Data-efficient graph learning: Problems, progress, and prospects

Received: 9 July 2024 | Accepted: 26 August 2024

DOI: 10.1002/aaai.12200

HIGHLIGHT

Data-efficient graph learning: Problems, progress, and prospects

Kaize Ding[1] | **Yixin Liu**[2] | **Chuxu Zhang**[3] | **Jianling Wang**[4]

[1]Northwestern University, Evanston, Illinois, USA

[2]Griffith University, Brisbane, Australia

[3]University of Connecticut, Storrs, Connecticut, USA

[4]Google DeepMind, Mountain View, California, USA

Correspondence
Kaize Ding, Northwestern University, Evanston, IL, USA.
Email: kaize.ding@northwestern.edu

Abstract

Graph-structured data, ranging from social networks to financial transaction networks, from citation networks to gene regulatory networks, have been widely used for modeling a myriad of real-world systems. As a prevailing model architecture to model graph-structured data, graph neural networks (GNNs) have drawn much attention in both academic and industrial communities in the past decades. Despite their success in different graph learning tasks, existing methods usually rely on learning from "big" data, requiring a large amount of labeled data for model training. However, it is common that real-world graphs are associated with "small" labeled data as data annotation and labeling on graphs is always time and resource-consuming. Therefore, it is imperative to investigate graph machine learning (graph ML) with low-cost human supervision for low-resource settings where limited or even no labeled data is available. This paper investigates a new research field—data-efficient graph learning, which aims to push forward the performance boundary of graph ML models with different kinds of low-cost supervision signals. Specifically, we outline the fundamental research problems, review the current progress, and discuss the future prospects of data-efficient graph learning, aiming to illuminate the path for subsequent research in this field.

INTRODUCTION

Recent years have witnessed rapid growth in our ability to generate and gather data from numerous platforms in the online world and various sensors in the physical world. Graphs serve as a common language for modeling a plethora of structured and relational systems, such as social networks, knowledge graphs, and academic graphs, where entities are denoted as nodes while their relations are denoted as edges. Modern graph learning algorithms, especially those empowered by graph neural networks (GNNs) (Shchur et al. 2018; Wu, Pan, Chen, et al. 2020) have received much research attention due to their significant impacts in addressing real-world problems. To harness the inherent structure among data, significant deep learning advances have been made in graph machine learning (graph ML), which has produced promising results in applications from diverse domains, ranging from cybersecurity (Zügner, Akbarnejad, and Günnemann 2018) to natural language processing (Ding, Wang, Li, Li, et al. 2020).

In general, existing graph ML algorithms focus on the setting where abundant human-annotated examples can be accessed during training. This assumption is often impractical since collecting ground-truth labels is a laborious process that requires intensive domain

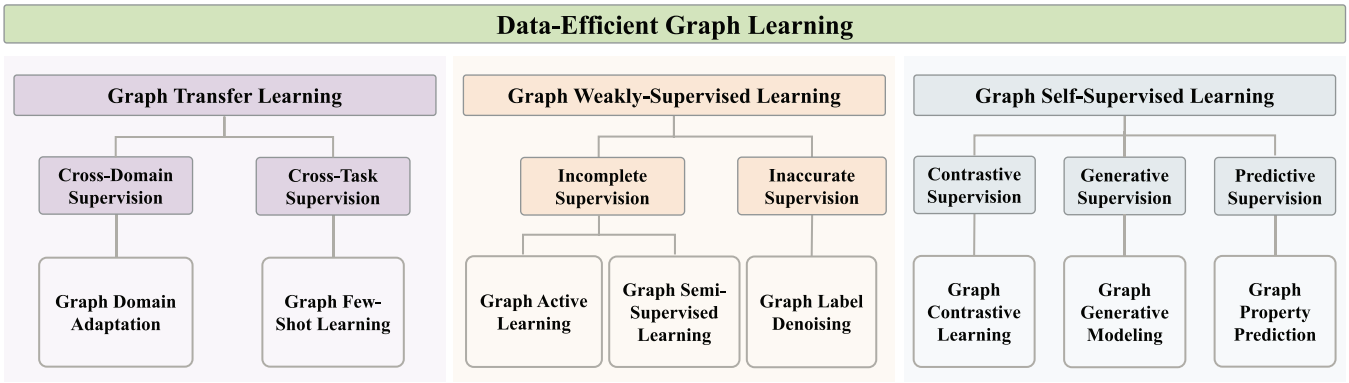

FIGURE 1 The proposed taxonomy of data-efficient graph learning.

knowledge, especially when considering the heterogeneity and complexity of graph-structured data (Yao et al. 2020; Ding, Wang, Li, Shu, et al. 2020). For example, molecular property testing for therapeutic activity requires a series of labor-intensive and resource-costly wet-lab experiments (Guo et al. 2021); recommender systems face cold-start challenges for new users or items coming to the system (Wang, Ding, and Caverlee 2021). As such, it is challenging yet imperative to study the problem of graph learning under low-resource settings, emphasizing the critical necessity of developing novel approaches that can operate effectively with limited or even no ground-truth labeled data.

To enhance our understanding of this new machine learning frontier, we propose a new taxonomy (as shown in Figure 1) that classifies the fundamental subproblems under data-efficient graph learning, which classifies existing works into three lines of research: (1) *graph transfer learning*, whose goal is to improve the data efficiency of graph ML models by transferring the supervision signals from other auxiliary domains or tasks; (2) *graph weakly-supervised learning*, which is a problem that focuses on learning data-efficient graph ML models for a specific task using inexpensive weak supervision signals, especially incomplete or inaccurate supervision; and (3) *graph self-supervised learning*, which aims to learn generalizable graph ML models without using any human supervisions. By learning from contrastive, generative, or predictive self-supervision, graph ML models can efficiently and effectively adapt to arbitrary down-stream tasks. To address each of the aforementioned fundamental problems, a series of research has been recently proposed, pushing forward the performance boundary of graph ML models with different types of low-cost supervision signals. Therefore, we systematically review the current progress and our contributions to this field. Finally, we discuss several potential directions in this area and provide insights for future research exploration.

GRAPH TRANSFER LEARNING

To compensate for the limited supervision signals in the target domain or task, we can leverage sufficient supervision signals or knowledge from auxiliary domains or tasks to enhance model training and improve data efficiency. graph transfer learning provides a solution by transferring the supervision signals from the auxiliary domains or tasks, enabling models to achieve strong performance under low-resource settings.

Graph ML with cross-domain supervision

Graph domain adaptation. As one important field in graph transfer learning, graph domain adaptation (GDA) focuses on transferring knowledge from well-labeled graphs in one or multiple domains to a sparsely labeled graph in another domain by bridging the distribution shifts between different graphs. In particular, most of the graph domain adaptation methods try to produce domain-invariant node representations either explicitly minimizing the distance between representations in two domains (Shen et al. 2020) or implicitly confounding a domain discriminator using adversarial learning (Wu, Pan, Zhou, et al. 2020; Zhang et al. 2021). For the predefined domain discrepancy metrics minimization models, node representations are first extracted from the GNN encoder, and then domain-invariant representations are learned by minimizing predefined domain discrepancy metrics, such as MMD (Kumagai and Iwata 2019), and so forth. Instead of explicitly minimizing domain discrepancies, adversarial learning methods integrate an encoder with a domain classifier that predicts the origin domain of the representation. Among them, UDAGCN (Wu, Pan, Zhou, et al. 2020) and AdaGCN (Dai et al. 2022) integrate graph convolution with adversarial training for graph transfer learning. Commander (Ding, Shu, et al. 2021) adopts similar idea

to solve the problem of cross-domain graph anomaly detection. ACDNE (Shen et al. 2020) utilizes the gradient reversal layer (GRL) (Ganin et al. 2016) to make node representation domain-invariant. ASN (Zhang et al. 2021) further improves node representations with attention mechanisms and disentanglement, where domain-private and domain-shared information are preserved.

