

# Tackling Negative Transfer on Graphs

Zehong Wang<sup>1</sup>, Zheyuan Zhang<sup>1</sup>, Chuxu Zhang<sup>2</sup> and Yanfang Ye<sup>1\*</sup>

<sup>1</sup>University of Notre Dame, Indiana, USA

<sup>2</sup>Brandeis University, Massachusetts, USA

{zwang43, zzhang42, yye7}@nd.edu, chuxuzhang@brandeis.edu

## Abstract

Transfer learning aims to boost the learning on the target task leveraging knowledge learned from other relevant tasks. However, when the source and target are not closely related, the learning performance may be adversely affected, a phenomenon known as negative transfer. In this paper, we investigate the negative transfer in graph transfer learning, which is important yet underexplored. We reveal that, unlike image or text, negative transfer commonly occurs in graph-structured data, even when source and target graphs share semantic similarities. Specifically, we identify that structural differences significantly amplify the dissimilarities in the node embeddings across graphs. To mitigate this, we bring a new insight: for semantically similar graphs, although structural differences lead to significant distribution shift in node embeddings, their impact on subgraph embeddings could be marginal. Building on this insight, we introduce two effective yet elegant methods, Subgraph Pooling (SP) and Subgraph Pooling++ (SP++), that transfer subgraph-level knowledge across graphs. We theoretically analyze the role of SP in reducing graph discrepancy and conduct extensive experiments to evaluate its superiority under various settings. Our code and datasets are available at: <https://github.com/Zehong-Wang/Subgraph-Pooling>.

## 1 Introduction

Graph Neural Networks (GNNs) are widely employed for graph mining tasks across various fields [Gaudelet *et al.*, 2021; Kipf and Welling, 2017; He *et al.*, 2020]. Despite their remarkable success in graph-structured datasets, these methods exhibit limitations in label sparse scenarios [Dai *et al.*, 2022a]. This restricts the applications of GNNs in real-world datasets where label acquisition is challenging or impractical. To address the issue, transfer learning [Zhuang *et al.*, 2020] emerges as a solution, which aims to transfer knowledge from a label-rich source graph to a label-sparse target graph through fine-tuning or prompting [Sun *et al.*, 2023].

However, the success of transfer learning is not always guaranteed [Wang *et al.*, 2019; Zhang *et al.*, 2022]. If the source and target lack sufficient similarity, transferring knowledge from such weakly related source may impair performance on the target. This phenomenon is known as negative transfer, which was initially analyzed by [Wang *et al.*, 2019]. By interpreting transfer learning as a specific generalization problem, it demonstrated that negative transfer is caused by the divergence between joint distributions of the source and target. To this end, researchers employed adversarial learning [Wu *et al.*, 2020], causal learning [Chen *et al.*, 2022], or domain regularizer [You *et al.*, 2023] to develop domain-invariant encoders, which reduce the distribution shift between the source and target.

In this paper, we conduct a systematic analysis of the negative transfer issue in GNNs, which is a lack of existing works. Our observations indicate that in graph datasets, negative transfer often occurs even when the source and target are semantically similar. This is in contrast to image and text datasets, where similar sources typically enhance the performance on targets [Zhuang *et al.*, 2020]. We consider this derives from the differences in graph structures between the source and target, which may lead to significant distribution shifts on node embeddings. For example, in financial transaction networks collected over different time intervals, transaction patterns can vary markedly due to the impact of social events or policy changes. These evolving patterns notably change the local structure of users, leading to a substantial divergence in user embeddings. To address this challenge, we introduce two straightforward yet effective methods called Subgraph Pooling (SP) and Subgraph Pooling++ (SP++) to reduce the discrepancy between graphs. Our major contributions are summarized as follows:

- **Negative Transfer in GNNs.** We systematically analyze the negative transfer in GNNs. We find that the structural difference between the source and target graphs intensifies distribution shifts on node embeddings, as the aggregation process of GNNs is highly sensitive to perturbations in graph structures. To address this issue, we present a novel insight: for semantically similar graphs, although structural differences lead to significant distribution shift in node embeddings, their impact on subgraph embeddings could be marginal.

\*Corresponding Author.

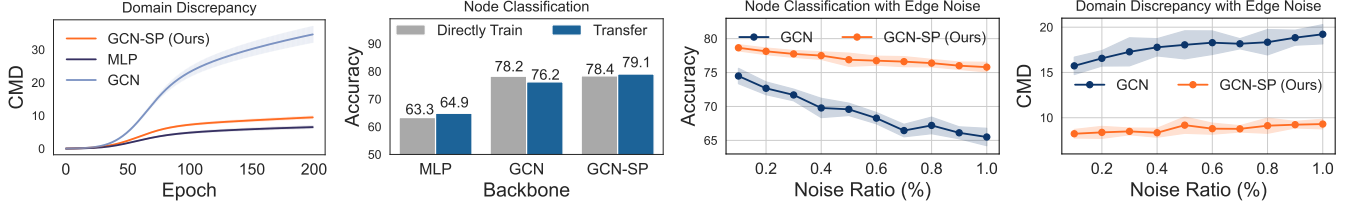


Figure 1: Structural differences between the source (DBLP) and target (ACM) amplify the distribution shift on nodes embeddings. **Left:** We illustrate the discrepancy (CMD value) between node embeddings of the source and target during pre-training, and compare the performance of direct training on the target (gray) and transferring knowledge from the source to the target (blue). A large discrepancy results in negative transfer. **Right:** We introduce structural noise in the target graph through random edge permutation. Even minor permutations can enlarge the discrepancy (and thus aggravate negative transfer) in vanilla GCN, yet our method effectively mitigates the issue.

- **Subgraph Pooling to Tackle Negative Transfer.** Building upon this insight, we introduce plug-and-play modules Subgraph Pooling (SP) and Subgraph Pooling++ (SP++) to mitigate the negative transfer. The key idea is to transfer subgraph information across source and target to prevent the distribution shift. Notably, we provide a comprehensive theoretical analysis to clarify the operational processes behind Subgraph Pooling.
- **Generality and Effectiveness.** Subgraph Pooling is straightforward to implement and introduces no additional parameters. It involves simple sampling and pooling operations, making it easily applicable to any GNN backbone. We conduct extensive experiments to demonstrate that our method can significantly surpass existing baselines under multiple transfer learning settings.

## 2 Negative Transfer in GNNs

### 2.1 Preliminary of Graph Transfer Learning

Semi-supervised graph learning is a common setting in real-world applications [Kipf and Welling, 2017]. In this work, we study negative transfer in semi-supervised graph transfer learning for node classification, while our analysis is also applicable for other transfer learning settings [Wu *et al.*, 2020]. Semi-supervised transfer learning focuses on transferring knowledge from a label-rich source  $\mathcal{D}_S$  to a label-sparse target  $\mathcal{D}_T$ . We represent the joint distribution over the source and target as  $P_S(\mathcal{X}, \mathcal{Y})$  and  $P_T(\mathcal{X}, \mathcal{Y})$ , respectively, where  $\mathcal{X}$  indicates the random input space and  $\mathcal{Y}$  is the output space. The labeled training instances are sampled as  $\mathcal{D}_S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s} \sim P_S(\mathcal{X}, \mathcal{Y})$  and  $\mathcal{D}_T^L = \{(x_i^t, y_i^t)\}_{i=1}^{n_t^L} \sim P_T(\mathcal{X}, \mathcal{Y})$ , while the unlabeled instances are  $\mathcal{D}_T^U = \{(x_i^t)\}_{i=1}^{n_t^U} \sim P_T(\mathcal{X})$ , combining to form  $\mathcal{D}_T = (\mathcal{D}_T^L, \mathcal{D}_T^U)$ . The objective is to develop a hypothesis function  $h : \mathcal{X} \rightarrow \mathcal{Y}$  that minimizes the empirical risk on the target  $R_T(h) = \mathbf{Pr}_{(x,y) \sim \mathcal{D}_T}(h(x) \neq y)$ .

Considering graph-structured data, a graph is represented as  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V} = \{v_1, \dots, v_n\}$  is the node set and  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$  is the edge set. Each node  $i \in \mathcal{V}$  is associated with node attributes  $\mathbf{x}_i \in \mathbb{R}^d$  and a class  $y_i \in \{1, \dots, C\}$ , with  $C$  being the total number of classes. Additionally, each graph has an adjacency matrix  $\mathbf{A} \in \{0, 1\}^{n \times n}$ , where  $\mathbf{A}_{ij} = 1$  iff  $(i, j) \in \mathcal{E}$ , otherwise  $\mathbf{A}_{ij} = 0$ . In the semi-supervised transfer setting, we have the source graph  $\mathcal{G}^s = (\mathcal{V}^s, \mathcal{E}^s)$  and target graph  $\mathcal{G}^t = (\mathcal{V}^t, \mathcal{E}^t)$ . For simplicity, we assume these graphs share the same feature space  $\mathbf{X}^s \in \mathbb{R}^{n_s \times d}$  and  $\mathbf{X}^t \in \mathbb{R}^{n_t \times d}$ ,

as well as a common label space  $y^s, y^t \in \{1, \dots, C\}$ . We employ a GNN backbone  $f(\cdot)$  to encode nodes into embeddings  $\mathbf{Z}^s, \mathbf{Z}^t$  and then use a classifier  $g(\cdot)$  for predictions. The joint distributions over the source and target graphs are  $P_S(\mathcal{Z}, \mathcal{Y})$  and  $P_T(\mathcal{Z}, \mathcal{Y})$ , where  $\mathcal{Z}$  denotes the node embedding space.

**Definition 1** (Semi-supervised Graph Transfer Learning). *The aim is to transfer knowledge from a label-rich source graph  $\mathcal{G}^s$  to a semantically similar label-sparse target graph  $\mathcal{G}^t$  for enhancing node classification performance. The joint distributions  $P(\mathcal{Z}, \mathcal{Y})$  over the source and target are different, where  $P_S(\mathcal{Y}|\mathcal{Z}) = P_T(\mathcal{Y}|\mathcal{Z})$  and  $P_S(\mathcal{Z}) \neq P_T(\mathcal{Z})$ .*

While the conditional distributions  $P(\mathcal{Y}|\mathcal{Z})$  over the source and target are identical, their marginal distributions  $P(\mathcal{Z})$  are different. To quantify this discrepancy, researchers utilize metrics such as Maximum Mean Discrepancy (MMD) [Gretton *et al.*, 2006], Center Moment Discrepancy (CMD) [Zellinger *et al.*, 2017], or Wasserstein Distance [Zhu *et al.*, 2023], to measure node similarities in complex spaces. We use CMD due to its computational efficiency:

$$d_{CMD} = \frac{1}{|n_s - n_t|} \|\mathbb{E}(\mathbf{Z}^s) - \mathbb{E}(\mathbf{Z}^t)\|_2 + \sum_{k=2}^K \frac{1}{|n_s - n_t|^k} \|c_k(\mathbf{Z}^s) - c_k(\mathbf{Z}^t)\|_2, \quad (1)$$

where  $c_k(\cdot)$  denotes the  $k$ -th order central moment (with  $K = 3$ ). A high CMD value indicates a considerable shift in marginal distributions between the source and target. This shift essentially results in a divergence between joint distributions  $P_S(\mathcal{Z}, \mathcal{Y})$  and  $P_T(\mathcal{Z}, \mathcal{Y})$ , which may hinder or even degrade performance on the target [Wang *et al.*, 2019].

