computer vision task by considering a large image dataset [45, 81]. It is possible to train such models on personal computers, as long as the model can fit on RAM, regardless how large the dataset is. This problem may be particularly challenging for graph embedding models, especially those which have parameters that scale with the number of nodes in the graph.

We foresee engineering and mathematical challenges in learning methods for large graphs, while still being operable on a single machine. We hope that researchers can focus on this direction to expose such learning tools to non-expert practitioners, such as a Neurologist wishing to analyze the sub-graph of the human brain given its neurons and synapses, stored as nodes and edges.

Molecule generation Learning on graphs has a great potential for helping molecular scientists to reduce cost and time in the laboratory. Researchers proposed methods for predicting quantum properties of molecules [46, 55] and for generating molecules with some desired properties [42, 88, 90, 124, 146]. A review of recent methods can be found in [48]. Many of these methods are concerned with manufacturing materials with certain properties (e.g. conductance and malleability), and others are concerned drug design [51, 71, 115].

Combinatorial optimization Computationally hard problems arise in a broad range of areas including routing science, cryptography, decision making and planning. Broadly speaking, a problem is computationally hard when the algorithms that compute the optimal solution scale poorly with the problem size. There has been recent interest in leveraging machine learning techniques (e.g. reinforcement learning) to solve combinatorial optimization problems and we refer to [17] for a review of these methods.

Many hard problems (e.g. SAT, vertex cover...) can be expressed in terms of graphs and more recently, there has been interest in leveraging graph embeddings to approximate solutions of NP-hard problems [74, 103, 112, 122]. These methods tackle computationally hard problems from a data-driven perspective, where given multiple instances of a problem, the task is to predict whether a particular instance (e.g. node) belongs to the optimal solution. One motivation for these approaches is the relational inductive biases found in GNNs which enable them to better represent graphs compared to standard neural networks (e.g. permutation invariance). While these data-driven methods are still outperformed by existing solvers, promising results show that GNNs can generalize to larger problem instances [103, 112]. We refer to the recent survey on neural symbolic learning by Lamb et al. [82] for an extensive review of GNN-based methods for combinatorial optimization.

Non-Euclidean embeddings As we saw in Section 4.1.2 and Section 5.6, an important aspect of graph embeddings is the underlying space geometry. Graphs are discrete, high-dimensional, non-Euclidean structures, and there is no straightforward way to encode this information into low-dimensional Euclidean embeddings that preserve the graph topology [22]. Recently, there has been interest and progress into learning non-Euclidean embeddings such as hyperbolic [101] or mixed-product space [63] embeddings. These non-Euclidean embeddings provide a promise for more

expressive embeddings, compared to Euclidean embeddings. For instance, hyperbolic embeddings can represent hierarchical data with much smaller distortion than Euclidean embeddings [119] and have led to state-of-the-art results in many modern applications such as link prediction in knowledge graphs [10, 30] and linguistics tasks [83, 130].

Two common challenges arise with non-Euclidean embeddings: precision issues (e.g. near the boundary of the Poincaré ball) in hyperbolic space [118, 148] and challenging Riemannian optimization [13, 19]. Additionally, it is also unclear how to pick the right geometry for a given input graph. While there exists some discrete measures for the tree-likeliness of graphs (e.g. Gromov's four-point condition [73] and others [1, 5, 35]), an interesting open research direction is how to pick or learn the right geometry for a given discrete graph.

Theoretical guarantees There have been significant advances in the design of graph embedding models, which improved over the state-of-the-art in many applications. However, there is still limited understanding about theoretical guarantees and limitations of graph embedding models. Understanding the representational power of GNNs is a nascent area of research, and recent works adapt existing results from learning theory to the problem of GRL [37, 54, 92, 94, 99, 136, 142]. The development of theoretical frameworks is critical to pursue in order to understand the theoretical guarantees and limitations of graph embedding methods.