In addition, meta-learning approaches have also been explored in GDA. Li and Hospedales (2020) extract comprehensive meta-knowledge from source domains to learn a well-generalized model initialization, facilitating fast adaptation to target domains. CrossHG-Meta (Zhang et al. 2022a) combines meta-learning and GDA by simulating source-target distributions within the inner-loop of MAML, addressing the inaccessibility of the target domain during meta-training. This approach splits available source domains into virtual source and target domains, eliminating marginal shifts between them through discrepancy measurement. CDTC (Zhang et al. 2023) addresses source-target shifts by selecting the most target-relevant meta-tasks in source domains for meta-training, requiring a few labeled target samples to form target prompt tasks. GDN (Ding, Zhou, et al. 2021) has experimentally demonstrated that meta-learning alone can enhance model performance in cross-domain node classification tasks.

Graph ML with cross-task supervision

Graph few-shot learning. Given the rapidly evolving nature of real-world graphs, many practical graph applications require a Graph ML model to possess the capability of dealing with *never-before-seen* node classes (e.g., newly created user interest group), relations (e.g., newly extracted relations in knowledge graph), or even graphs (e.g., a protein—protein interaction graph from a new organism), using only a handful of labeled data samples. Despite that humans are capable of learning new tasks rapidly by utilizing what they learned in the past, current AI techniques cannot rapidly generalize from a few examples. As such, it is challenging yet imperative to study the problem of *graph few-shot learning* that involves transferring knowledge from auxiliary tasks that have a disjoint label space from the target task. In general, graph few-shot learning methods can be divided into the following two categories: (1) meta-initialization methods and (2) metric learning methods.

As a straightforward application of model-agnostic meta-learning (MAML) (Finn, Abbeel, and Levine 2017) on graph-structured data, Meta-GNN (Zhou, Cao, et al. 2019) adopts a GNN encoder to learn transferable node representations for few-shot node classification. AMM-GNN (Wang et al. 2020) uses an attribute-level attention mechanism to characterize the feature distribution differences between tasks and learn more meaningful transferable knowledge across tasks. Subsequently, RALE (Liu, Fang, et al. 2021) highlights the importance of learning the dependencies among the nodes in a task, and proposes to use hub nodes to update the representations learned by a GNN. Moreover, HG-Meta (Zhang et al. 2022b) proposes to address few-shot node classification on heterogeneous graphs by modeling both graph structure heterogeneity and task diversity. Apart from node classification, Meta-KGR (Lv et al. 2019) proposes to tackle few-shot multihop relation prediction by adopting MAML to learn effective meta parameters from high-frequency relations that could quickly adapt to few-shot relations. Later, FIRE (Zhang et al. 2020) extends Meta-KGR with a heterogeneous neighbor aggregator and a search space pruning strategy. More recently, ADK-KG (Zhang, Qian, et al. 2022) further improves FIRE by enhancing the neighbor aggregator with node text content and augmenting MAML with task weight. To solve the few-shot graph classification problem, Chauhan, Nathani, and Kaul (2019) make use of the graph's spectral measures to generate a set of super-classes and a super-graph to better model the latent relations between classes. Based on that, the GNN model is pretrained to learn a good initialization and can be fine-tuned with a few examples during the test phase. Moreover, to control the meta-learner's adaptation step, AS-MAML (Ma et al. 2020) leverages a novel step controller using reinforcement learning, improving the robustness and generalization of the meta-learner.

As the first metric learning endeavor, GPN (Ding, Wang, Li, Shu, et al. 2020) is proposed for few-shot node classification, which improves upon the previous method with a prototype representation for each class and adopts meta-learning to learn the prototype representation as the weighted average of representations of each class. MetaTNE (Lan et al. 2020) is able to automatically capture the relationships between the graph structure and the node labels as prior knowledge, which is leveraged to help recognize novel labels with only a few support nodes. It is designed for plain graphs and is more applicable to content-less scenarios. Unlike GFL using whole graph information, G-Meta (Huang and Zitnik 2020) uses local subgraphs to transfer subgraph-specific information and learns transferable knowledge faster via meta gradients. Moreover, HAG-Meta (Tan et al. 2022) leverages both node-level attention and task-level attention to improve class prototype computation, which is further trained via an incremental learning paradigm. Similarly, TENT (Wang et al. 2022) leverages node-, class-, and task-level adaptations in each meta-task to advance the model generalization performance on meta-test tasks. Unlike previous efforts focusing on the few-shot node classification task, Meta-Graph (Bose et al. 2019) combines

meta-learning and GCN to solve few-shot link prediction across multiple graphs. GMatching (Xiong et al. 2018) proposes to solve the one-shot relation prediction problem on knowledge graphs. Later, FSRL (Zhang et al. 2020) is proposed to extend GMatching to the few-shot scenario by attentively aggregating all support samples of each relation and improving node embedding formulation with a heterogeneous neighbor aggregator. FAAN (Sheng et al. 2020) further improves FSRL by designing an adaptive attentional network to learn adaptive node and reference representations. Furthermore, REFORM (Wang, Huang, et al. 2021) introduces an error mitigation module to alleviate the negative impact of errors incorporated into the problem of knowledge graph completion. More recently, we have tried to improve the effectiveness of graph few-shot learning methods when facing different challenges, such as weak supervision (Ding, Wang, et al. 2024; Wang et al. 2023), catastrophic forgetting issue (Tan et al. 2022; Xu et al. 2022), and so forth.

GRAPH WEAKLY SUPERVISED LEARNING

Existing graph ML algorithms are mainly developed for the supervised or semi-supervised setting where the input graph usually has relatively plenty of labeled instances. However, weakly supervised graph ML, in particular where only *incomplete* and *inaccurate* supervision signals are provided, remains largely understudied in the community.

Graph ML with incomplete supervision

As one major challenge in weakly supervised learning, incomplete supervision concerns the situation where the model is given a small amount of labeled data, which is insufficient to train a good learner. As abundant unlabeled data are commonly available in this setting, the key of addressing incomplete supervision is to maximize the effective utilization of structural and feature information of unlabeled data. Despite that GNNs (Kipf and Welling 2017) have become successful attempts to revisit semi-supervised learning with neural networks, the main concern in semi-supervised node classification, that is, the shortage of supervision information, has not been fully addressed.

Graph semi-supervised learning. When the labeled training data are extremely limited on graphs, the performance of GNNs commonly faces a catastrophic decline as other neural endeavors. The major reason is that conventional GNNs with shallow architecture will severely restrict the efficient propagation of label signals, but when a GNN model is equipped with many layers, it will usually suffer from the notorious over-smoothing problem (Li, Han, and Wu 2018), resulting in the restriction of the performance of GCNs in the case of few labels. To counter this, researchers also try to increase the message-passing range or receptive fields of GNNs by proposing different techniques, such as adding advanced normalizations (Zhao and Akoglu 2019; Li, Muller, et al. 2019), decoupling the feature transformation and propagation steps (Klicpera, Bojchevski, and Günnemann 2019; Liu, Gao, and Ji 2020; Dong et al. 2021). For example, APPNP (Klicpera, Bojchevski, and Günnemann 2019) propagates the neural predictions via personalized PageRank, which can preserve the node's local information while increasing the receptive fields. DAGNN (Liu, Gao, and Ji 2020) decouples the propagation and transformation steps and then utilizes an adaptive adjustment mechanism to balance the information from the local and global neighborhoods of each node. GLP and IGCN (Li, Wu, et al. 2019) are two models that combine label propagation and GCN from a unifying graph filtering perspective.