### 2.2 Why Negative Transfer Happens?

Although negative transfer typically occurs between two weakly related domains, it remains a prevalent issue in GNNs, even when the source and target are semantically similar. This issue is attributed to the sensitivity of GNNs to graph structures. Specifically, differences in structural distribution between the source and target can lead to distinct marginal distributions  $P_S(\mathcal{Z})$  and  $P_T(\mathcal{Z})$ , thereby diverging the corresponding joint distributions. To support this claim, we analyze the influence of structures on graph discrepancy (CMD value) and transfer learning performance, as illustrated in Figure 1 (Left). Particularly, when graph structure is not considered (using MLP), the discrepancy remains relatively low, ensuring the performance gain of transfer learning. Conversely,

incorporating structural information through GNNs can increase the discrepancy, resulting in negative transfer.

Based on the observations, we consider that GNNs may project semantically similar graphs into distinct spaces, unless their structures are very similar. To further reveal the phenomenon, we delve into the aggregation process of GNNs. For any GNN architecture, each node is associated with a computational tree, through which messages are passed and aggregated from leaves to the root. Only closely aligned structures can lead to similar computational tree distributions across graphs, thereby ensuring closely matched node embeddings. However, this requirement is often impractical in many graph datasets. Even minor perturbations in the graph structure can dramatically alter the computational tree, either by dropping critical branches or by introducing noisy connections. Furthermore, a single perturbation can impact the computational trees of multiple nodes, thus altering the computational tree distributions across the graph. We demonstrate the impact of structure perturbations in Figure 1 (Right). In conclusion, structural differences between the source and target result in distinct computational tree distributions, culminating in a significant distribution shift in node embeddings.

### 2.3 Analysing The Impact of Structure

The above analysis suggests that mitigating the impact of graph structure on node embeddings is crucial for alleviating negative transfer. Existing works implicitly or explicitly handle this issue. For instance, some researchers utilize adversarial learning [Wu *et al.*, 2020; Dai *et al.*, 2022b] or domain regularizers [You *et al.*, 2023] to develop domain-invariant GNN encoders, which consistently project graphs with different structures into a unified embedding space. However, these methods lack generalizability to new, unseen graphs and are sensitive to structural perturbations. Alternatively, another line of work employs causal learning [Wu *et al.*, 2022b; Chen *et al.*, 2022] or augmentation [Liu *et al.*, 2022] to train encoders robust to structural distribution shift. Yet, these methods essentially generate additional training graphs to enhance robustness against minor structural perturbations, instead of considering the fundamental nature of graph structures. Unlike these two approaches, we present a novel insight to solve the issue: for semantically similar graphs, although structural differences lead to significant distribution shift in node embeddings, their impact on subgraph embeddings could be marginal. To better describe this phenomenon, we introduce node-level and subgraph-level discrepancy as metrics to evaluate the influence of graph structures.

**Definition 2** (Node-level Discrepancy). *For nodes  $u \in \mathcal{V}^s$  in source graph and  $v \in \mathcal{V}^t$  in target graph, we have*

$$\mathbb{E}_{u \in \mathcal{V}^s, v \in \mathcal{V}^t} \frac{\mathbf{z}_u^T \mathbf{z}_v}{\mathbf{z}_u^T \mathbf{z}_v} \geq \lambda, \quad (2)$$

where  $\lambda$  denotes the node-level discrepancy.

**Definition 3** (Subgraph-level Discrepancy). *For node  $u \in \mathcal{V}^s$  with surrounding subgraph  $\mathcal{S}_u^s = (\mathcal{V}_u^s, \mathcal{E}_u^s)$  and node  $v \in \mathcal{V}^t$  with surrounding subgraph  $\mathcal{S}_v^t = (\mathcal{V}_v^t, \mathcal{E}_v^t)$ , we have*

$$\mathbb{E}_{u \in \mathcal{V}^s, v \in \mathcal{V}^t} \left\| \frac{1}{n_u^s + 1} \sum_{i \in \mathcal{V}_u^s} \mathbf{z}_i - \frac{1}{m_v^t + 1} \sum_{j \in \mathcal{V}_v^t} \mathbf{z}_j \right\| \leq \epsilon \quad (3)$$

	ACM $\rightarrow$ DBLP	DBLP $\rightarrow$ ACM	Arxiv T1	Arxiv T3
$\lambda$	2.413	2.353	2.134	2.683
$\epsilon$ ( $k$ -hop)	0.212	0.380	0.191	0.203
$\epsilon$ (RW)	0.166	0.322	0.184	0.212

Table 1: Although node-level discrepancy ( $\lambda$ ) between source and target is high, the subgraph-level discrepancy ( $\epsilon$ ) remains low.  $k$ -hop and RW (Random Walk) indicate two subgraph sampling methods.

where  $n_u^s = |\mathcal{V}_u^s|$ ,  $m_v^t = |\mathcal{V}_v^t|$ , and  $\epsilon$  denotes the subgraph-level discrepancy.

Intuitively, a high  $\lambda$  value suggests a significant distinction in node embeddings between the source and target, which potentially leads to negative transfer. On the other hand, a low value of  $\epsilon$  indicates similar subgraph embeddings across the source and target, which potentially prevents negative transfer. We demonstrate the impact of graph structures on these two measurements using real-world datasets, as detailed in Table 1. Although the node embeddings are distinct between the source and target owing to the impact of structural differences (as indicated by high  $\lambda$ ), the subgraph embeddings remain similar across graphs (as indicated by low  $\epsilon$ ). Drawing on these insights, we propose to directly transfer the subgraph information across graphs to enhance transfer learning performance by mitigating the impact of graph structures.

## 3 Overcoming Negative Transfer

### 3.1 Subgraph Pooling

We start by presenting the final objective for node-level graph transfer learning. The goal is to minimize the empirical risk (loss) on the target distribution, expressed as

$$\min \mathbb{E}_{(z,y) \sim P_T(\mathcal{Z}, \mathcal{Y})} [\mathcal{L}(g(z), y)], \quad (4)$$

where  $P_T(\mathcal{Z}, \mathcal{Y})$  represents the joint distribution over the target graph,  $z$  denotes the node embeddings encoded by a GNN backbone  $f_\Theta(\cdot)$  with parameters  $\Theta$ , and  $g(\cdot)$  is a linear classifier. These two components are represented as

$$\hat{\mathbf{Y}} = g(\mathbf{Z}), \mathbf{Z} = f(\mathbf{A}, \mathbf{X}, \Theta). \quad (5)$$

However, owing to the scarcity of labels, we cannot describe the real joint distribution  $P_T(\mathcal{Z}, \mathcal{Y})$  [Wenzel *et al.*, 2022]. Consequently, directly optimizing Eq. 4 on the target graph may lead to overfitting [Mallinar *et al.*, 2022]. To address this, we pre-train the backbone and classifier on a similar, label-rich source graph to achieve a suitable initialization, aiming to project the source and target into similar embedding spaces. Nevertheless, training such encoder is challenging, as structural differences enlarge the disparity between the marginal distributions  $P_S(\mathcal{Z})$  and  $P_T(\mathcal{Z})$ . To overcome the limitation, we introduce Subgraph Pooling (SP), a plug-and-play method that leverages subgraph information to diminish the discrepancy between the source and target. This approach is based on the following assumption.

**Assumption 1.** *For two semantically similar graphs, the subgraph-level discrepancy  $\epsilon$  is small enough.*

**Remark 1.** *We empirically validate the assumption in Table 1 and consider it matches many real-world graphs. For example, considering papers from two citation networks, if they*

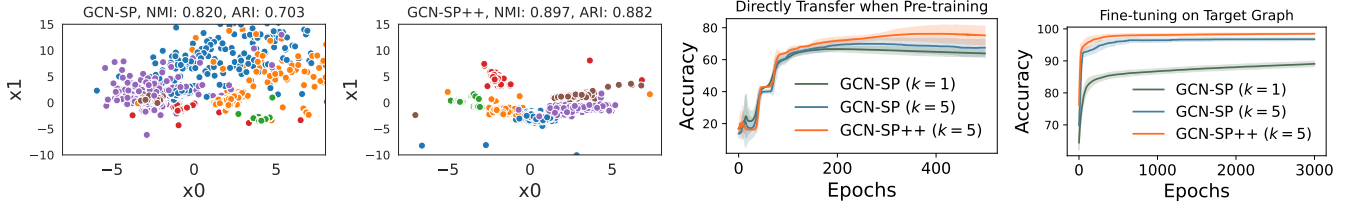


Figure 2: Subgraph Pooling++ (SP++) mitigates the risk of over-smoothing derived from a large pooling kernel. We conduct transfer learning from ACM to DBLP. **Left:** Illustration of the subgraph embeddings on the target graph with  $k = 5$ , where SP++ leveraging RW sampler has a clearer boundary. **Right:** Transfer learning performance during pre-training and fine-tuning, where SP++ achieves better.

share the same research field, they tend to have similar local structures, e.g., neighbors, since papers within the same domain often reference a core set of foundational works. Additionally, this pattern extends to social networks, where individuals with similar interests or professional backgrounds are likely to have comparable connection patterns, reflecting shared community norms or communication channels.

The key idea of Subgraph Pooling is to transfer subgraph-level knowledge across graphs. The method is applicable for arbitrary GNNs by adding a subgraph pooling layer at the end of backbone. Specifically, in the SP layer, we first sample the subgraphs around nodes and then perform pooling to generate subgraph embeddings for each node. The choice of sampling and pooling functions can be arbitrary. Here we consider a straightforward sampling method, defined as the  $k$ -hop subgraph around each node:

$$\mathcal{N}_s(i) = \text{Sample}_{k\text{-hop}}(\mathcal{G}, i). \quad (6)$$

Subsequently, we pool the subgraph for each node:

$$\mathbf{h}_i = \frac{1}{|\mathcal{N}_s(i)| + 1} \sum_{j \in \mathcal{N}_s(i) \cup i} w_{ij} \mathbf{z}_j. \quad (7)$$

where  $\mathbf{h}_i \in \mathbf{H}$  represents the subgraph embeddings (the new embeddings for each node), utilized in training the classifier  $g(\cdot)$ .  $w_{ij}$  denotes the pooling weight, which can be either learnable or fixed. Empirically, the MEAN pooling function is effective enough to achieve desirable transfer performance.