Acknowledgements

We thank Meissane Chami, Aram Galstyan, Megan Leszczynski, John Palowitch, Laurel Orr, and Nimit Sohoni for their helpful feedback and discussions. We also thank Carlo Vittorio Cannistraci, Thomas Kipf, Luis Lamb, Bruno Ribeiro and Petar Veličković for their helpful feedback and comments on the first version of this work. We gratefully acknowledge the support of DARPA under Nos. FA86501827865 (SDH) and FA86501827882 (ASED); NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and 1937301 (RTML); ONR under No. N000141712266 (Unifying Weak Supervision); the Moore Foundation, NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, the Okawa Foundation, American Family Insurance, Google Cloud, Swiss Re, the HAI-AWS Cloud Credits for Research program, TOTAL, and members of the Stanford DAWN project: Teradata, Facebook, Google, Ant Financial, NEC, VMWare, and Infosys. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of DARPA, NIH, ONR, or the U.S. Government.

References

[1] Muad Abu-Ata and Feodor F Dragan. Metric tree-like structures in real-world networks: an empirical study. *Networks*, 67 (1):49–68, 2016.

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

[3] Sami Abu-El-Haija, Bryan Perozzi, Rami Al-Rfou, and Alexander A Alemi. Watch your step: Learning node embeddings via graph attention. In *Advances in Neural Information Processing Systems*, pages 9180–9190, 2018.

[4] Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Nazanin Alipourfard, Kristina Lerman, Hrayr Harutyunyan, Greg Ver Steeg, and Aram Galstyan. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *International Conference on Machine Learning*, pages 21–29, 2019.

[5] Aaron B Adcock, Blair D Sullivan, and Michael W Mahoney. Tree-like structure in large social and information networks. In *2013 IEEE 13th International Conference on Data Mining*, pages 1–10. IEEE, 2013.

[6] Amr Ahmed, Nino Shervashidze, Shravan Narayanamurthy, Vanja Josifovski, and Alexander J Smola. Distributed large-scale natural graph factorization. In *Proceedings of the 22nd international conference on World Wide Web*, pages 37–48. ACM, 2013.

[7] Rami Al-Rfou, Dustin Zelle, and Bryan Perozzi. Ddgk: Learning graph representations for deep divergence graph kernels. *Proceedings of the 2019 World Wide Web Conference on World Wide Web*, 2019.

[8] Gregorio Alanis-Lobato, Pablo Mier, and Miguel A Andrade-Navarro. Efficient embedding of complex networks to hyperbolic space via their laplacian. *Scientific reports*, 6:30108, 2016.

[9] Luis B Almeida. A learning rule for asynchronous perceptrons with feedback in a combinatorial environment. In *Proceedings, 1st First International Conference on Neural Networks*, volume 2, pages 609–618. IEEE, 1987.

[10] Ivana Balazevic, Carl Allen, and Timothy Hospedales. Multi-relational poincaré graph embeddings. In *Advances in Neural Information Processing Systems*, pages 4463–4473, 2019.

[11] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems*, pages 4502–4510, 2016.

[12] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[13] Gary Becigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *International Conference on Learning Representations*, 2018.

[14] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591, 2002.

[15] Mikhail Belkin and Partha Niyogi. Semi-supervised learning on riemannian manifolds. *Machine learning*, 56(1-3):209–239, 2004.

[16] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

[17] Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: a methodological tour d'horizon. *arXiv preprint arXiv:1811.06128*, 2018.

[18] Rianne van den Berg, Thomas N Kipf, and Max Welling. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263*, 2017.

[19] Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9): 2217–2229, 2013.

[20] Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 3189–3197, 2016.

[21] Avishek Joey Bose and William Hamilton. Compositional fairness constraints for graph embeddings. *arXiv preprint arXiv:1905.10674*, 2019.

[22] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[23] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann Lecun. Spectral networks and locally connected networks on graphs international conference on learning representations (iclr2014). *CBLS, April*, 2014.

[24] Thang D Bui, Sujith Ravi, and Vivek Ramavajjala. Neural graph learning: Training neural networks using graphs. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 64–71, 2018.

[25] Hongyun Cai, Vincent W Zheng, and Kevin Chang. A comprehensive survey of graph embedding: problems, techniques and applications. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[26] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 891–900. ACM, 2015.

[27] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Deep neural networks for learning graph representations. In *AAAI*, pages 1145–1152, 2016.