In the meantime, self-training (Yarowsky 1995), also known as pseudo-labeling (Lee 2013), where one imputes labels on unlabeled data based on a teacher model trained with limited labeled data, has been applied to improve GNNs to solve the problem of semi-supervised node classification. To overcome the limits and realize the full potential of the GCN model, Li, Han, and Wu (2018) first propose a co-training approach and a self-training approach to train GCNs. The accumulative error of pseudo-labels will gradually increase and degrade the model. M3S (Sun, Lin, and Zhu 2020) is a multistage self-training model that utilizes a clustering method (i.e., DeepCluster) to eliminate the pseudo labels that may be incorrect. Similarly, CGCN (Hui, Zhu, and Hu 2020) generates pseudo labels by combining variational graph auto-encoder with Gaussian mixture models. Similar ideas can also be found in Zhou, Shi, et al. (2019) and Dai, Aggarwal, and Wang (2021). More recently, Dong et al. (2021) propose to generate pseudo labels via label propagation and then transform the features via simple neural networks, which combine the power of graph self-training and deeper GNNs in a decoupled architecture.

To leverage the benefits of both long-range message passing and pseudo labeling, we propose Meta-PN (Ding et al. 2022), which adopts meta-learning to enable an adaptive label propagation process. The generated soft pseudo labels not only capture informative local and global structure information, but more importantly, have aligned data usage with the gold-labeled nodes. Similarly, in our work AGST (Ding, Nouri et al. 2024), we propose a weakly supervised contrastive learning algorithm to optimize the semantic alignment between the few labeled nodes and

generated pseudo-labeled nodes by encouraging intraclass compactness and interclass separability in the latent feature space. Moreover, we investigate the incomplete weak information problem, which has incomplete structure, incomplete features, and incomplete labels together, and propose a new solution to solve it by leveraging the idea of long-range propagation and consistency training (Liu, Ding, Wang, et al. 2023).

Graph active learning. The goal of active learning is to make use of expert knowledge more efficiently by having them label data points which are estimated to be most valuable to the model. In each trial, the active learner selects new data instance(s) to be labeled. Recently, there are also a handful of studies focusing on active learning for graph-structured data. Some earlier studies (Gu et al. 2013) are developed based on the graph homophily assumption that neighboring nodes are more likely to have the same label, and utilize theories of graph signal processing to select nodes for active learning. More recent works utilize the expressive power of GNNs to design more informative selection criteria. For example, SEAL (Li, Yin, and Chen 2019) integrates deep neural networks and adversarial learning with active learning, the divergence of unlabeled data and labeled ones is used for selecting the query nodes. More recently, we propose a novel multitask active graph anomaly detection framework (Chang et al. 2024) to detect anomalies within a limited labeling budget, which actively queries nodes with the guidance of supervision signals from auxiliary task.

Notably, multi-armed bandit (MAB) and reinforcement learning have been proven as effective solutions for addressing graph active learning problems. ANRMAB (Gao et al. 2018) uses the combination of different heuristics, and dynamically adjusts the combination weights based on a multi-armed bandit (MAB) framework. Similarly, ActiveHNE (Chen et al. 2019) tackles active learning on heterogeneous graphs by posing it as an MAB problem. GraphUCB (Ding, Li, and Liu 2019) is an MAB-based algorithm for solving graph-based anomaly detection. However, these methods either greedily choose the most informative node in each single query step (AGE, GraphUCB) or maximize a surrogate reward signal (ANRMAB and ActiveHNE), but fail to directly optimize the long-term performance gain of the whole query sequence. Hu, Xiong, et al. (2020) propose GPA, which learns transferrable knowledge across different graphs, and is able to label nodes on the target graph in a zero-shot setting. They train the graph policy network on multiple training graphs where node labels are available, and directly transfer the learned policy to perform active learning on unseen test graphs where no labels are provided initially.

Graph ML with inaccurate supervision

To address the data insufficiency problem, one alternative solution is to adopt automatic labeling tools based on heuristics, crowd-sourcing, or weak-learners to obtain more labeled data. Though using such weakly labeled data is more practical, it poses an inaccurate supervision challenge since the supervision data are not always clean-labeled. In the meantime, the performance of learning models could be noticeably degraded in the presence of label noise, especially in an adversary environment (Hendrycks et al. 2018).

Graph label denoising. To tackle the inaccurate supervision challenge, label denoising has become essential for learning with weak supervision. As a generalization of neural networks on graph-structured data, GNNs are also vulnerable to noisy labels (NT et al. 2019; Dai, Aggarwal, and Wang 2021). On the one hand, the smoothing operation during message-passing can be disrupted when the training data are corrupted with label noise; on the other hand, the noisy label information will pass to the unlabeled nodes during the learning process, which severely degrades the performances of GNNs. As the training proceeds, GCNs would completely fit the noisy labels, resulting in degraded performance and poor generalization (NT et al. 2019). However, very few efforts are taken to address the problem of learning GNNs on graphs with noisy labels. The existing label denoising techniques, however, cannot be directly applied to graph-structured data due to the challenges in terms of label sparsity and label dependency. To mitigate the detrimental effects of noisy labels in graph representation learning, the research on graph label denoising has received increasing endeavors. For example, NT et al. first applies the backward loss correction procedure to GNNs to estimate a noise transition matrix. GraphLT (Zhuang and Hasan 2020) is a new graph label transition model that can improve the GNN robustness by transiting the node latent representation based on dynamic conditional label transition. UnionNET (Li, Yin, and Chen 2021) is a unified robust training framework for GNNs that performs sample reweighting and label correction simultaneously. Dai, Aggarwal, and Wang investigate a novel framework that could achieve robustness towards noisy labels in graphs by carefully connecting unlabeled nodes with (pseudo) labeled nodes. PI-GNN (Du et al. 2021) leverages explicit pairwise interactions to mitigate the effect of noisy labels on graphs.

Though pseudo labeling has shown to be empirically effective in improving model performance in semi-supervised node classification, its effectiveness largely relies on the quality of the obtained pseudo labels (Liu, Hu, et al. 2022; Ding et al. 2022). In practice, pseudo

labels could be quite noisy, especially under the setting where the training labels are not clean. Therefore, how to cope with the generated noisy pseudo labels along with the given noisy labels by leveraging the knowledge inherent in the graph itself becomes the main challenge. Specifically, we develop a new GNN learning backbone that is inherently robust to learning with noisy labels on graphs. With just a simple architecture, this new model inherently addresses the label noise memorization and propagation issues (Ding, Ma, et al. 2024). Meanwhile, the proposed reliable graph pseudo labeling algorithm can effectively leverage the knowledge of unlabeled nodes while mitigating the adverse effects of noisy pseudo labels by considering the neighbor uncertainty of predicted pseudo label.

GRAPH SELF-SUPERVISED LEARNING

To alleviate the demand for massive labeled data and provide sufficient supervision in Graph ML, graph self-supervised learning (Graph SSL) has been introduced and achieved great success in representation learning and model pretraining. Unlike the traditional supervised paradigm that relies on annotated labels as supervision signals, graph SSL extracts supervision signals from the graph data itself and employs carefully crafted pretext tasks for model training. With distinct designs of pretext graph learning tasks, graph SSL approaches can be divided into three categories: (1) *graph contrastive learning*; (2) *graph generative modeling*, and (3) *graph property prediction*.