The integration of the SP layer into GNN architectures does not substantially increase time complexity. We have three-fold considerations. Firstly, the SP layer functions as a non-parametric GNN layer, thus imposing no additional burden on model optimization and enjoying the computational efficiency of existing GNN libraries [Fey and Lenssen, 2019]. Secondly, the sampling operation relies solely on the graph structure and can be performed in pre-processing. Finally, our empirical observations indicate that sampling low-order neighborhoods ( $k = 1, 2$ ) is sufficient for achieving optimal transfer learning performance, which ensures computational efficiency in both sampling and pooling.

Our SP layer leverages subgraph information to reduce the discrepancy (CMD value) between node embeddings in the source and target, thereby enhancing transfer learning performance, which is illustrated in Figure 1 (Left). Furthermore, the use of subgraph information also reduces the sensitivity to structural perturbations, as evidenced in Figure 1 (Right). We also provide a theoretical analysis to explain how subgraph information reduces the graph discrepancy.

**Theorem 1.** For node  $u \in \mathcal{V}^s$  in the source graph and  $v \in \mathcal{V}^t$  in the target graph, considering the MEAN pooling function, the subgraph embeddings are  $\mathbf{h}_u = \frac{\mathbf{z}_u + \sum_{i \in \mathcal{N}_s(u)} \mathbf{z}_i}{n+1}$ ,  $\mathbf{h}_v = \frac{\mathbf{z}_v + \sum_{j \in \mathcal{N}_s(v)} \mathbf{z}_j}{m+1}$  where  $n = |\mathcal{N}_s(u)|$ ,  $m = |\mathcal{N}_s(v)|$ . We have

$$\|\mathbf{h}_u - \mathbf{h}_v\| \leq \|\mathbf{z}_u - \mathbf{z}_v\| - \Delta, \quad (8)$$

where  $\Delta = \frac{(n\|\mathbf{z}_u - \mathbf{z}_v\| - \frac{m-n}{m+1}\|\mathbf{z}_v\|)}{n+1}$  denotes the discrepancy margin.

**Corollary 1.** If either of the following conditions is satisfied ( $|\mathcal{N}_s(u)| \geq |\mathcal{N}_s(v)|$  or  $|\mathcal{N}_s(u)|$  is sufficiently large), the inequality  $\|\mathbf{h}_u - \mathbf{h}_v\| \leq \|\mathbf{z}_u - \mathbf{z}_v\|$  strictly holds.

**Corollary 2.** If the following condition is satisfied ( $|\mathcal{N}_s(u)| < |\mathcal{N}_s(v)|$ ), the inequality  $\|\mathbf{h}_u - \mathbf{h}_v\| \leq \|\mathbf{z}_u - \mathbf{z}_v\|$  strictly holds when  $\lambda \geq 2$ , even in extreme case where  $|\mathcal{N}_s(u)| \rightarrow 0$  and  $|\mathcal{N}_s(v)| \rightarrow \infty$ .

*Proof.* All proofs are presented in Appendix A.

**Remark 2.** Based on the theoretical results, we can readily prove that the distance of  $k$ -th order central moment between two graphs can be reduced, i.e.,  $\|c_k(\mathbf{H}^s) - c_k(\mathbf{H}^t)\| \leq \|c_k(\mathbf{Z}^s) - c_k(\mathbf{Z}^t)\|$ . This implies that the SP layer indeed decreases the discrepancy (CMD value) between two graphs.

### 3.2 Subgraph Pooling++

The performance of Subgraph Pooling highly depends on the choice of subgraph sampling function. Employing an inappropriate sampling function can impair the distinguishability of the learned embeddings. For instance, the basic  $k$ -hop sampler might cause distinct nodes to share an identical subgraph, collapsing the embeddings into a single point. This can result in the potential over-smoothing problem [Zhao and Akoglu, 2020; Keriven, 2022; Huang *et al.*, 2023] since these nodes may have different labels.

To address the limitation, we propose an advanced method called Subgraph Pooling++ (SP++) that leverages Random Walk (RW) [Huang *et al.*, 2021] to sample subgraphs. We use the same hyper-parameter  $k$  to define the maximum walk length, restricting the sampling process within  $k$ -hop subgraphs. The RW sampler is defined as

$$\mathcal{N}_r(i) = \text{Sample}_{\text{RW}}(\mathcal{G}, i). \quad (9)$$

The RW sampler mitigates the over-smoothing by imposing nodes to share different subgraphs. Inherently,  $k$ -hop sampler

Backbone	Model	DBLP $\rightarrow$ ACM				ACM $\rightarrow$ DBLP				Rank
		$q = 0.1\%$	$q = 0.5\%$	$q = 1\%$	$q = 10\%$	$q = 0.1\%$	$q = 0.5\%$	$q = 1\%$	$q = 10\%$	
GCN	No Transfer	48.44 $\pm$ 2.50	62.70 $\pm$ 2.91	68.63 $\pm$ 2.51	78.23 $\pm$ 0.41	39.12 $\pm$ 6.52	92.14 $\pm$ 1.77	95.61 $\pm$ 1.06	97.19 $\pm$ 0.18	4.8
	ERM	73.36 $\pm$ 0.88	74.08 $\pm$ 0.67	75.18 $\pm$ 0.54	76.19 $\pm$ 0.92	70.52 $\pm$ 0.91	80.88 $\pm$ 1.48	81.76 $\pm$ 0.74	83.07 $\pm$ 0.90	4.6
	Multi-task	70.10 $\pm$ 5.50	70.96 $\pm$ 7.94	74.35 $\pm$ 2.87	76.32 $\pm$ 2.79	74.51 $\pm$ 0.58	80.21 $\pm$ 0.97	80.24 $\pm$ 1.03	84.56 $\pm$ 1.04	5.3
	EERM	56.94 $\pm$ 6.49	59.39 $\pm$ 6.33	64.32 $\pm$ 6.93	67.96 $\pm$ 7.30	59.29 $\pm$ 6.23	70.10 $\pm$ 5.39	77.39 $\pm$ 3.05	90.03 $\pm$ 5.30	6.5
	GTrans	72.20 $\pm$ 0.19	73.70 $\pm$ 1.93	75.10 $\pm$ 0.11	77.53 $\pm$ 1.94	80.97 $\pm$ 1.84	88.84 $\pm$ 1.29	94.00 $\pm$ 3.09	95.19 $\pm$ 0.69	3.9
	GNN-SP	74.51 $\pm$ 1.23	75.63 $\pm$ 1.61	75.64 $\pm$ 0.89	79.18 $\pm$ 0.40	<b>84.11 <math>\pm</math> 2.00</b>	<b>96.40 <math>\pm</math> 1.65</b>	96.41 $\pm$ 1.52	97.54 $\pm$ 1.01	1.8
	GNN-SP++	<b>74.68 <math>\pm</math> 1.07</b>	<b>76.41 <math>\pm</math> 1.83</b>	<b>77.06 <math>\pm</math> 0.90</b>	<b>79.20 <math>\pm</math> 0.23</b>	81.69 $\pm$ 5.96	95.42 $\pm$ 2.74	<b>96.66 <math>\pm</math> 1.47</b>	<b>98.20 <math>\pm</math> 0.54</b>	1.3
GAT	No Transfer	48.11 $\pm$ 2.89	62.52 $\pm$ 2.50	68.50 $\pm$ 2.13	78.32 $\pm$ 0.32	40.30 $\pm$ 0.11	94.78 $\pm$ 2.24	96.68 $\pm$ 1.33	97.31 $\pm$ 0.28	4.9
	ERM	68.48 $\pm$ 2.91	72.60 $\pm$ 2.15	72.67 $\pm$ 1.65	73.10 $\pm$ 2.02	<u>75.38 <math>\pm</math> 2.32</u>	85.76 $\pm$ 1.82	86.35 $\pm$ 1.42	87.99 $\pm$ 1.76	4.3
	Multi-task	67.72 $\pm$ 4.69	69.37 $\pm$ 2.94	70.72 $\pm$ 3.19	73.64 $\pm$ 3.80	71.34 $\pm$ 2.16	81.91 $\pm$ 2.73	81.74 $\pm$ 2.78	85.10 $\pm$ 2.99	6.0
	EERM	67.49 $\pm$ 2.89	69.69 $\pm$ 1.85	71.89 $\pm$ 2.48	74.48 $\pm$ 2.95	72.15 $\pm$ 2.91	79.80 $\pm$ 3.20	82.11 $\pm$ 1.34	89.48 $\pm$ 2.68	5.4
	GTrans	67.39 $\pm$ 2.09	71.36 $\pm$ 0.49	72.99 $\pm$ 1.34	75.36 $\pm$ 2.56	74.83 $\pm$ 1.92	85.03 $\pm$ 1.39	93.15 $\pm$ 1.54	95.59 $\pm$ 0.53	4.3
	GNN-SP	70.85 $\pm$ 2.83	75.98 $\pm$ 1.03	76.56 $\pm$ 0.72	78.56 $\pm$ 0.67	<b>77.43 <math>\pm</math> 6.47</b>	95.44 $\pm$ 1.36	96.55 $\pm$ 0.96	97.63 $\pm$ 0.80	2.0
	GNN-SP++	<b>71.88 <math>\pm</math> 1.25</b>	<b>76.27 <math>\pm</math> 1.33</b>	<b>77.14 <math>\pm</math> 0.79</b>	<b>79.02 <math>\pm</math> 0.42</b>	75.14 $\pm$ 7.21	<b>95.94 <math>\pm</math> 1.37</b>	<b>97.08 <math>\pm</math> 1.07</b>	<b>98.31 <math>\pm</math> 0.20</b>	1.3
SGC	No Transfer	45.87 $\pm$ 5.79	62.40 $\pm$ 2.77	68.51 $\pm$ 2.41	78.49 $\pm$ 0.35	39.47 $\pm$ 3.88	92.37 $\pm$ 2.25	93.36 $\pm$ 1.10	96.13 $\pm$ 0.16	5.3
	ERM	73.44 $\pm$ 0.87	74.37 $\pm$ 0.80	74.63 $\pm$ 0.81	75.23 $\pm$ 1.04	70.07 $\pm$ 0.73	81.65 $\pm$ 1.61	81.43 $\pm$ 1.34	82.93 $\pm$ 0.89	4.8
	Multi-task	71.00 $\pm$ 0.62	71.76 $\pm$ 1.11	72.12 $\pm$ 2.39	74.71 $\pm$ 2.08	73.53 $\pm$ 1.16	79.35 $\pm$ 0.70	83.76 $\pm$ 1.06	84.27 $\pm$ 0.67	5.9
	EERM	72.45 $\pm$ 0.50	72.95 $\pm$ 1.20	74.55 $\pm$ 0.85	74.90 $\pm$ 0.58	74.35 $\pm$ 0.74	80.89 $\pm$ 0.58	82.12 $\pm$ 0.69	87.40 $\pm$ 0.45	4.8
	GTrans	71.72 $\pm$ 0.39	72.02 $\pm$ 1.95	73.97 $\pm$ 2.10	74.03 $\pm$ 0.50	80.29 $\pm$ 0.49	92.64 $\pm$ 0.61	93.68 $\pm$ 1.05	94.84 $\pm$ 1.54	4.4
	GNN-SP	74.94 $\pm$ 1.24	75.67 $\pm$ 0.90	77.10 $\pm$ 1.12	<b>79.35 <math>\pm</math> 0.46</b>	<u>82.86 <math>\pm</math> 1.06</u>	<b>96.07 <math>\pm</math> 1.83</b>	<u>96.20 <math>\pm</math> 1.53</u>	<u>96.28 <math>\pm</math> 1.48</u>	1.8
	GNN-SP++	<b>74.99 <math>\pm</math> 0.57</b>	<b>76.59 <math>\pm</math> 1.72</b>	<b>77.30 <math>\pm</math> 1.30</b>	79.15 $\pm$ 0.41	<b>83.97 <math>\pm</math> 1.93</b>	95.75 $\pm$ 1.69	<b>96.83 <math>\pm</math> 1.70</b>	<b>97.09 <math>\pm</math> 1.67</b>	1.3

$q$  denotes the ratio of training nodes in the target graph. For example,  $q = 10\%$  indicates 10 percent of nodes in the target graph are used for fine-tuning.