[28] Benjamin Paul Chamberlain, James Clough, and Marc Peter Deisenroth. Neural embeddings of graphs in hyperbolic space. *arXiv preprint arXiv:1705.10359*, 2017.

[29] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 4869–4880, 2019.

[30] Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[31] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[32] Haochen Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. A tutorial on network embeddings. *arXiv preprint arXiv:1808.02590*, 2018.

[33] Haochen Chen, Bryan Perozzi, Yifan Hu, and Steven Skiena. Harp: Hierarchical representation learning for networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[34] Haochen Chen, Xiaofei Sun, Yingtao Tian, Bryan Perozzi, Muhao Chen, and Steven Skiena. Enhanced network embeddings via exploiting edge labels. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1579–1582, 2018.

[35] Wei Chen, Wenjie Fang, Guangda Hu, and Michael W Mahoney. On the hyperbolicity of small-world and treelike random graphs. *Internet Mathematics*, 9(4):434–491, 2013.

[36] Zhengdao Chen, Joan Bruna Estrach, and Lisha Li. Supervised community detection with line graph neural networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[37] Zhengdao Chen, Soledad Villar, Lei Chen, and Joan Bruna. On the equivalence between graph isomorphism testing and function approximation with gnns. In *Advances in Neural Information Processing Systems*, pages 15894–15902, 2019.

[38] Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2019. URL https://arxiv.org/pdf/1905.07953.pdf.

[39] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[40] Michael AA Cox and Trevor F Cox. Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer, 2008.

[41] Daniel Fernando Daza Cruz, Thomas Kipf, and Max Welling. A modular framework for unsupervised graph representation learning. 2019.

[42] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

[43] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, page 107–113, 2008. doi: 10.1145/1327452.1327492. URL https://doi.org/10.1145/1327452.1327492.

[44] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pages 3844–3852, 2016.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[46] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

[47] Vijay Prakash Dwivedi, Chaitanya K Joshi, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *arXiv preprint arXiv:2003.00982*, 2020.

[48] Daniel C Elton, Zois Boukouvalas, Mark D Fuge, and Peter W Chung. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, 4(4):828–849, 2019.

[49] Alessandro Epasto and Bryan Perozzi. Is a single embedding enough? learning node representations that capture multiple social contexts. In *The World Wide Web Conference*, WWW '19, page 394–404, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313660. URL https://doi.org/10.1145/3308558.3313660.

[50] Alessandro Epasto, Silvio Lattanzi, and Renato Paes Leme. Ego-splitting framework: From non-overlapping to overlapping clusters. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 145–154, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098054. URL https://doi.org/10.1145/3097983.3098054.

[51] Qingyuan Feng, Evgenia Dueva, Artem Cherkasov, and Martin Ester. Padme: A deep learning-based framework for drug-target interaction prediction. *arXiv preprint arXiv:1807.09741*, 2018.

[52] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

[53] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[54] Vikas K Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. *arXiv preprint arXiv:2002.06157*, 2020.

[55] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272. JMLR. org, 2017.

[56] Primož Godec. https://towardsdatascience.com/graph-embeddings-the-summary-cc6075aba007, 2018.

[57] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

[58] Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE, 2005.

[59] Palash Goyal and Emilio Ferrara. Gem: a python package for graph embedding methods. *Journal of Open Source Software*, 3(29):876, 2018.

[60] Palash Goyal and Emilio Ferrara. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94, 2018.

[61] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

[62] Aditya Grover, Aaron Zweig, and Stefano Ermon. Graphite: Iterative generative modeling of graphs. In *International Conference on Machine Learning*, pages 2434–2444, 2019.

[63] Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. *International Conference on Learning Representations*, 2018.

[64] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017.

[65] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

[66] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

[67] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

[68] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[69] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*, 2020.

[70] Di Huang, Zihao He, Yuzhong Huang, Kexuan Sun, Sami Abu-El-Haija, Bryan Perozzi, Kristina Lerman, Fred Morstatter, and Aram Galstyan. Graph embedding with personalized context distribution. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 655–661, 2020.

[71] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, 2018.