Graph ML with contrastive supervision

Graph contrastive learning. Graph contrastive learning (GCL) methods learn by predicting the agreement between two elements (i.e., nodes, subgraphs, and/or graphs) in graph data. Specifically, the agreement between elements with similar semantic information (denoted as a *positive pair*) is maximized, while those with unrelated semantic information (denoted as a *negative pair*) are minimized. The elements for contrast can be heuristically defined as specific segments in graph data or synthesized through data augmentations. Then, the maximization and minimization of agreement between elements can be achieved by contrastive objective functions, such as InfoNCE (van den Oord, Li and Vinyals 2018; He et al. 2020) and Jensen–Shannon divergence (Hjelm et al. 2018). In this subsection, we categorize GCL methods into two subclasses based on whether two contrastive elements belong to the same scale.

The first group of GCL methods conducts contrastive learning between two elements on the same scale. Following the breakthroughs in contrastive visual feature learning (He et al. 2020; Chen et al. 2020), the pioneering studies generate augmented examples from the input graph and view two augmented examples that are from the same original sample as a positive pair, while those from different original samples are negative. For instance, GCC (Qiu et al. 2020) uses subgraphs as contrastive instances to pretrain the graph encoder and employs random walk with restart (RWR) as the augmentation strategy to generate subgraphs. GraphCL (You, Chen, Sui, et al. 2020) aims to learn graph-level representations via graph-level contrastive learning, where the augmented graph samples are acquired by four augmentation strategies: node dropping, edge perturbation, attribute masking, and subgraph sampling. GRACE (Zhu et al. 2020) focuses on node-level contrastive learning and considers both intra-view and inter-view negative pairs for contrast purposes. Two augmentation strategies, edge removing and node feature masking, are considered to generate augmented views of graph data.

Parallelly to same-scale GCL, another line of GCL studies focuses on learning representations by contrasting the graph elements in different scales, for example, node-graph and node-subgraph contrasting. In general, most of these methods also adopt the idea of MI maximization but leverage the Jensen–Shannon divergence as their MI estimator. For example, DGI (Velickovic, Fedus, and Hamilton 2019) and InfoGraph (Sun et al. 2019) perform contrasting between graph- and node-level representations. GMI (Peng et al. 2020) proposes to jointly maximize feature MI (between the node's embedding and raw features of its neighbors) and edge MI (embedding of two adjacent nodes) for graph representation learning. MVGRL (Hassani and Khasahmadi 2020) suggests a multiview contrasting where the original graph structure and graph diffusion are regarded as two different views. Context prediction proposed by Hu, Liu, et al. (2020) enhances the agreement between nodes and context subgraphs whose embedding is acquired by an auxiliary GNN. SLiCE (Wang, Agarwal, et al. 2021) introduces a contextual node prediction task that maximizes the occurrence probability between a context graph and a randomly masked node within this subgraph.

Despite their effectiveness, the above methods fail to learn high-quality representations in several situations. To learn informative representations for graphs with low homophily, in our work GREET (Liu, Zheng, et al. 2023), we propose an edge discriminating strategy to define homophilic and heterophilic views in GCL and employ low-pass and high-pass GNNs to generate representations for two views, respectively. Also, to capture global

knowledge with GCL, we propose S^3-CL (Ding et al. 2023), a novel GCL approach to capture both structural and semantic knowledge via Bayesian prototype inference and multiscale feature propagation.

On the basis of GCL, we further develop innovative solutions for various graph learning tasks. For graph anomaly detection task, we first propose an unsupervised detection method termed CoLA (Liu, Li, et al. 2021) with cross-scale contrast, and then develop a multiscale GCL framework for the same task (Jin et al. 2021). We also integrate GCL with information bottleneck theory to boost the explainability of graph anomaly detection models (Liu, Ding, Lu, et al. 2023). For graph structure learning task, we develop the first unsupervised structure learning method for graph structure refinement and inference based on same-scale GCL (Liu, Zheng, et al. 2022). Meanwhile, aiming to identify the out-of-distribution samples from graph data, we also develop an out-of-distribution detection method based on hierarchical GCL (Liu, Ding, Liu, et al. 2023). Our recent work GRENADE (Li, Ding, and Lee 2023) leverages GCL to align the representations learned from different modalities in text-attributed graphs.

Graph ML with generative supervision

Graph generative modeling. The generative self-supervision aims to reconstruct the input data and use the input data as their supervision signals. Following the idea of Autoencoder (Baldi 2012), which learns to compress data vectors into low-dimensional representations with the encoder network and then try to rebuild the input vectors with the decoder network, a series of research on leveraging generative self-supervision has been proposed.

The early generative graph SSL approaches learn by reconstructing the structural information in graph data, that is, adjacency matrix. As one representative work, GAE (Kipf and Welling 2016) is the simplest instance of the structure generation method, and VGAE (Kipf and Welling 2016) further integrates the idea of variational autoencoder into GAE. Following VGAE, ARGA/ARVGA (Pan et al. 2018) regularizes the GAE/VGAE model with generative adversarial networks (GANs), which enforces the distribution of latent representations to match the Gaussian prior. Instead of rebuilding the full graph, another solution is to reconstruct the masked edges. Denoising link reconstruction (Hu et al. 2019) randomly drops existing edges to obtain the perturbed graph and tries to recover the discarded connections with a pairwise similarity-based decoder trained by a BCE loss. Similar ideas can be found in Zhu, Du, and Yan (2020) and Jin et al. (2020).

Generating features in graph data can also serve as powerful supervision signals in graph SSL. Some pioneering studies (Hu, Dong, et al. 2020; You, Chen, Wang, et al. 2020; Jin et al. 2020) try to use feature generation as one of the generative graph SSL strategies, but these methods have limited performance in graph representation learning. Motivated by the success of MaskAutoencoder (MAE) (He et al. 2022) in visual representation learning, some recent works based on the learning paradigm of MAE display better performance. For example, GraphMAE (Hou et al. 2022) employs a GNN decoder and remask mechanism for hidden features, achieving superior performance on node- and graph-level tasks. The following work, GraphMAE2 (Hou et al. 2023), further equips GraphMAE with a latent reconstruction target and multiview remask mechanism. GiGaMAE (Shi et al. 2023) follows the paradigm of MAE with both feature generation and structure generation, yielding its strong capability in diverse downstream tasks.

Despite their effectiveness, the existing generative methods can only deal with homogeneous graphs but fail to handle heterogeneous graphs. To bridge the gap, we propose HGMAE (Tian et al. 2023), a generative graph SSL method designed for heterogeneous graphs, with specifically designed metapath masking and adaptive attribute masking mechanisms. Meanwhile, we produce a unified framework for MAE-based methods termed UGMAE (Tian et al. 2024), which consists of several useful components, including adaptive feature mask generator, ranking-based structure reconstruction, bootstrapping-based similarity, and consistency assurance. UGMAE achieves state-of-the-art performance on various learning tasks.

Based on generative Graph SSL techniques, we also develop a series of approaches for graph anomaly detection, which is a critical task in graph learning. Equipped with structure and feature generation, DOMINANT (Ding, Li, Bhanushali, et al. 2019) is an anomaly detection approach that can identify anomalous nodes from a graph without the guidance of labels. We also propose an inductive graph anomaly detection approach, AEGIS (Ding, Li, et al. 2021), which combines generative graph SSL with adversarial learning to identify anomalies in unseen testing graphs.