Table 2: Node classification performance on Citation dataset. Rank indicates the average rank of all settings.

aims to cluster nodes with similar localized structural distributions. RW sampler further enhances the distinctiveness between structurally distant nodes, thereby creating more distinguishable clusters (Figure 2 (Left)). This improved distinguishability helps the classifier to capture meaningful information in prediction (Figure 2 (Right)).

Another approach to mitigate the risk of over-smoothing is to design advanced pooling functions. For example, we can employ attention mechanism [Lee *et al.*, 2019] or hierarchical pooling [Wu *et al.*, 2022a] to adaptively assign pooling weights  $w_{ij}$  to nodes within a subgraph, thus preserving the uniqueness of subgraph embeddings. However, empirical evidence suggests that these advanced methods offer no significant advantage compared to basic MEAN pooling (Sec. 4.3). Moreover, there is a concern regarding the efficiency of complicated pooling functions, as they could potentially increase computational and optimization efforts at each epoch. In contrast, the proposed RW sampling can be efficiently executed during pre-processing. To illustrate how RW sampling alleviates over-smoothing, we provide a concrete example below.

**Example 1.** Considering two nodes  $u, v \in \mathcal{V}^s$  in the source graph with  $k$ -hop sampler. Suppose  $u, v$  share an identical  $k$ -hop subgraph yet different labels, i.e.,  $\mathcal{N}_s(u) = \mathcal{N}_s(v)$  and  $y_u \neq y_v$ , employing RW to sample neighborhoods  $\mathcal{N}_r(u)$  and  $\mathcal{N}_r(v)$ ,  $\mathcal{N}_r(u) \neq \mathcal{N}_r(v)$ , can achieve lower empirical risk.

**Illustration.** Let  $\mathbf{h}_u = \sum_{i \in \mathcal{N}_s(u) \cup u} \mathbf{z}_i / (|\mathcal{N}_s(u)| + 1)$  and  $\mathbf{h}_v = \sum_{j \in \mathcal{N}_s(v) \cup v} \mathbf{z}_j / (|\mathcal{N}_s(v)| + 1)$  as subgraph embeddings for node  $u, v$  via MEAN pooling, where  $\mathbf{h}_u = \mathbf{h}_v$ . The empirical risk with classifier  $g(\cdot)$  is given by:

$$R_S = \frac{1}{2} ((g(\mathbf{h}_u) - y_u)^2 + (g(\mathbf{h}_v) - y_v)^2). \quad (10)$$

For simplicity, we use the mean square loss. We consider that  $R_S$  is minimized when  $g(\mathbf{h}_u) = y_u$  and  $g(\mathbf{h}_v) = y_v$ , but it is impossible to find a classifier  $g(\cdot)$  that projects a single vector into different labels. However, by applying RW

to sample subgraphs with  $\mathcal{N}_r(u)$  and  $\mathcal{N}_r(v)$ , we can obtain distinct subgraph embeddings  $\bar{\mathbf{h}}_u \neq \bar{\mathbf{h}}_v$ . Then, it becomes feasible to find a classifier  $g^*(\cdot)$  such that  $g^*(\bar{\mathbf{h}}_u) = y_u$  and  $g^*(\bar{\mathbf{h}}_v) = y_v$  to minimize  $R_S$ . In the extreme case, the subgraphs sampled by RW might be the same as  $k$ -hop, leading to  $\bar{\mathbf{h}}_u = \bar{\mathbf{h}}_v$ . To mitigate the issue, we can control the walk length and sampling frequency to maintain the distinctiveness of the sampled subgraphs. Therefore, utilizing RW sampler can lead to a lower empirical risk.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets.** We use Citation network [Wu *et al.*, 2020], consisting of ACMv9 and DBLPv8; Airport network [Ribeiro *et al.*, 2017], including Brazil, USA, and Europe; Twitch network [Rozemberczki *et al.*, 2021] collected from six countries, including DE, EN, ES, FR, PT, RU; citation network Arxiv [Hu *et al.*, 2020] consisting papers with varying publish times; and dynamic financial network Elliptic [Weber *et al.*, 2019] that contains dozens of graph snapshots where each node is a Bitcoin transaction.

**Baselines.** We use four different GNNs as backbones: GCN [Kipf and Welling, 2017], SAGE [Hamilton *et al.*, 2017], GAT [Veličković *et al.*, 2018], and SGC [Wu *et al.*, 2019]. We compare our proposed Subgraph Pooling with No Transfer (directly training on target), Empirical Risk Minimization (pre-training on source and fine-tuning on target), Multi-task (jointly training on source and target), EERM [Wu *et al.*, 2022b], and recent SOTA method GTrans [Jin *et al.*, 2023]. We also compare various domain adaptation methods, including DANN [Ganin *et al.*, 2016], CDAN [Long *et al.*, 2018], UDAGCN [Wu *et al.*, 2020], MIXUP [Wang *et al.*, 2021], EGI [Zhu *et al.*, 2021b], SR-GNN [Zhu *et al.*, 2021a], GRADE [Wu *et al.*, 2023a], SSReg [You *et al.*, 2023], and StruRW [Liu *et al.*, 2023a].



Model	Brazil → Europe	USA → Europe	Brazil → USA	Europe → USA	USA → Brazil	Europe → Brazil	Rank
No Transfer	48.63 ± 3.70		59.18 ± 1.76		52.36 ± 6.46		2.8
ERM	45.00 ± 2.95	39.29 ± 3.96	47.83 ± 1.92	47.79 ± 4.66	39.53 ± 7.70	44.62 ± 4.24	6.8
Multi-task	48.55 ± 1.48	47.61 ± 2.02	48.73 ± 2.01	50.96 ± 2.12	52.17 ± 2.13	52.92 ± 6.07	4.3
EERM	48.77 ± 2.85	46.88 ± 4.70	48.91 ± 4.19	48.36 ± 3.74	45.67 ± 3.68	46.65 ± 5.93	4.8
GTrans	48.50 ± 1.31	47.49 ± 2.41	48.84 ± 0.99	48.88 ± 1.25	52.30 ± 1.50	53.00 ± 4.12	4.5
GNN-SP	48.76 ± 2.61	<b>51.30 ± 2.22</b>	46.06 ± 5.44	49.85 ± 5.55	<u>55.47 ± 5.90</u>	<u>54.72 ± 5.48</u>	3.2
GNN-SP++	<b>50.90 ± 3.93</b>	50.40 ± 2.27	<b>51.06 ± 6.17</b>	<b>53.87 ± 5.99</b>	<b>57.08 ± 6.13</b>	<b>55.23 ± 9.36</b>	1.5

Table 3: Node classification performance across Airport networks with GCN backbone.

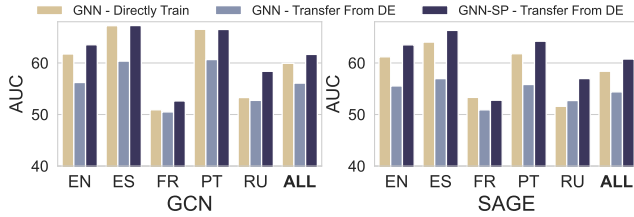


Figure 3: Node classification performance on Twitch.

**Settings.** We pre-train the model on the source with 60 percent of labeled nodes and adapt the model to the target. The adaptation involves three settings: (1) directly applying the pre-trained model without any fine-tuning (Without FT); (2) fine-tuning the last layer (classifier) of the model (FT Last Layer); (3) fine-tuning all parameters (Fully FT). We take split 10/10/80 to form train/valid/test sets on the target graph. Specific hyper-parameters are presented in the Appendix C.

## 4.2 Node Classification

**One-to-One Transfer.** We use Citation and Airport as benchmarks and employ 2-layer GNNs with 64 dimensions as backbones. Considering the limited number of parameters, we apply the FT Last Layer setting. The transfer learning results on Citation over three GNN backbones are presented in Table 2. Our proposed GNN-SP outperforms other baselines across all settings and is even better than the model trained from scratch (No Transfer), demonstrating the capability of overcoming negative transfer. Note that GNN-SP performs better than the advanced GNN-SP++ with extremely limited labels. It is may because (1) the dataset is small and sparse ( $\sim 5k$  to  $\sim 7k$  nodes and  $\sim 20k$  to  $\sim 30k$  edges) where RW sampler fails to model the real localized structures; and (2) the label sparsity exacerbates overfitting.

The performance on Airport is presented in Table 3. Our method surpasses all baselines and achieves an average Rank of 1.5, which is notably higher than the No Transfer. Note that transferring knowledge to USA results in negative transfer for all baselines, likely due to its significantly larger size compared to the other two graphs. Specifically, USA contains 1,190 nodes and 28,388 edges, whereas Europe and Brazil contain only 399 nodes & 12,385 edges, and 131 nodes & 2,137 edges, respectively. The smaller graphs provide limited patterns and may introduce unexpected biases during pre-training. Additionally, we observe that GNN-SP++ performs better than GNN-SP, likely because these graphs are densely connected, leading to the issue that various nodes share identical  $k$ -hop subgraphs.