[72] Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

[73] Edmond Jonckheere, Poonsuk Lohsoonthorn, and Francis Bonahon. Scaled gromov hyperbolic graphs. *Journal of Graph Theory*, 2008.

[74] Elias Khalil, Hanjun Dai, Yuyu Zhang, Bistra Dilkina, and Le Song. Learning combinatorial optimization algorithms over graphs. In *Advances in Neural Information Processing Systems*, pages 6348–6358, 2017.

[75] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[76] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[77] Robert Kleinberg. Geographic routing using hyperbolic space. In *IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications*, pages 1902–1909. IEEE, 2007.

[78] Ioannis Konstas, Vassilios Stathopoulos, and Joemon M. Jose. On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 195–202, 2009.

[79] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

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

[81] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[82] Luis Lamb, Artur Garcez, Marco Gori, Marcelo Prates, Pedro Avelar, and Moshe Vardi. Graph neural networks meet neural-symbolic computing: A survey and perspective. *arXiv preprint arXiv:2003.00330*, 2020.

[83] Matthew Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3231–3241, 2019.

[84] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[85] Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. PyTorch-BigGraph: A Large-scale Graph Embedding System. In *Proceedings of the 2nd SysML Conference*, Palo Alto, CA, USA, 2019.

[86] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014.

[87] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

[88] Yujia Li, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter Battaglia. Learning deep generative models of graphs. *arXiv preprint arXiv:1803.03324*, 2018.

[89] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[90] Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander Gaunt. Constrained graph variational autoencoders for molecule design. In *Advances in Neural Information Processing Systems*, pages 7795–7804, 2018.

[91] Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In *Advances in Neural Information Processing Systems*, pages 8228–8239, 2019.

[92] Andreas Loukas. What graph neural networks cannot learn: depth vs width. *arXiv preprint arXiv:1907.03199*, 2019.

[93] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

[94] Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2018.

[95] Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 37–45, 2015.

[96] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[97] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[98] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5115–5124, 2017.

[99] Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4602–4609, 2019.

[100] Alessandro Muscoloni, Josephine Maria Thomas, Sara Ciucci, Ginestra Bianconi, and Carlo Vittorio Cannistraci. Machine learning meets complex networks via coalescent embedding in the hyperbolic space. *Nature communications*, 8(1):1–19, 2017.

[101] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, pages 6338–6347, 2017.

[102] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pages 3779–3788, 2018.

[103] Alex Nowak, Soledad Villar, Afonso S Bandeira, and Joan Bruna. Revised note on learning algorithms for quadratic assignment with graph neural networks. *arXiv preprint arXiv:1706.07450*, 2017.

[104] Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1105–1114. ACM, 2016.

[105] John Palowitch and Bryan Perozzi. Monet: Debiasing graph embeddings via the metadata-orthogonal training unit. *arXiv preprint arXiv:1909.11793*, 2019.

[106] Fragkiskos Papadopoulos, Maksim Kitsak, M Ángeles Serrano, Marián Boguná, and Dmitri Krioukov. Popularity versus similarity in growing networks. *Nature*, 489(7417):537–540, 2012.

[107] Fragkiskos Papadopoulos, Constantinos Psomas, and Dmitri Krioukov. Network mapping by replaying hyperbolic growth. *IEEE/ACM Transactions on Networking*, 23(1):198–211, 2014.

[108] Zhen Peng, Wenbing Huang, Minnan Luo, Qinghua Zheng, Yu Rong, Tingyang Xu, and Junzhou Huang. Graph Representation Learning via Graphical Mutual Information Maximization. In *Proceedings of The Web Conference*, 2020. doi: https://doi.org/10.1145/3366423.3380112.

[109] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[110] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[111] Fernando J Pineda. Generalization of back propagation to recurrent and higher order neural networks. In *Neural information processing systems*, pages 602–611, 1988.

[112] Marcelo Prates, Pedro HC Avelar, Henrique Lemos, Luis C Lamb, and Moshe Y Vardi. Learning to solve np-complete problems: A graph neural network for decision tsp. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4731–4738, 2019.

[113] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 459–467, 2018.