Graph ML with predictive supervision

Graph property prediction. Predictive self-supervision signals from the node-, link-, and graph-level properties can be obtained from the graph data freely. These methods have a similar training paradigm with supervised learning since both of them learn with "sample-label" pairs. Their difference lies in how the label is obtained: In supervised learning, the manual label is human-annotated which

often needs expensive costs; in predictive graph SSL, the pseudo label is self-generated automatically without any data labeling cost.

Jin et al. (2020) first propose different graph property prediction pretexts: (1) NodeProperty, which aims to predict some representative node properties, such as node degree, to form a local-structure aware pretext task; (2) Distance2Cluster, which predicts the distance between unlabeled nodes and predefined clusters in the graph. It encourages the node representation to be aware of its global positioning while training; and (3) PairwiseAttrSim, which enforces the similarity of a pair of nodes to be close to their feature similarity on a set of representative node pairs (i.e., with the highest or lowest attribute similarity). Similarly, GROVER (Rong et al. 2020) performs self-supervised learning on molecular graph data with two predictive learning tasks. The first task (i.e., contextual property prediction) aims to predict the statistical properties of anchor nodes' subgraphs (e.g., node-edge-counts), while the second task (i.e., graph-level motif prediction) is to identify whether a graph contains specific motifs predefined by domain knowledge. Centrality score ranking (Hu et al. 2019) leverages node centrality (e.g., eigencentrality and subgraph centrality) to rank nodes and estimate the ranking score for any two nodes in the graph to preserve different-level graph information.

In addition, for methods built on the idea of clustering, it is intuitive to use obtained pseudo labels as the predictive self-supervision. M3S (Sun, Lin, and Zhu 2020) iteratively trains an encoder by leveraging DeepCluster (Caron et al. 2018) to assign pseudo labels to those unlabeled nodes at each training stage. You, Chen, Wang, et al. propose node clustering to use the indices of precomputed clusters as self-supervised labels, which is similar to the cluster preserving introduced by Hu et al. (2019). In addition to feature-based clustering, You, Chen, Wang, et al. (2020) also introduce another structural-aware task termed graph partitioning to group nodes based on the inherent topology where the connections across subsets are minimized. After this, partition indices are assigned as node labels in different subsets, as in node clustering. To absorb the advantages of both attributive and structural-based clustering, CAGNN (Zhu et al. 2021) first assigns cluster indices as pseudo labels and then refines the clusters by minimizing the intercluster edges.

FUTURE RESEARCH CHALLENGES

In this section, we discuss three future research directions and corresponding challenges to provide a deeper understanding of data-efficient graph learning.

Theoretical Understanding of data-efficient graph learning

While transfer learning, weakly supervised learning and self-supervised learning have achieved promising results in data-efficient graph ML, their full potential have yet to be fully optimized due to the unique characteristics and diverse properties of graph-structured data (e.g., size, density, homophily). In the future, we should keep exploring and developing more fundamental and theoretical understanding of these methodologies considering the specialties of graph-structured data. This includes identifying their theoretical limits under various low-resource graph scenarios and establishing rigorous guarantees like generalization bounds and convergence analysis. Such explorations will not only enhance our comprehension of data-efficient graph learning but also guide the development of novel and generalizable approaches with enhanced robustness. Furthermore, it can enable us to extend our findings in graph ML to the broader machine learning community and foster cross-disciplinary advancements.

Data-efficient graph learning for complex graphs

The complexity of real-world graphs, characterized by their scale, heterogeneity, and dynamic nature, poses significant challenges for data-efficient graph learning. Real-world graphs often contain heterogeneous nodes and edges of different types (i.e., users, items, and relationships) and they evolve over time as relationships and preferences change. Future research should focus on developing data-efficient graph ML algorithms that can efficiently handle different types of complex graphs such as heterogeneous graphs (Zhang et al. 2022a), hypergraphs (Ding, Wang, Li, Li, et al. 2020), and spatio-temporal graphs (Sheth et al. 2023). Innovative algorithm design will be crucial to maintain both efficiency and effectiveness as complexity increases. Additionally, to support and evaluate progress, establishing standardized benchmarks and evaluation metrics specifically designed for complex graphs is urgently needed. This will enable fair comparisons of proposed methods and promote solutions tailored to the challenges posed by real-world graph data.

Data-efficient graph learning with large language models

The comprehensive reasoning capabilities encoded within LLMs (Brown et al. 2020; Achiam et al. 2023) grant them the potential to provide low-cost supervision signals for graph ML in low-resource scenarios. LLMs have demonstrated efficacy as data augmenters across diverse domains (Wang et al. 2024; Li et al. 2024; Ding, Qin, et al. 2024). One promising approach is to firstly convert graph structures into textual descriptions or leveraging existing textual

attributes associated with nodes and edges, and then enable LLMs to augment existing supervision through synthetic labels, links, or attributes. However, maintaining the quality and accuracy of generated annotations while mitigating the risk of hallucinations remains a critical challenge. Another approach is to utilize LLMs as agents with expert knowledge during the graph active learning process. To harness the power of LLMs for data-efficient graph ML, we need to focus on developing effective methods for transforming graph structures into meaningful text representations, ensuring the fidelity and quality of these representations, and investigating optimal strategies to integrate LLMs' capabilities within graph ML frameworks.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

ORCID

Kaize Ding https://orcid.org/0000-0001-6684-6752

REFERENCES

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, et al. 2023. "Gpt-4 Technical Report." *arXiv preprint arXiv:2303.08774*.

Baldi, Pierre. 2012. "Autoencoders, Unsupervised Learning, and Deep Architectures." In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning, JMLR Workshop and Conference Proceedings*, 37–49.

Bose, Avishek Joey, Ankit Jain, Piero Molino, and William L Hamilton. 2019. "Meta-Graph: Few Shot Link Prediction via Meta Learning." *arXiv preprint arXiv:1912.09867*.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models are Few-Shot Learners." In *NeurIPS*.

Caron, Mathilde, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. "Deep Clustering for Unsupervised Learning of Visual Features." In *ECCV*.

Chang, Wenjing, Kay Liu, Kaize Ding, Philip S Yu, and Jianjun Yu. 2024. "Multitask Active Learning for Graph Anomaly Detection." *arXiv preprint arXiv:2401.13210*.

Chauhan, Jatin, Deepak Nathani, and Manohar Kaul. 2019. "Few-Shot Learning on Graphs via Super-Classes Based on Graph Spectral Measures." In *ICLR*.

Chen, Xia, Guoxian Yu, Jun Wang, Carlotta Domeniconi, Zhao Li, and Xiangliang Zhang. 2019. "ActiveHNE: Active Heterogeneous Network Embedding." In *IJCAI*.

Chen, Ming, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. 2020. "Simple and Deep Graph Convolutional Networks." In *ICML*.

Dai, Enyan, Charu Aggarwal, and Suhang Wang. 2021. "NRGNN: Learning a Label Noise-Resistant Graph Neural Network on Sparsely and Noisily Labeled Graphs." In *KDD*.

Dai, Quanyu, Xiao-Ming Wu, Jiaren Xiao, Xiao Shen, and Dan Wang. 2022. "Graph Transfer Learning via Adversarial Domain Adaptation with Graph Convolution." *TKDE*.

Ding, Kaize, Jundong Li, Rohit Bhanushali, and Huan Liu. 2019. "Deep Anomaly Detection on Attributed Networks." In *Proceedings of the 2019 SIAM International Conference on Data Mining*, 594–602. SIAM.

Ding, Kaize, Jundong Li, and Huan Liu. 2019. "Interactive Anomaly Detection on Attributed Networks." In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 357–65.

Ding, Kaize, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. 2020. "Be More with Less: Hypergraph Attention Networks for Inductive Text Classification." In *EMNLP*.