**One-to-Multi Transfer.** We use Twitch as the benchmark, which consists of six graphs with different sizes and data distributions. We pre-train the model on DE and fine-tune on

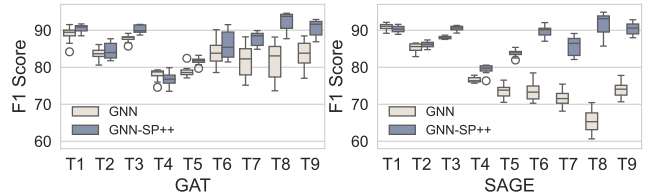


Figure 4: Node classification performance on Elliptic.

Model	Time 1	Time 2	Time 3	Time 4	Rank
No Transfer	69.60 ± 0.31				2.8
ERM	65.73 ± 0.57	66.18 ± 0.48	68.67 ± 0.32	70.33 ± 0.29	4.5
Multi-task	50.32 ± 2.17	52.77 ± 2.82	60.02 ± 0.99	67.62 ± 0.75	6.8
EERM	55.25 ± 2.03	57.47 ± 0.59	63.25 ± 0.54	65.26 ± 0.63	6.3
GTrans	65.95 ± 0.12	66.64 ± 0.51	69.51 ± 0.39	71.54 ± 0.30	3.3
GCN-SP	<u>67.76 ± 0.23</u>	<u>68.36 ± 0.33</u>	69.03 ± 0.63	69.75 ± 0.56	3.5
GCN-SP++	<b>71.43 ± 0.52</b>	<b>72.75 ± 1.24</b>	<b>74.04 ± 0.83</b>	<b>75.17 ± 0.21</b>	1.0

Table 4: Node classification on Arxiv with GCN backbone.

other graphs (EN, ES, FR, PT, RU). We employ ROC-AUC as metric and adopt 2-layer GCN and SAGE as backbones. Figure 3 shows GNN-SP outperforms standard GNN with up to 8% improvements on ROC-AUC under FT Last Layer setting and achieves better performance than the model directly trained on the target over 10 out of 12 settings. The results validate the generalizability of GNN-SP to multiple graphs.

**Transfer with Dynamic Shift.** In this scenario, the model is pre-trained on datasets collected from the past, and then fine-tuned on future data to evaluate its capability in handling temporal distribution shifts. We first adopt a dynamic financial network Elliptic with splitting 5/5/33 snapshots for train/valid/test and F1-score for evaluation. Figure 4 presents the results where the test snapshots are grouped into 9 folds in chronological order. Our GNN-SP outperforms standard GNN up to 10% and 24% improvements over GAT and SAGE backbones, respectively.

Additionally, we use Arxiv as another temporal dataset, where nodes represent papers published from 2005 to 2020 and edges indicate citations. Based on the publication time, we collect five sub-graphs, represented as Time 1 (2005 - 2007), Time 2 (2008 - 2010), Time 3 (2011 - 2014), Time 4 (2015 - 2017), and Time 5 (2018 - 2020). We use the first four graphs as sources and the last one as the target. The results are presented in Table 4 where our GNN-SP++ achieves significant improvements over all baselines. We note that the temporal distribution shift is marginal when the source and target are temporally proximate, resulting in improved transfer learning performance. Additionally, the performance of GNN-SP++ is considerably superior to GNN-SP, further validating the efficacy of the RW sampler.

	ACM $\rightarrow$ DBLP	DBLP $\rightarrow$ ACM	Twitch-All	Arxiv-T1	Arxiv-T3
GCN + Fully FT	97.75 $\pm$ 0.16	80.03 $\pm$ 0.22	60.59 $\pm$ 1.13	69.29 $\pm$ 0.16	69.70 $\pm$ 0.39
GCN-SP + FT Last Layer	98.20 $\pm$ 0.54	79.20 $\pm$ 0.73	61.66 $\pm$ 0.92	71.43 $\pm$ 0.52	74.04 $\pm$ 0.83
GCN-SP + Fully FT	<b>98.66 <math>\pm</math> 0.29</b>	<b>80.82 <math>\pm</math> 0.59</b>	<b>61.77 <math>\pm</math> 0.98</b>	<b>73.12 <math>\pm</math> 0.93</b>	<b>75.01 <math>\pm</math> 0.62</b>

Table 5: Comparison between fine-tuning the model classifier (FT Last Layer) and fine-tuning the whole model (Fully FT).

Model	ACM & DBLP		Arxiv	
	A $\rightarrow$ D	D $\rightarrow$ A	Time 1	Degree
GCN*	59.02 $\pm$ 1.04	59.20 $\pm$ 0.70	28.08 $\pm$ 0.24	57.41 $\pm$ 0.14
GAT*	61.67 $\pm$ 3.54	62.18 $\pm$ 7.04	32.32 $\pm$ 1.10	58.10 $\pm$ 0.15
DANN	59.02 $\pm$ 7.79	65.77 $\pm$ 0.46	24.33 $\pm$ 1.19	56.13 $\pm$ 0.18
CDAN	60.56 $\pm$ 4.38	64.35 $\pm$ 0.83	25.85 $\pm$ 1.15	56.43 $\pm$ 0.45
UDAGCN	59.62 $\pm$ 2.86	64.74 $\pm$ 2.51	25.64 $\pm$ 3.04	55.77 $\pm$ 0.83
EERM	40.88 $\pm$ 5.10	51.71 $\pm$ 5.07	-	-
MIXUP	49.93 $\pm$ 0.89	63.36 $\pm$ 0.66	28.04 $\pm$ 0.18	59.22 $\pm$ 0.22
EGI*	49.03 $\pm$ 1.50	64.40 $\pm$ 1.03	25.59 $\pm$ 0.25	56.93 $\pm$ 0.23
SR-GNN*	62.49 $\pm$ 1.96	63.32 $\pm$ 1.49	25.44 $\pm$ 0.30	56.98 $\pm$ 0.12
GRADE*	67.29 $\pm$ 2.04	64.13 $\pm$ 3.12	25.69 $\pm$ 0.12	57.49 $\pm$ 0.39
SSReg*	69.04 $\pm$ 2.95	65.93 $\pm$ 1.05	27.93 $\pm$ 0.29	56.67 $\pm$ 0.33
StruRW	70.19 $\pm$ 2.10	65.07 $\pm$ 1.98	28.46 $\pm$ 0.18	57.45 $\pm$ 0.15
GCN-SP++	<b>75.88 <math>\pm</math> 5.57</b>	<b>71.32 <math>\pm</math> 1.33</b>	<b>40.41 <math>\pm</math> 1.07</b>	64.35 $\pm$ 0.41
GAT-SP++	73.78 $\pm$ 8.13	67.05 $\pm$ 4.97	36.38 $\pm$ 2.50	<b>65.53 <math>\pm</math> 0.60</b>

The reported results are from [Liu *et al.*, 2023a]. \* indicates the results of our implementations based on the official code.

Table 6: Node classification results without fine-tuning.

### 4.3 Ablation Study

**Transfer without Fine-tuning.** In previous settings, we fine-tune the model classifier to evaluate the quality of knowledge encoded in backbones. Following [Liu *et al.*, 2023a], we take a further step by directly employing the pre-trained model on target graph without fine-tuning. Table 6 presents the transfer learning performance on *Citation* and *Arxiv* benchmarks. Note that we adopt another domain adaptation setting (Degree) from [Gui *et al.*, 2022]. It is obvious that our proposed SP layer significantly improves the transfer learning performance by enhancing the quality of the encoder.

**Transfer with Fully Fine-tuning.** We also analyze the transfer learning results when fine-tuning the whole model. Table 5 shows the results that even if GNN-SP only fine-tunes the last layer, it still outperforms standard GNN with fully fine-tuning. If we fine-tune the whole model of GNN-SP, the model performance can be further improved, especially on large-scaled *Arxiv* dataset.

**Pooling Methods.** Apart from adopting multiple backbones, we also evaluate model performance with different pooling functions, including MEAN, ATTN, MAX, and GCN. Figure 5 presents the experimental results over *Citation* network with GCN backbone. We observe the basic MEAN pooling outperforms complicated GCN and ATTN that adaptively determine the pooling weights  $w_{ij}$ . It is might because the complicated methods introduce extra inductive bias with unexpected noises.

## 5 Related Works

Existing studies are typically categorized based on the availability of the target graph during the pre-training phase.

**Pre-training with Target Graph.** Researchers developed methods to explicitly align source and target graphs during pre-training, ensuring that  $P_S(\mathcal{Z}) = P_T(\mathcal{Z})$ . For example, [Zhang *et al.*, 2019; Dai *et al.*, 2022b] adopt adversarial learning [Ganin *et al.*, 2016] to train domain-invariant encoder, and following UDAGCN [Wu *et al.*, 2020] incor-

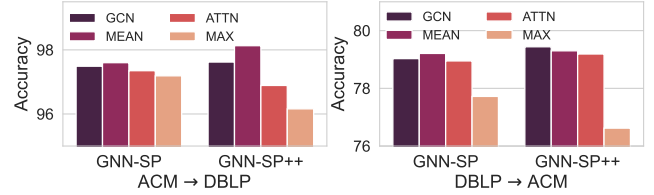


Figure 5: Ablation on *Citation* with different pooling functions.

porates attention to further enhance expressiveness. Alternatively, one can employ regularizers [Zhu *et al.*, 2021a; Zhu *et al.*, 2023; Shi *et al.*, 2023], such as MMD and CMD, to constrain the discrepancy between the source and target. To facilitate this process, new graph-specific discrepancy metrics are proposed, including tree mover distance [Chuang and Jegelka, 2022], subtree discrepancy [Wu *et al.*, 2023a], and spectral regularizer [You *et al.*, 2023]. Additionally, [Liu *et al.*, 2023a] emphasizes the most relevant instances in the source to better match distributions of the source and target. While these methods effectively reduce the distribution shift between source and target, the target graph may be unavailable during pre-training in many real-world scenarios.

**Pre-training without Target Graph.** Existing works aim to train GNNs that can be transferred to unseen target graphs. For example, EERM [Wu *et al.*, 2022b] and following works [Chen *et al.*, 2022; Wu *et al.*, 2022c; Wu *et al.*, 2023b; Yu *et al.*, 2023] utilize causal learning to develop environment-invariant encoder. GTrans [Jin *et al.*, 2023] transforms the target graph at test-time to align the source and target. Moreover, various studies employ augmentation to enhance the robustness of encoder against permutations [Verma *et al.*, 2021; Wang *et al.*, 2021; Liu *et al.*, 2022; Han *et al.*, 2022; Liu *et al.*, 2023b; Guo *et al.*, 2023] or apply disentangle learning to extract domain-invariant semantics [Ma *et al.*, 2019; Liu *et al.*, 2020]. To understand the transferability of GNNs, [Ruiz *et al.*, 2020; Levie *et al.*, 2021; Bevilacqua *et al.*, 2021; Cao *et al.*, 2023] interpret graphs as the combination of graphons, while [Han *et al.*, 2021; Zhu *et al.*, 2021b; Sun *et al.*, 2023; Qiu *et al.*, 2020] adopt self-supervised learning to identify transferable structures. Despite these efforts to enhance transferability, they cannot well address the negative transfer issue. To this end, we systematically analyze why the negative transfer happens and provide insights to solve this issue.