[114] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Chi Wang, Kuansan Wang, and Jie Tang. Netsmf: Large-scale network embedding as sparse matrix factorization. In *The World Wide Web Conference*, WWW '19, page 1509–1520, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313446. URL https://doi.org/10.1145/3308558.3313446.

[115] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein–ligand scoring with convolutional neural networks. *Journal of Chemical Information and Modeling*, 57(4):942–957, 2017. doi: 10.1021/acs.jcim.6b00740. URL https://doi.org/10.1021/acs.jcim.6b00740. PMID: 28368587.

[116] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500): 2323–2326, 2000.

[117] Benedek Rozemberczki, Ryan Davies, Rik Sarkar, and Charles Sutton. Gemsec: Graph embedding with self clustering. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, page 65–72, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368681. doi: 10.1145/3341161.3342890. URL https://doi.org/10.1145/3341161.3342890.

[118] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Re. Representation tradeoffs for hyperbolic embeddings. In *International Conference on Machine Learning*, pages 4460–4469, 2018.

[119] Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pages 355–366. Springer, 2011.

[120] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[121] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018.

[122] Daniel Selsam, Matthew Lamm, Benedikt Bünz, Percy Liang, Leonardo de Moura, and David L Dill. Learning a sat solver from single-bit supervision. *arXiv preprint arXiv:1802.03685*, 2018.

[123] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

[124] Martin Simonovsky and Nikos Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. *arXiv preprint arXiv:1802.03480*, 2018.

[125] Koustuv Sinha, Shagun Sodhani, Joelle Pineau, and William L Hamilton. Evaluating logical generalization in graph neural networks. *arXiv preprint arXiv:2003.06560*, 2020.

[126] Balasubramaniam Srinivasan and Bruno Ribeiro. On the equivalence between positional node embeddings and structural graph representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJxzFySKwH.

[127] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539, 2006.

[128] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.

[129] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[130] Alexandru Tifrea, Gary Becigneul, and Octavian-Eugen Ganea. Poincare glove: Hyperbolic word embeddings. In *International Conference on Learning Representations*, 2018.

[131] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, Alexander Bronstein, and Emmanuel Müller. Netlsd: Hearing the shape of a graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2347–2356, 2018.

[132] Anton Tsitsulin, Marina Munkhoeva, and Bryan Perozzi. Just slaq when you approximate: Accurate spectral distances for web-scale graphs. In *Proceedings of The Web Conference 2020*, WWW '20, page 2697–2703, 2020.

[133] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[134] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[135] Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019.

[136] Saurabh Verma and Zhi-Li Zhang. Stability and generalization of graph convolutional neural networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1539–1548, 2019.

[137] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

[138] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234. ACM, 2016.

[139] Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, et al. Deep graph library: Towards efficient and scalable deep learning on graphs. *arXiv preprint arXiv:1909.01315*, 2019.

[140] Jason Weston, Frédéric Ratle, and Ronan Collobert. Deep learning via semi-supervised embedding. In *Proceedings of the 25th international conference on Machine learning*, pages 1168–1175. ACM, 2008.

[141] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*, 2019.

[142] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[143] Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 40–48. JMLR. org, 2016.

[144] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[145] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4800–4810. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7729-hierarchical-graph-representation-learning-with-differentiable-pooling.pdf.

[146] Jiaxuan You, Rex Ying, Xiang Ren, William L Hamilton, and Jure Leskovec. Graphrnn: A deep generative model for graphs. *arXiv preprint arXiv:1802.08773*, 2018.

[147] Jiaxuan You, Rex Ying, and Jure Leskovec. Position-aware graph neural networks. *arXiv preprint arXiv:1906.04817*, 2019.

[148] Tao Yu and Christopher M De Sa. Numerically accurate hyperbolic embeddings using tiling-based models. In *Advances in Neural Information Processing Systems*, pages 2023–2033, 2019.

[149] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. Network representation learning: A survey. *IEEE Transactions on Big Data*, 2018.

[150] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[151] Ziwei Zhang, Peng Cui, and Wenwu Zhu. Deep learning on graphs: A survey. *arXiv preprint arXiv:1812.04202*, 2018.

[152] Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328, 2004.

[153] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.

[154] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.