Ding, Kaize, Jianling Wang, Jundong Li, Kai Shu, Chenghao Liu, and Huan Liu. 2020. "Graph Prototypical Networks for Few-Shot Learning on Attributed Networks." In *CIKM*.

Ding, Kaize, Jundong Li, Nitin Agarwal, and Huan Liu. 2021. "Inductive Anomaly Detection on Attributed Networks." In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 1288–94.

Ding, Kaize, Kai Shu, Xuan Shan, Jundong Li, and Huan Liu. 2021. "Cross-Domain Graph Anomaly Detection." *TNNLS* 33: 2406–15.

Ding, Kaize, Qinghai Zhou, Hanghang Tong, and Huan Liu. 2021. "Few-Shot Network Anomaly Detection via Cross-Network Meta-Learning." *arXiv preprint arXiv:2102.11165*.

Ding, Kaize, Jianling Wang, James Caverlee, and Huan Liu. 2022. "Meta Propagation Networks for Graph Few-Shot Semi-Supervised Learning." In *AAAI*.

Ding, Kaize, Yancheng Wang, Yingzhen Yang, and Huan Liu. 2023. "Eliciting Structural and Semantic Global Knowledge in Unsupervised Graph Contrastive Learning." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 7378–86.

Ding, Bosheng, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. "Data Augmentation Using LLms: Data Perspectives, Learning Paradigms and Challenges." *arXiv preprint arXiv:2403.02990*.

Ding, Kaize, Xiaoxiao Ma, Yixin Liu, and Shirui Pan. 2024. "Divide and Denoise: Empowering Simple Models for Robust Graph Learning Against Label Noise." In *KDD*.

Ding, Kaize, Elnaz Nouri, Guoqing Zheng, Huan Liu, and Ryen White. 2024. "Toward Robust Graph Semi-Supervised Learning Against Extreme Data Scarcity." *TNNLS*.

Ding, Kaize, Jianling Wang, Jundong Li, James Caverlee, and Huan Liu. 2024. "Robust Graph Meta-Learning for Weakly Supervised Few-Shot Node Classification." *TKDD* 18, 1–18.

Dong, Hande, Jiawei Chen, Fuli Feng, Xiangnan He, Shuxian Bi, Zhaolin Ding, and Peng Cui. 2021. "On the Equivalence of Decoupled Graph Convolution Network and Label Propagation." In *The Web Conference*.

Du, Xuefeng, Tian Bian, Yu Rong, Bo Han, Tongliang Liu, Tingyang Xu, Wenbing Huang, and Junzhou Huang. 2021. "PI-GNN: A Novel Perspective on Semi-Supervised Node Classification Against Noisy Labels." *arXiv preprint arXiv:2106.07451*.

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. 2017. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks." In *ICML*.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. "Domain-Adversarial Training of Neural Networks." *JMLR* 17: 2096–30.

Gao, Li, Hong Yang, Chuan Zhou, Jia Wu, Shirui Pan, and Yue Hu. 2018. "Active Discriminative Network Representation Learning." In *IJCAI*.

Gu, Quanquan, Charu Aggarwal, Jialu Liu, and Jiawei Han. 2013. "Selective Sampling on Graphs for Classification." In *KDD*.

Guo, Zhichun, Chuxu Zhang, Wenhao Yu, John Herr, Olaf Wiest, Meng Jiang, and Nitesh V Chawla. 2021. "Few-Shot Graph Learning for Molecular Property Prediction." In *The Web Conference*.

Hassani, Kaveh, and Amir Hosein Khasahmadi. 2020. "Contrastive Multi-View Representation Learning on Graphs." In *ICLR*.

He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. "Momentum Contrast for Unsupervised Visual Representation Learning." In *CVPR*.

He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. "Masked Autoencoders are Scalable Vision Learners." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

Hendrycks, Dan, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. 2018. "Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise." In *NeurIPS*.

Hjelm, R Devon, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. "Learning Deep Representations by Mutual Information Estimation and Maximization." In *ICLR*.

Hou, Zhenyu, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. 2022. "GraphMAE: Self-Supervised Masked Graph Autoencoders." In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 594–604.

Hou, Zhenyu, Yufei He, Yukuo Cen, Xiao Liu, Yuxiao Dong, Evgeny Kharlamov, and Jie Tang. 2023. "GraphMAE2: A Decoding-Enhanced Masked Self-Supervised Graph Learner." In *TheWebConf*.

Hu, Ziniu, Changjun Fan, Ting Chen, Kai-Wei Chang, and Yizhou Sun. 2019. "Pre-Training Graph Neural Networks for Generic Structural Feature Extraction." *arXiv preprint arXiv:1905.13728*.

Hu, Shengding, Zheng Xiong, Meng Qu, Xingdi Yuan, Marc-Alexandre Côté, Zhiyuan Liu, and Jian Tang. 2020. "Graph Policy Network for Transferable Active Learning on Graphs." In *NeurIPS*.

Hu, Weihua, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. 2020. "Strategies for Pre-Training Graph Neural Networks." In *ICLR*.

Hu, Ziniu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. 2020. "GPT-GNN: Generative Pre-Training of Graph Neural Networks." In *KDD*.

Huang, Kexin, and Marinka Zitnik. 2020. "Graph Meta Learning via Local Subgraphs." In *NeurIPS*.

Hui, Binyuan, Pengfei Zhu, and Qinghua Hu. 2020. "Collaborative Graph Convolutional Networks: Unsupervised Learning Meets Semi-Supervised Learning." In *AAAI*.

Jin, Wei, Tyler Derr, Haochen Liu, Yiqi Wang, Suhang Wang, Zitao Liu, and Jiliang Tang. 2020. "Self-Supervised Learning on Graphs: Deep Insights and New Direction." *arXiv preprint arXiv:2006.10141*.

Jin, Ming, Yixin Liu, Yu Zheng, Lianhua Chi, Yuan-Fang Li, and Shirui Pan. 2021. "ANEMONE: Graph Anomaly Detection with Multi-Scale Contrastive Learning." In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3122–26.

Kipf, Thomas N., and Max Welling. 2016. "Variational Graph Auto-Encoders." *arXiv preprint arXiv:1611.07308*.

Kipf, Thomas N., and Max Welling. 2017. "Semi-Supervised Classification with Graph Convolutional Networks." In *ICLR*.

Klicpera, Johannes, Aleksandar Bojchevski, and Stephan Günnemann. 2019. "Predict Then Propagate: Graph Neural Networks Meet Personalized Pagerank." In *ICLR*.

Kumagai, Atsutoshi, and Tomoharu Iwata. 2019. "Unsupervised Domain Adaptation by Matching Distributions Based on the Maximum Mean Discrepancy via Unilateral Transformations." In *AAAI*.

Lan, Lin, Pinghui Wang, Xuefeng Du, Kaikai Song, Jing Tao, and Xiaohong Guan. 2020. "Node Classification on Graphs with Few-Shot Novel Labels via Meta Transformed Network Embedding." In *NeurIPS*.

Lee, Dong-Hyun,. 2013. "Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks." In *ICML Workshop*.

Li, Da, and Timothy Hospedales. 2020. "Online Meta-Learning for Multi-Source and Semi-Supervised Domain Adaptation." In *European Conference on Computer Vision*, 382–403. Springer.

Li, Qimai, Zhichao Han, and Xiao-Ming Wu. 2018. "Deeper Insights Into Graph Convolutional Networks for Semi-Supervised Learning." In *AAAI*.