## 6 Conclusion

In this paper, we explore the negative transfer in GNNs and introduce Subgraph Pooling, a simple yet effective method, to mitigate the issue. Our method transfers the subgraph-level knowledge to reduce the discrepancy between the source and target graphs, and is applicable for any GNN backbone without introducing extra parameters. We provide a theoretical analysis to demonstrate how the model works and conduct extensive experiments to evaluate its superiority under various transfer learning settings.

## References

- [Bevilacqua *et al.*, 2021] Beatrice Bevilacqua, Yangze Zhou, and Bruno Ribeiro. Size-invariant graph representations for graph classification extrapolations. In *ICML*, 2021.
- [Cao *et al.*, 2023] Yuxuan Cao, Jiarong Xu, Carl Yang, Jiaan Wang, Yunchao Zhang, Chunping Wang, Lei CHEN, and Yang Yang. When to pre-train graph neural networks? from data generation perspective! In *KDD*, 2023.
- [Chen *et al.*, 2022] Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. *NeurIPS*, 2022.
- [Chuang and Jegelka, 2022] Ching-Yao Chuang and Stefanie Jegelka. Tree mover’s distance: Bridging graph metrics and stability of graph neural networks. In *NeurIPS*, 2022.
- [Dai *et al.*, 2022a] Enyan Dai, Wei Jin, Hui Liu, and Suhang Wang. Towards robust graph neural networks for noisy graphs with sparse labels. In *WSDM*, 2022.
- [Dai *et al.*, 2022b] Quanyu Dai, Xiao-Ming Wu, Jiaren Xiao, Xiao Shen, and Dan Wang. Graph transfer learning via adversarial domain adaptation with graph convolution. *TKDE*, 2022.
- [Fey and Lenssen, 2019] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv*, 2019.
- [Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.
- [Gauzelet *et al.*, 2021] Thomas Gauzelet, Ben Day, Arian R Jamasb, Jyothish Soman, Cristian Regep, Gertrude Liu, Jeremy BR Hayter, Richard Vickers, Charles Roberts, Jian Tang, et al. Utilizing graph machine learning within drug discovery and development. *Briefings in bioinformatics*, 2021.
- [Gretton *et al.*, 2006] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *NeurIPS*, 2006.
- [Gui *et al.*, 2022] Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A graph out-of-distribution benchmark. In *NeurIPS*, 2022.
- [Guo *et al.*, 2023] Yuxin Guo, Cheng Yang, Yuluo Chen, Jixi Liu, Chuan Shi, and Junping Du. A data-centric framework to endow graph neural networks with out-of-distribution detection ability. In *KDD*, 2023.
- [Hamilton *et al.*, 2017] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017.
- [Han *et al.*, 2021] Xueting Han, Zhenhuan Huang, Bang An, and Jing Bai. Adaptive transfer learning on graph neural networks. In *KDD*, 2021.
- [Han *et al.*, 2022] Xiaotian Han, Zhimeng Jiang, Ninghao Liu, and Xia Hu. G-mixup: Graph data augmentation for graph classification. In *ICML*, 2022.
- [He *et al.*, 2020] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*, 2020.
- [Hu *et al.*, 2020] 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. In *NeurIPS*, 2020.
- [Huang *et al.*, 2021] Zexi Huang, Arlei Silva, and Ambuj Singh. A broader picture of random-walk based graph embedding. In *KDD*, 2021.
- [Huang *et al.*, 2023] Wenbing Huang, Yu Rong, Tingyang Xu, Fuchun Sun, and Junzhou Huang. Tackling over-smoothing for general graph convolutional networks. *TPAMI*, 2023.
- [Jin *et al.*, 2023] Wei Jin, Tong Zhao, Jiayuan Ding, Yozen Liu, Jiliang Tang, and Neil Shah. Empowering graph representation learning with test-time graph transformation. In *ICLR*, 2023.
- [Keriven, 2022] Nicolas Keriven. Not too little, not too much: a theoretical analysis of graph (over)smoothing. In *LoG*, 2022.
- [Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.
- [Lee *et al.*, 2019] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *ICML*, 2019.
- [Levie *et al.*, 2021] Ron Levie, Wei Huang, Lorenzo Bucci, Michael Bronstein, and Gitta Kutyniok. Transferability of spectral graph convolutional neural networks. *JMLR*, 2021.
- [Liu *et al.*, 2020] Yanbei Liu, Xiao Wang, Shu Wu, and Zhitao Xiao. Independence promoted graph disentangled networks. In *AAAI*, 2020.
- [Liu *et al.*, 2022] Songtao Liu, Rex Ying, Hanze Dong, Lanqing Li, Tingyang Xu, Yu Rong, Peilin Zhao, Junzhou Huang, and Dinghao Wu. Local augmentation for graph neural networks. In *ICML*, 2022.
- [Liu *et al.*, 2023a] Shikun Liu, Tianchun Li, Yongbin Feng, Nhan Tran, Han Zhao, Qiang Qiu, and Pan Li. Structural re-weighting improves graph domain adaptation. In *ICML*, 2023.
- [Liu *et al.*, 2023b] Yixin Liu, Kaize Ding, Huan Liu, and Shirui Pan. Good-d: On unsupervised graph out-of-distribution detection. In *WSDM*, 2023.
- [Long *et al.*, 2018] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018.
- [Ma *et al.*, 2019] Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. Disentangled graph convolutional networks. In *ICML*, 2019.



- [Mallinar *et al.*, 2022] Neil Mallinar, James Simon, Amirhesam Abedsoltan, Parthe Pandit, Misha Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. In *NeurIPS*, 2022.
- [Qiu *et al.*, 2020] Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc: Graph contrastive coding for graph neural network pre-training. In *KDD*, 2020.
- [Ribeiro *et al.*, 2017] Leonardo FR Ribeiro, Pedro HP Saverese, and Daniel R Figueiredo. struc2vec: Learning node representations from structural identity. In *KDD*, 2017.
- [Rozemberczki *et al.*, 2021] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding. *Journal of Complex Networks*, 2021.
- [Ruiz *et al.*, 2020] Luana Ruiz, Luiz Chamon, and Alejandro Ribeiro. Graphon neural networks and the transferability of graph neural networks. In *NeurIPS*, 2020.
- [Shi *et al.*, 2023] Boshen Shi, Yongqing Wang, Fangda Guo, Jiangli Shao, Huawei Shen, and Xueqi Cheng. Improving graph domain adaptation with network hierarchy. In *CIKM*, 2023.
- [Sun *et al.*, 2023] Xiangguo Sun, Hong Cheng, Jia Li, Bo Liu, and Jihong Guan. All in one: Multi-task prompting for graph neural networks. In *KDD*, 2023.
- [Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.
- [Verma *et al.*, 2021] Vikas Verma, Meng Qu, Kenji Kawaguchi, Alex Lamb, Yoshua Bengio, Juho Kannala, and Jian Tang. Graphmix: Improved training of gnns for semi-supervised learning. In *AAAI*, 2021.
- [Wang *et al.*, 2019] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *CVPR*, 2019.
- [Wang *et al.*, 2021] Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. Mixup for node and graph classification. In *WWW*, 2021.
- [Weber *et al.*, 2019] Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I Weidele, Claudio Bellei, Tom Robinson, and Charles E Leiserson. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv*, 2019.
- [Wenzel *et al.*, 2022] Florian Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. In *NeurIPS*, 2022.
- [Wu *et al.*, 2019] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *ICML*, 2019.
- [Wu *et al.*, 2020] Man Wu, Shirui Pan, Chuan Zhou, Xiaojun Chang, and Xingquan Zhu. Unsupervised domain adaptive graph convolutional networks. In *WWW*, 2020.
- [Wu *et al.*, 2022a] Junran Wu, Xueyuan Chen, Ke Xu, and Shangzhe Li. Structural entropy guided graph hierarchical pooling. In *ICML*, 2022.
- [Wu *et al.*, 2022b] Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. In *ICLR*, 2022.
- [Wu *et al.*, 2022c] Yingxin Wu, Xiang Wang, An Zhang, Xiangan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *ICLR*, 2022.
- [Wu *et al.*, 2023a] Jun Wu, Jingrui He, and Elizabeth Ainsworth. Non-iid transfer learning on graphs. In *AAAI*, 2023.
- [Wu *et al.*, 2023b] Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. Energy-based out-of-distribution detection for graph neural networks. In *ICLR*, 2023.
- [You *et al.*, 2023] Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. Graph domain adaptation via theory-grounded spectral regularization. In *ICLR*, 2023.
- [Yu *et al.*, 2023] Junchi Yu, Jian Liang, and Ran He. Mind the label shift of augmentation-based graph ood generalization. In *CVPR*, 2023.
- [Zellinger *et al.*, 2017] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (CMD) for domain-invariant representation learning. In *ICLR*, 2017.
- [Zhang *et al.*, 2019] Yizhou Zhang, Guojie Song, Lun Du, Shuwen Yang, and Yilun Jin. Dane: domain adaptive network embedding. In *IJCAI*, 2019.
- [Zhang *et al.*, 2022] Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 2022.
- [Zhao and Akoglu, 2020] Lingxiao Zhao and Leman Akoglu. Pairnorm: Tackling oversmoothing in gnns. In *ICLR*, 2020.
- [Zhu *et al.*, 2021a] Qi Zhu, Natalia Ponomareva, Jiawei Han, and Bryan Perozzi. Shift-robust gnns: Overcoming the limitations of localized graph training data. In *NeurIPS*, 2021.
- [Zhu *et al.*, 2021b] Qi Zhu, Carl Yang, Yidan Xu, Haonan Wang, Chao Zhang, and Jiawei Han. Transfer learning of graph neural networks with ego-graph information maximization. In *NeurIPS*, 2021.
- [Zhu *et al.*, 2023] Qi Zhu, Yizhu Jiao, Natalia Ponomareva, Jiawei Han, and Bryan Perozzi. Explaining and adapting graph conditional shift. *arXiv*, 2023.
- [Zhuang *et al.*, 2020] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 2020.

## A Proof

**Theorem 1.** For node  $u \in \mathcal{V}^s$  in the source graph and  $v \in \mathcal{V}^t$  in the target graph, considering the MEAN pooling function, the subgraph embeddings are  $\mathbf{h}_u = \frac{\mathbf{z}_u + \sum_{i \in \mathcal{N}_s(u)} \mathbf{z}_i}{n+1}$ ,  $\mathbf{h}_v = \frac{\mathbf{z}_v + \sum_{j \in \mathcal{N}_s(v)} \mathbf{z}_j}{m+1}$  where  $n = |\mathcal{N}_s(u)|$ ,  $m = |\mathcal{N}_s(v)|$ . We have

$$\|\mathbf{h}_u - \mathbf{h}_v\| \leq \|\mathbf{z}_u - \mathbf{z}_v\| - \Delta, \quad (11)$$

where  $\Delta = \frac{(n\|\mathbf{z}_u - \mathbf{z}_v\| - \frac{m-n}{m+1}\|\mathbf{z}_v\|)}{n+1}$  denotes the discrepancy margin.