Li, Guohao, Matthias Muller, Ali Thabet, and Bernard Ghanem. 2019. "DeepGCNs: Can GCNs go as Deep as CNNs?" In *CVPR*.

Li, Qimai, Xiao-Ming Wu, Han Liu, Xiaotong Zhang, and Zhichao Guan. 2019. "Label Efficient Semi-Supervised Learning via Graph Filtering." In *CVPR*.

Li, Yayong, Jie Yin, and Ling Chen. 2019. "Semi-Supervised Adversarial Active Learning on Attributed Graphs." *arXiv preprint arXiv:1908.08169*.

Li, Yayong, Jie Yin, and Ling Chen. 2021. "Unified Robust Training for Graph Neural Networks Against Label Noise." *arXiv preprint arXiv:2103.03414*.

Li, Yichuan, Kaize Ding, and Kyumin Lee. 2023. "GRENADE: Graph-Centric Language Model for Self-Supervised Representation Learning on Text-Attributed Graphs." In *EMNLP Findings*.

Li, Yichuan, Kaize Ding, Jianling Wang, and Kyumin Lee. 2024. "Empowering Large Language Models for Textual Data Augmentation." *arXiv preprint arXiv:2404.17642*.

Liu, Meng, Hongyang Gao, and Shuiwang Ji. 2020. "Towards Deeper Graph Neural Networks." In *KDD*.

Liu, Yixin, Zhao Li, Shirui Pan, Chen Gong, Chuan Zhou, and George Karypis. 2021. "Anomaly Detection on Attributed Networks via Contrastive Self-Supervised Learning." *IEEE Transactions on Neural Networks and Learning Systems* 33(6): 2378–92.

Liu, Zemin, Yuan Fang, Chenghao Liu, and Steven C. H. Hoi. 2021. "Relative and Absolute Location Embedding for Few-Shot Node Classification on Graph." In *AAAI*.

Liu, Hongrui, Binbin Hu, Xiao Wang, Chuan Shi, Zhiqiang Zhang, and Jun Zhou. 2022. "Confidence May Cheat: Self-Training on Graph Neural Networks Under Distribution Shift." In *The Web Conference*.

Liu, Yixin, Yu Zheng, Daokun Zhang, Hongxu Chen, Hao Peng, and Shirui Pan. 2022. "Towards Unsupervised Deep Graph Structure Learning." In *Proceedings of the ACM Web Conference 2022*, 1392–403.

Liu, Yixin, Kaize Ding, Huan Liu, and Shirui Pan. 2023. "GOOD-D: On Unsupervised Graph Out-of-Distribution Detection." In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 339–47.

Liu, Yixin, Kaize Ding, Qinghua Lu, Fuyi Li, Leo Yu Zhang, and Shirui Pan. 2023. "Towards Self-Interpretable Graph-Level Anomaly Detection." In *Advances in Neural Information Processing Systems*, vol. 36.

Liu, Yixin, Kaize Ding, Jianling Wang, Vincent Lee, Huan Liu, and Shirui Pan. 2023. "Learning Strong Graph Neural Networks With Weak Information." In *KDD*.

Liu, Yixin, Yizhen Zheng, Daokun Zhang, Vincent CS Lee, and Shirui Pan. 2023. "Beyond Smoothing: Unsupervised Graph Representation Learning with Edge Heterophily Discriminating." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 4516–24.

Lv, Xin, Yuxian Gu, Xu Han, Lei Hou, Juanzi Li, and Zhiyuan Liu. 2019. "Adapting Meta Knowledge Graph Information for Multi-Hop Reasoning Over Few-Shot Relations." In *EMNLP*.

Ma, Ning, Jiajun Bu, Jieyu Yang, Zhen Zhang, Chengwei Yao, Zhi Yu, Sheng Zhou, and Xifeng Yan. 2020. "Adaptive-Step Graph Meta-Learner for Few-Shot Graph Classification." In *CIKM*.

NT, Hoang, Choong Jun Jin, and Tsuyoshi Murata. 2019. "Learning Graph Neural Networks with Noisy Labels." *arXiv preprint arXiv:1905.01591*.

Pan, Shirui, Ruiqi Hu, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. 2018. "Adversarially Regularized Graph Autoencoder for Graph Embedding." In *IJCAI*.

Peng, Zhen, Wenbing Huang, Minnan Luo, Qinghua Zheng, Yu Rong, Tingyang Xu, and Junzhou Huang. 2020. "Graph Representation Learning via Graphical Mutual Information Maximization." In *The Web Conference*.

Qiu, Jiezhong, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. 2020. "GCC: Graph Contrastive Coding for Graph Neural Network Pre-Training." In *KDD*.

Rong, Yu, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. "Self-Supervised Graph Transformer on Large-Scale Molecular Data." In *NeurIPS*.

Shchur, Oleksandr, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. 2018. "Pitfalls of Graph Neural Network Evaluation." *arXiv preprint arXiv:1811.05868*.

Shen, Xiao, Quanyu Dai, Sitong Mao, Fu-lai Chung, and Kup-Sze Choi. 2020. "Network Together: Node Classification via Cross-Network Deep Network Embedding." *TNNLS*.

Sheng, Jiawei, Shu Guo, Zhenyu Chen, Juwei Yue, Lihong Wang, Tingwen Liu, and Hongbo Xu. 2020. "Adaptive Attentional Network for Few-Shot Knowledge Graph Completion." In *EMNLP*.

Sheth, Paras, Ahmadreza Mosallanezhad, Kaize Ding, Reepal Shah, John Sabo, Huan Liu, and K Selçuk Candan. 2023. "Streams: Towards Spatio-Temporal Causal Discovery With Reinforcement Learning for Streamflow Rate Prediction." In *CIKM*.

Shi, Yucheng, Yushun Dong, Qiaoyu Tan, Jundong Li, and Ninghao Liu. 2023. "GiGaMAE: Generalizable Graph Masked Autoencoder via Collaborative Latent Space Reconstruction." In *CIKM*.

Sun, Fan-Yun, Jordan Hoffmann, Vikas Verma, and Jian Tang. 2019. "Infograph: Unsupervised and Semi-Supervised Graph-Level Representation Learning via Mutual Information Maximization." In *ICLR*.

Sun, Ke, Zhouchen Lin, and Zhanxing Zhu. 2020. "Multi-Stage Self-Supervised Learning for Graph Convolutional Networks on Graphs with Few Labeled Nodes." In *AAAI*.

Tan, Zhen, Kaize Ding, Ruocheng Guo, and Huan Liu. 2022. "Graph Few-Shot Class-Incremental Learning." In *WSDM*.

Tian, Yijun, Kaiwen Dong, Chunhui Zhang, Chuxu Zhang, and Nitesh V Chawla. 2023. "Heterogeneous Graph Masked Autoencoders." In *AAAI*.

Tian, Yijun, Chuxu Zhang, Ziyi Kou, Zheyuan Liu, Xiangliang Zhang, and Nitesh V. Chawla. 2024. "UGMAE: A Unified Framework for Graph Masked Autoencoders." *arXiv preprint arXiv:2402.08023*.

van den Oord, Aaron, Yazhe Li, and Oriol Vinyals. 2018. "Representation Learning with Contrastive Predictive Coding." *arXiv preprint arXiv:1807.03748*.

Velickovic, Petar, William Fedus, and William L Hamilton. 2019. "Deep Graph Infomax." In *ICML*.

Wang, Ning, Minnan Luo, Kaize Ding, Lingling Zhang, Jundong Li, and Qinghua Zheng. 2020. "Graph Few-Shot Learning With Attribute Matching." In *CIKM*.