*Proof.* To demonstrate this theorem, we analyze the distance between the subgraph embeddings  $\|\mathbf{h}_u - \mathbf{h}_v\|$  in terms of the node embeddings:

$$\left\| \frac{\mathbf{z}_u + \sum_{i \in \mathcal{N}_s(u)} \mathbf{z}_i}{n+1} - \frac{\mathbf{z}_v + \sum_{j \in \mathcal{N}_s(v)} \mathbf{z}_j}{m+1} \right\| = \left\| \frac{(m+1)(\mathbf{z}_u + \sum_i \mathbf{z}_i) - (n+1)(\mathbf{z}_v + \sum_j \mathbf{z}_j)}{(m+1)(n+1)} \right\| \quad (12)$$

$$= \left\| \frac{(m+1)\mathbf{z}_u - (n+1)\mathbf{z}_v + (m+1)\sum_i \mathbf{z}_i - (n+1)\sum_j \mathbf{z}_j}{(m+1)(n+1)} \right\| \quad (13)$$

$$\leq \underbrace{\left\| \frac{(m+1)\mathbf{z}_u - (n+1)\mathbf{z}_v}{(m+1)(n+1)} \right\|}_{(a)} + \underbrace{\left\| \frac{(m+1)\sum_i \mathbf{z}_i - (n+1)\sum_j \mathbf{z}_j}{(m+1)(n+1)} \right\|}_{(b)} \quad (14)$$

This inequality results from the triangle inequality. Simplifying these terms separately, we have the term (a) as:

$$\left\| \frac{(m+1)\mathbf{z}_u - (n+1)\mathbf{z}_v}{(m+1)(n+1)} \right\| = \frac{\|(m+1)(\mathbf{z}_u - \mathbf{z}_v) + (m-n)\mathbf{z}_v\|}{(m+1)(n+1)} \quad (15)$$

$$\leq \frac{\|\mathbf{z}_u - \mathbf{z}_v\|}{n+1} + \frac{(m-n)\|\mathbf{z}_v\|}{(m+1)(n+1)} \quad (16)$$

$$= \|\mathbf{z}_u - \mathbf{z}_v\| - \|\mathbf{z}_u - \mathbf{z}_v\| + \frac{\|\mathbf{z}_u - \mathbf{z}_v\| + \frac{m-n}{m+1}\|\mathbf{z}_v\|}{n+1} \quad (17)$$

$$= \|\mathbf{z}_u - \mathbf{z}_v\| - \frac{(n\|\mathbf{z}_u - \mathbf{z}_v\| - \frac{m-n}{m+1}\|\mathbf{z}_v\|)}{n+1} \quad (18)$$

$$= \|\mathbf{z}_u - \mathbf{z}_v\| - \Delta \quad (19)$$

where  $\Delta = \frac{(n\|\mathbf{z}_u - \mathbf{z}_v\| - \frac{m-n}{m+1}\|\mathbf{z}_v\|)}{n+1}$  indicates the discrepancy gap handled by our proposed Subgraph Pooling. Additionally, the term (b) can be simplified as:

$$\left\| \frac{(m+1)\sum_{i \in \mathcal{N}_s(u)} \mathbf{z}_i - (n+1)\sum_{j \in \mathcal{N}_s(v)} \mathbf{z}_j}{(m+1)(n+1)} \right\| = \left\| \frac{\sum_{i \in \mathcal{N}_s(u)} \mathbf{z}_i}{n+1} - \frac{\sum_{j \in \mathcal{N}_s(v)} \mathbf{z}_j}{m+1} \right\| \leq \epsilon \quad (20)$$

For this term, under Assumption 1, it is sufficiently small to be disregarded. Combining these analyses, we derive  $\|\mathbf{h}_u - \mathbf{h}_v\| \leq \|\mathbf{z}_u - \mathbf{z}_v\| - \Delta$ . While  $\Delta$  may not always be positive, it is essential for understanding the impact of SP. To further substantiate this, we present two corollaries.

**Corollary 1.** If either of the following conditions is satisfied ( $|\mathcal{N}_s(u)| \geq |\mathcal{N}_s(v)|$  or  $|\mathcal{N}_s(u)|$  is sufficiently large), the inequality  $\|\mathbf{h}_u - \mathbf{h}_v\| \leq \|\mathbf{z}_u - \mathbf{z}_v\|$  strictly holds.

*Proof.* To establish that  $\Delta \geq 0$ , we need to demonstrate that:

$$n\|\mathbf{z}_u - \mathbf{z}_v\| \geq \frac{m-n}{m+1}\|\mathbf{z}_v\| \quad (21)$$

$$\frac{\|\mathbf{z}_u - \mathbf{z}_v\|}{\|\mathbf{z}_v\|} \geq \frac{(m-n)}{n(m+1)}, \quad (22)$$

where  $n = |\mathcal{N}_s(u)|$  and  $m = |\mathcal{N}_s(v)|$ .

We consider the following two cases:

1. In cases where the source graph is richer than the target, namely,  $|\mathcal{N}_s(u)| \geq |\mathcal{N}_s(v)|$ , we have  $\frac{(m-n)}{n(m+1)} \leq 0$ . Under these conditions,  $\frac{\|\mathbf{z}_u - \mathbf{z}_v\|}{\|\mathbf{z}_v\|} \geq \frac{(m-n)}{n(m+1)}$  is strictly valid, given that  $\|\mathbf{z}_u - \mathbf{z}_v\| \geq 0$  and  $\|\mathbf{z}_v\| \geq 0$ .
2. When both source and target graphs are substantially rich,  $\frac{(m-n)}{n(m+1)} \leq 0$  remains strictly valid. In the extreme case where  $n \rightarrow \infty$  and  $m \rightarrow \infty$ , it follows that  $\lim_{n \rightarrow \infty, m \rightarrow \infty} \frac{(m-n)}{n(m+1)} = 0$ . This suggests that when the source graph is adequately rich, essential patterns can be transferred effectively, regardless of the scale of the target.

**Corollary 2.** *If the following condition is satisfied ( $|\mathcal{N}_s(u)| < |\mathcal{N}_s(v)|$ ), the inequality  $\|\mathbf{h}_u - \mathbf{h}_v\| \leq \|\mathbf{z}_u - \mathbf{z}_v\|$  strictly holds when  $\lambda \geq 2$ , even in extreme case where  $|\mathcal{N}_s(u)| \rightarrow 0$  and  $|\mathcal{N}_s(v)| \rightarrow \infty$ .*

*Proof.* Assuming the target graph is richer than the source, indicated by  $|\mathcal{N}_s(v)| \geq |\mathcal{N}_s(u)|$ , it is necessary to validate that  $\|\mathbf{z}_u - \mathbf{z}_v\| \geq \|\mathbf{z}_v\|$  to maintain the inequality. This validation becomes crucial in the extreme case where  $n \rightarrow 0$  and  $m \rightarrow \infty$ , leading to  $\lim_{n \rightarrow 0, m \rightarrow \infty} \frac{(m-n)}{n(m+1)} = 1$ . We must demonstrate that  $\frac{\|\mathbf{z}_u - \mathbf{z}_v\|}{\|\mathbf{z}_v\|} \geq \frac{(m-n)}{n(m+1)}$  is valid under these circumstances, implying  $\frac{\|\mathbf{z}_u - \mathbf{z}_v\|}{\|\mathbf{z}_v\|} \geq 1$ . Therefore, when the source is sparse, it may lack sufficient information for effective transfer.

$$\|\mathbf{z}_u - \mathbf{z}_v\| \geq \|\mathbf{z}_v\| \quad (23)$$

$$\Rightarrow \|\mathbf{z}_u - \mathbf{z}_v\|^2 \geq \|\mathbf{z}_v\|^2 \quad (24)$$

$$\Rightarrow \|\mathbf{z}_u\|^2 + \|\mathbf{z}_v\|^2 - 2\|\mathbf{z}_u\|\|\mathbf{z}_v\|\cos\theta \geq \|\mathbf{z}_v\|^2 \quad (25)$$

$$\Rightarrow \|\mathbf{z}_u\|^2 \geq 2\|\mathbf{z}_u\|\|\mathbf{z}_v\|\cos\theta \quad (26)$$

$$\Rightarrow \frac{\mathbf{z}_u^T \mathbf{z}_u}{\mathbf{z}_u^T \mathbf{z}_v} \geq 2 \quad (27)$$

where  $\cos\theta = \mathbf{z}_x^T \mathbf{z}_y / \|\mathbf{z}_x\|\|\mathbf{z}_y\|$ . According to Definition 2, this inequality is satisfied when  $\lambda \geq 2$ .

## B Dataset Details

Dataset	Setting	# Nodes	# Edges	# Classes	# Density	Avg Degree	Max Degree	Metric
Citation	ACMv9	7,410	29,456	6	0.00054	3.98	108	Accuracy
	DBLPv8	5,578	20,158	6	0.00065	3.61	254	
Airport	USA	1,190	28,388	4	0.02006	23.86	239	Accuracy
	Europe	399	12,385	4	0.07799	31.04	203	
	Brazil	131	2,137	4	0.12548	16.31	80	
Twitch	DE	9,498	315,774	2	0.00350	33.25	4,260	ROC-AUC
	EN	7,126	77,774	2	0.00153	10.91	721	
	ES	4,648	123,412	2	0.00571	26.55	1,023	
	FR	6,551	231,883	2	0.00540	35.40	2,041	
	PT	1,912	64,510	2	0.01766	33.74	768	
	RU	4,385	78,993	2	0.00411	18.01	1,230	
Arxiv	Time 1 [2005 - 2007]	4,980	10,086	40	0.00041	2.03	14	Accuracy
	Time 2 [2008 - 2010]	12,974	32,215	40	0.00019	2.48	43	
	Time 3 [2011 - 2014]	41,125	140,526	40	0.00008	3.42	76	
	Time 4 [2015 - 2017]	90,941	462,438	40	0.00006	5.09	222	
	Time 5 [2018 - 2020]	169,343	1,335,586	40	0.00005	7.89	437	
	Degree	169,343	2,484,941	40	0.00009	14.67	13,162	

Table 7: The statistics of Citation, Airport, Twitch, and Arxiv networks.

In this section, we detail the datasets utilized in our experiments, with their statistics summarized in Table 7.