Wang, Jianling, Kaize Ding, and James Caverlee. 2021. "Sequential Recommendation for Cold-Start Users with Meta Transitional Learning." In *SIGIR*.

Wang, Ping, Khushbu Agarwal, Colby Ham, Sutanay Choudhury, and Chandan K Reddy. 2021. "Self-Supervised Learning of Contextual Embeddings for Link Prediction in Heterogeneous Networks." In *The Web Conference*.

Wang, Song, Xiao Huang, Chen Chen, Liang Wu, and Jundong Li. 2021. "Reform: Error-Aware Few-Shot Knowledge Graph Completion." In *CIKM*.

Wang, Song, Kaize Ding, Chuxu Zhang, Chen Chen, and Jundong Li. 2022. "Task-Adaptive Few-Shot Node Classification." In *KDD*.

Wang, Song, Yushun Dong, Kaize Ding, Chen Chen, and Jundong Li. 2023. "Few-Shot Node Classification with Extremely Weak Supervision." In *WSDM*.

Wang, Jianling, Haokai Lu, James Caverlee, Ed H. Chi, and Minmin Chen. 2024. "Large Language Models as Data Augmenters for Cold-Start Item Recommendation." In *Companion Proceedings of the ACM on Web Conference 2024*.

Wu, Man, Shirui Pan, Chuan Zhou, Xiaojun Chang, and Xingquan Zhu. 2020. "Unsupervised Domain Adaptive Graph Convolutional Networks." In *The Web Conference*.

Wu, Zonghan, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2020. "A Comprehensive Survey on Graph Neural Networks." *TNNLS*.

Xiong, Wenhan, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2018. "One-Shot Relational Learning for Knowledge Graphs." In *EMNLP*.

Xu, Zhe, Kaize Ding, Yu-Xiong Wang, Huan Liu, and Hanghang Tong. 2022. "Generalized Few-Shot Node Classification." In *ICDM*.

Yao, Huaxiu, Chuxu Zhang, Ying Wei, Meng Jiang, Suhang Wang, Junzhou Huang, Nitesh Chawla, and Zhenhui Li. 2020. "Graph Few-Shot Learning via Knowledge Transfer." In *AAAI*.

Yarowsky, David. 1995. "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods." In *ACL*.

You, Yuning, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. "Graph Contrastive Learning With Augmentations." In *NeurIPS*.

You, Yuning, Tianlong Chen, Zhangyang Wang, and Yang Shen. 2020. "When Does Self-Supervision Help Graph Convolutional Networks?" In *ICML*.

Zhang, Chuxu, Huaxiu Yao, Chao Huang, Meng Jiang, Zhenhui Li, and Nitesh V. Chawla. 2020. "Few-Shot Knowledge Graph Completion." In *AAAI*.

Zhang, Xiaowen, Yuntao Du, Rongbiao Xie, and Chongjun Wang. 2021. "Adversarial Separation Network for Cross-Network Node Classification." In *CIKM*.

Zhang, Qiannan, Xiaodong Wu, Qiang Yang, Chuxu Zhang, and Xiangliang Zhang. 2022a. "Few-Shot Heterogeneous Graph Learning via Cross-Domain Knowledge Transfer." In *KDD*.

Zhang, Qiannan, Xiaodong Wu, Qiang Yang, Chuxu Zhang, and Xiangliang Zhang. 2022b. "HG-Meta: Graph Meta-Learning Over Heterogeneous Graphs." In *SDM*. SIAM.

Zhang, Yiming, Yiyue Qian, Yanfang Ye, and Chuxu Zhang. 2022. "Adapting Distilled Knowledge for Few-Shot Relation Reasoning over Knowledge Graphs." In *SDM*. SIAM.

Zhang, Qiannan, Shichao Pei, Qiang Yang, Chuxu Zhang, Nitesh V. Chawla, and Xiangliang Zhang. 2023. "Cross-Domain Few-Shot Graph Classification with a Reinforced Task Coordinator." In *AAAI*.

Zhao, Lingxiao, and Leman Akoglu. 2019. "PairNorm: Tackling Oversmoothing in GNNs." In *ICLR*.

Zhou, Fan, Chengtai Cao, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Ji Geng. 2019. "Meta-GNN: On Few-Shot Node Classification in Graph Meta-Learning." In *CIKM*.

Zhou, Ziang, Jieming Shi, Shengzhong Zhang, Zengfeng Huang, and Qing Li. 2019. "Effective Semi-Supervised Node Classification on Few-Labeled Graph Data." *arXiv preprint arXiv:1910.02684*.

Zhu, Qikui, Bo Du, and Pingkun Yan. 2020. "Self-Supervised Training of Graph Convolutional Networks." *arXiv preprint arXiv:2006.02380*.

Zhu, Yanqiao, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2020. "Deep Graph Contrastive Representation Learning." In *ICML Workshop*.

Zhu, Yanqiao, Yichen Xu, Feng Yu, Shu Wu, and Liang Wang. 2021. "CAGNN: Cluster-Aware Graph Neural Networks for Unsupervised Graph Representation Learning." In *TIST*.

Zhuang, Jun, and Mohammad AI Hasan. 2020. "Anti-Perturbation of Online Social Networks by Graph Label Transition." *arXiv preprint arXiv:2010.14121*.

Zügner, Daniel, Amir Akbarnejad, and Stephan Günnemann. 2018. "Adversarial Attacks on Neural Networks for Graph Data." In *KDD*.

How to cite this article: Ding, K., Y. Liu, C. Zhang, and J. Wang. 2024. "Data-efficient graph learning: Problems, progress, and prospects." *AI Magazine* 1–12. https://doi.org/10.1002/aaai.12200

AUTHOR BIOGRAPHIES

Kaize Ding is an Assistant Professor in the Department of Statistics and Data Science at Northwestern University. His research interests are generally in data mining, machine learning, and large foundation models, with a special focus on reliable and efficient AI. His work has been published in top-tier venues (e.g., AAAI, EMNLP, KDD, NeurIPS, and TheWebConf), and has been recognized with several prestigious awards, including the Amazon Research Award, the AAAI New Faculty Highlights, SDM Best Posters Award, Best Paper Award at the Trustworthy Learning on Graphs workshop, and so forth.

Yixin Liu is currently a postdoctoral research fellow at Griffith University. He received his Ph.D. in Artificial Intelligence (AI) from Monash University, Australia, and his Bachelor's and Master's degrees from Beihang University, China. His research focuses on data mining, machine learning, graph analytics, and anomaly detection. To date, he has published more than 20 research papers in top-tier journals and conferences, including IEEE TKDE, IEEE TNNLS, NeurIPS, KDD, AAAI, and The Web Conference. He was awarded the Google Ph.D. Fellowship in 2022.

Chuxu Zhang is an Associate Professor in the Department of Computer Science and Engineering at the University of Connecticut. His research lies in the intersection of artificial intelligence, graph machine learning, large language models, and societal applications. Recently, he has focused on developing resource-efficient and safe learning models and algorithms on graph, language, and multimodality data. He has more than 100 papers published in major conferences such as ICML, NeurIPS, ICLR, KDD, WWW, AAAI, IJCAI, and so forth. He is the recipient of multiple awards including the NSF CAREER Award (2024), the Frontiers of Science Award (2024), and the Best Paper (Candidate) Awards in CIKM 2021, WWW 2019, and WAIM 2016.

Jianling Wang is a Senior Research Scientist working at Google DeeMind. She obtained her Ph.D. degree from the Department of Computer Science and Engineering at Texas A&M University, advised by Prof. James Caverlee. Her research interests generally include data mining and machine learning, with a particular focus on recommendation systems and graph neural networks.