- **Citation.** The Citation networks consist of the ACM and DBLP datasets, each sourced from distinct academic databases. In these networks, nodes correspond to academic papers, while edges represent citations between them. Both ACM and DBLP demonstrate comparable sizes and densities. The datasets are categorized into six research fields, namely Database, Data Mining, Artificial Intelligence, Computer Vision, Information Security, and High Performance Computing. Despite the scale of DBLP is relatively smaller compared to ACM, it has a higher maximum degree of 254, as opposed to 108 of ACM. This difference is attributed to the variation in structural distribution between the two datasets.
- **Airport.** The Airport networks include three distinct datasets from various regions, where nodes symbolize airports, and edges represent flight connections. The labels in these datasets indicate the airport activity level, quantified by the number of flights or passengers. Compared to other datasets, the density of the Airport is much higher where the number of edges is much higher than the number of nodes. It is noteworthy that the USA dataset might encapsulate more intricate patterns due to its larger size compared to Europe and Brazil.
- **Twitch.** Twitch dataset contains six networks from different regions, including DE, EN, ES, FR, PT, and RU. In these networks, nodes correspond to users on the streaming platform, and edges reflect friendships. The primary task is to predict mature content within the network. The model is initially pre-trained on the comparatively larger DE network and subsequently transferred to other networks, each possessing distinct distributions on network scales and densities. These networks are leveraged to test model expressiveness on various target graphs.
- **Arxiv.** Arxiv is another citation network where nodes represent papers published from 2005 to 2020, and edges correspond to citations between these papers. The network is released by OGB. We divided the original dataset into five

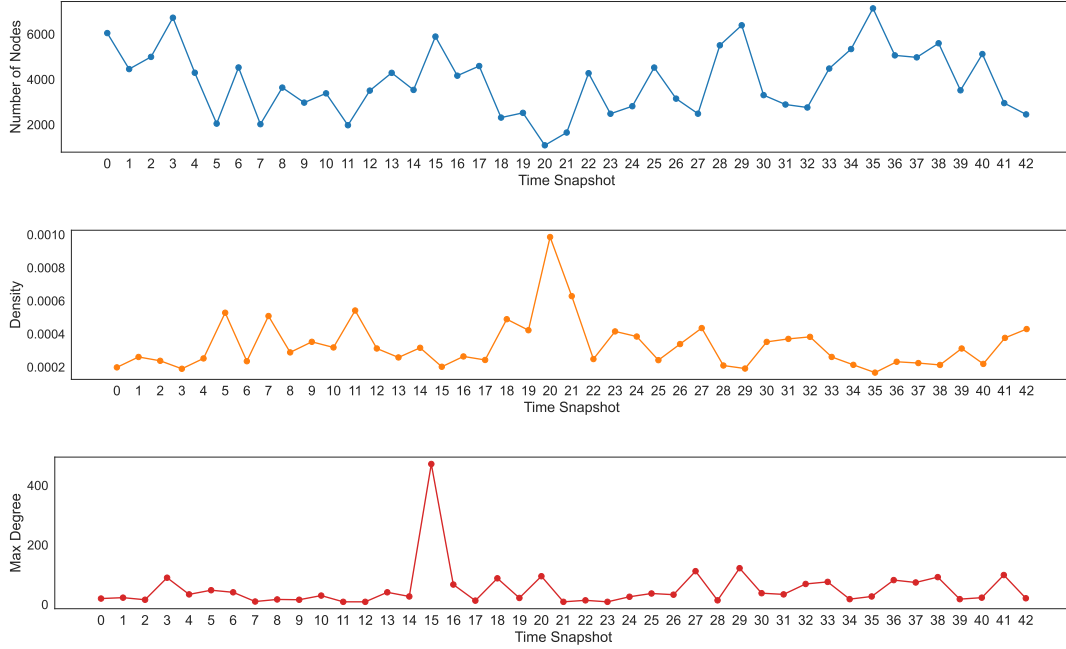


Figure 6: Statistics of `Elliptic` dataset.

distinct sub-datasets, each segmented according to the time of paper publication, shown in Table 7. We transfer knowledge from the previous four time periods to the final one, each characterized by differing temporal distributions. Generally, temporally close networks share relatively similar distributions.

- `Elliptic`. `Elliptic` consists of 43 snapshots, where the dataset distributions of each one are illustrated in Figure 6. Specifically, we present the number of nodes, density, and the max degree of each snapshot. We can observe a clear temporal distribution shift across these graphs. We use a 5/5/33 split for train/valid/test sets to evaluate the model robustness on long-range temporal distribution shift.

## C Hyper-parameters

To prevent the impact of randomness, we run each model 10 times and report the mean and standard deviation. For all baseline models, we standardized the hidden dimension to 64, the learning rate to 0.001 (1e-3), and the number of backbone layers to 2. In the case of attention-based methods, we configured the number of attention heads to 4. For additional configurations, we adhered to the specifications outlined in their respective original papers. The comprehensive hyper-parameter setting for our proposed Subgraph Pooling method is detailed in Table 8.

	Citation	Airport	Twitch	Arxiv	Elliptic	Facebook
Hidden Dimension		64 for all datasets				
Activation		PReLU used for all datasets				
# Encoder Layers	2	2	2	2	2	5
Normalization		-				
Pre-train Learning Rate	1e-3	1e-3	1e-3	1e-3	1e-3	1e-2
Pre-train Epochs	200 (D $\rightarrow$ A) & 500 (A $\rightarrow$ D)	500	100	500	500	200
Pre-train Weight Decay		0 for all datasets				
Fine-tune Learning Rate	1e-3	1e-3	1e-3	1e-3	-	-
Fine-tune Epochs	3000	3000	3000	3000	-	-
Fine-tune Weight Decay		1e-5 for all datasets				
k (SP)	2	2	1	1	-	1
k (SP++)	3	3	-	3	3	10
repeat (SP++)	100	100	-	50	50	10
Pooling	MEAN	ATTN	GCN	GCN	GCN	MEAN
Optimizer		AdamW used for all datasets				
Early Stop		200 for all datasets				

Table 8: Hyper-parameters.

## D Additional Experiments

### D.1 Facebook Dataset

We conduct experiments on `Facebook` dataset. This dataset contains 100 snapshots of Facebook friendship networks from 2005, with each network representing users from a specific American university. For our experiments, we selected 14 networks, including those from John Hopkins, Caltech, Amherst, Bingham, Duke, Princeton, WashU, Brandeis, Carnegie, Penn, Brown, Texas, Cornell, and Yale. The test datasets are Penn, Brown, and Texas, while Cornell and Yale are utilized for evaluation. The training sets are combinations of the remaining nine graphs. These datasets display unique sizes, densities, and degree distributions, as depicted in Figure 7. Notably, the distributions of the testing graphs differ significantly from those of the training and validation sets, providing a rigorous assessment of out-of-distribution (OOD) performance. Moreover, we use the No Fine-tune setting due to the absence of training data on the target graphs. The effectiveness of our methods is evident in Table 9, where our GNN-SP outperformed ERM and EERM in 8 out of 9 settings. This result highlights the robust transferability of our approach across multiple sources with varied distributions and multiple target environments.

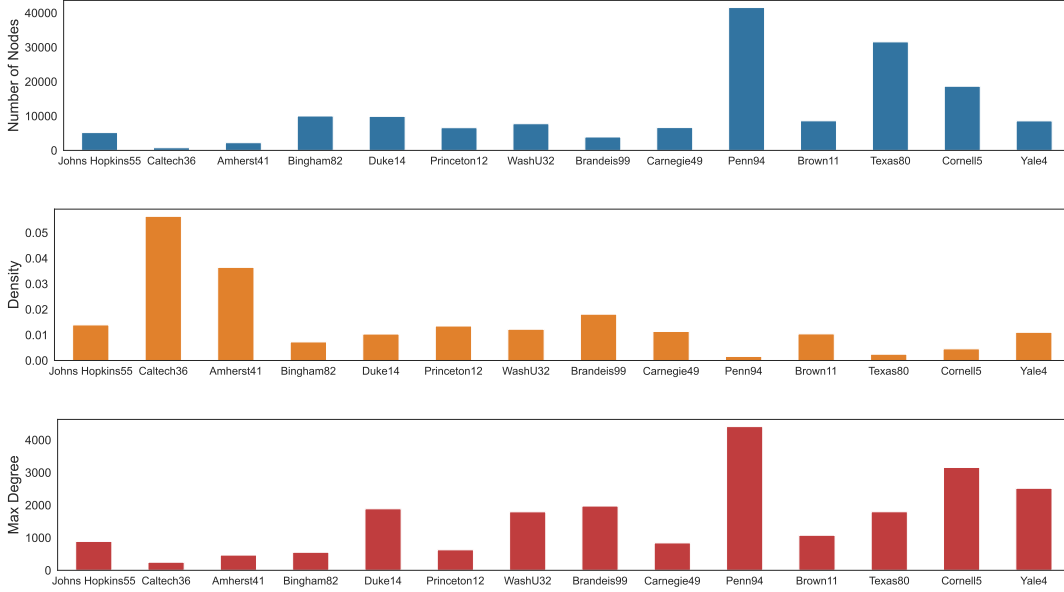


Figure 7: Statistics of `Facebook` dataset.

Combinations	Penn			Brown			Texas		
	ERM	EERM	GNN-SP	ERM	EERM	GNN-SP	ERM	EERM	GNN-SP
John Hopkins + Caltech + Amherst	50.48±1.09	50.64±0.25	<b>52.36±0.54</b>	54.53±3.93	56.73±0.23	<b>56.86±0.04</b>	53.23±4.49	55.57±0.75	<b>55.93±1.00</b>
Bingham + Duke + Princeton	50.17±0.65	50.67±0.79	<b>51.81±0.52</b>	50.43±4.58	52.76±3.40	<b>56.34±0.05</b>	50.19±5.81	53.82±4.88	<b>56.34±0.05</b>
WashU + Brandeis + Carnegie	50.83±0.17	<b>51.52±0.87</b>	50.85±0.05	54.61±4.75	55.15±3.22	<b>56.96±0.02</b>	56.25±0.13	56.12±0.42	<b>56.32±0.05</b>

Table 9: Node classification performance across `Facebook` networks with GCN backbone.

### D.2 More Analysis

**Varying Training Ratio.** The transfer learning performance highly relies on the number of training instances available on the target. To evaluate the sensitivity to varying training sample sizes, we consider four training ratios  $\{0.1\%, 0.5\%, 1\%, 10\%\}$ , as shown in Table 2 in the main body. Our propose GNN-SP outperforms all baselines. We observe that transfer learning models achieve better performance than the model trained from scratch (No Transfer) with extremely limited labels. This is because the knowledge gained from the source provides beneficial regularization during the fine-tuning process. However, when the training ratio reaches  $q = 10\%$ , standard GNNs acquire sufficient knowledge to surpass other baselines, with the exception of our GNN-SP.