KDD2023 AUGUST 6-10 LONG BEACH, CA

When to Pre-Train Graph Neural Networks? From Data Generation Perspective!

Yuxuan Cao*[1], Jiarong Xu*[2], Carl Yang[3], Jiaan Wang[4], Yunchao Zhang[1], Chunping Wang[5], Lei Chen[5], Yang Yang[1]

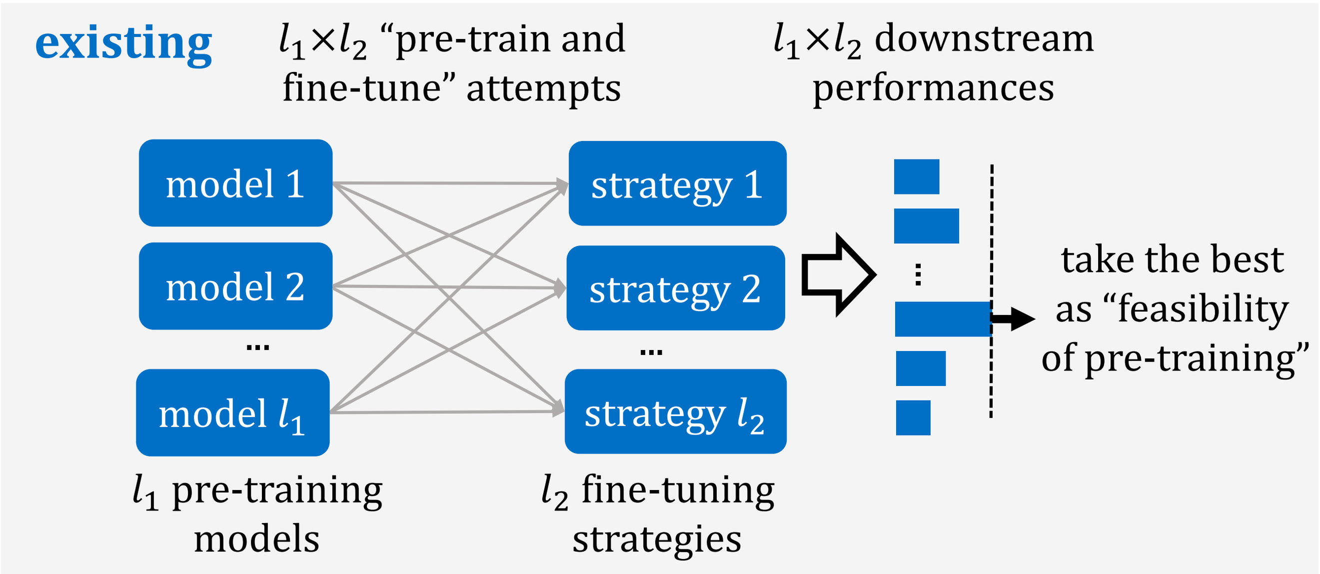
1 Zhejiang University, 2 Fudan University, 3 Emory University, 4 Soochow University, 5 Finvolution Group * Equal contribution.

Motivation

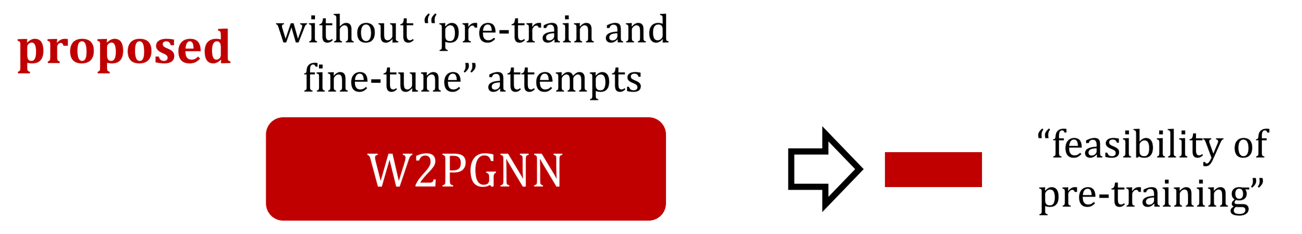
■ To avoid the negative transfer, recent efforts focus on **what to pre-train** and **how to pre-train**. However, the transferability from pre-training data to downstream data cannot be guaranteed in some cases.

■ It is a necessity to understand **when to pre-train**, i.e., under what situations the "graph pre-train and fine-tune" paradigm should be adopted.

■ Existing methods train and evaluate on **all candidates of pre-training models and fine-tuning strategies**, which is very costly. We propose a W2PGNN framework to answer **when to pre-train GNNs** from a **graph data generation perspective.**



(a) Existing methods make costly "pre-train and fine-tune" attempts.



(b) W2PGNN tells the feasibility of pre-training before "pre-train and fine-tune".

W2PGNN Framework

Application Cases

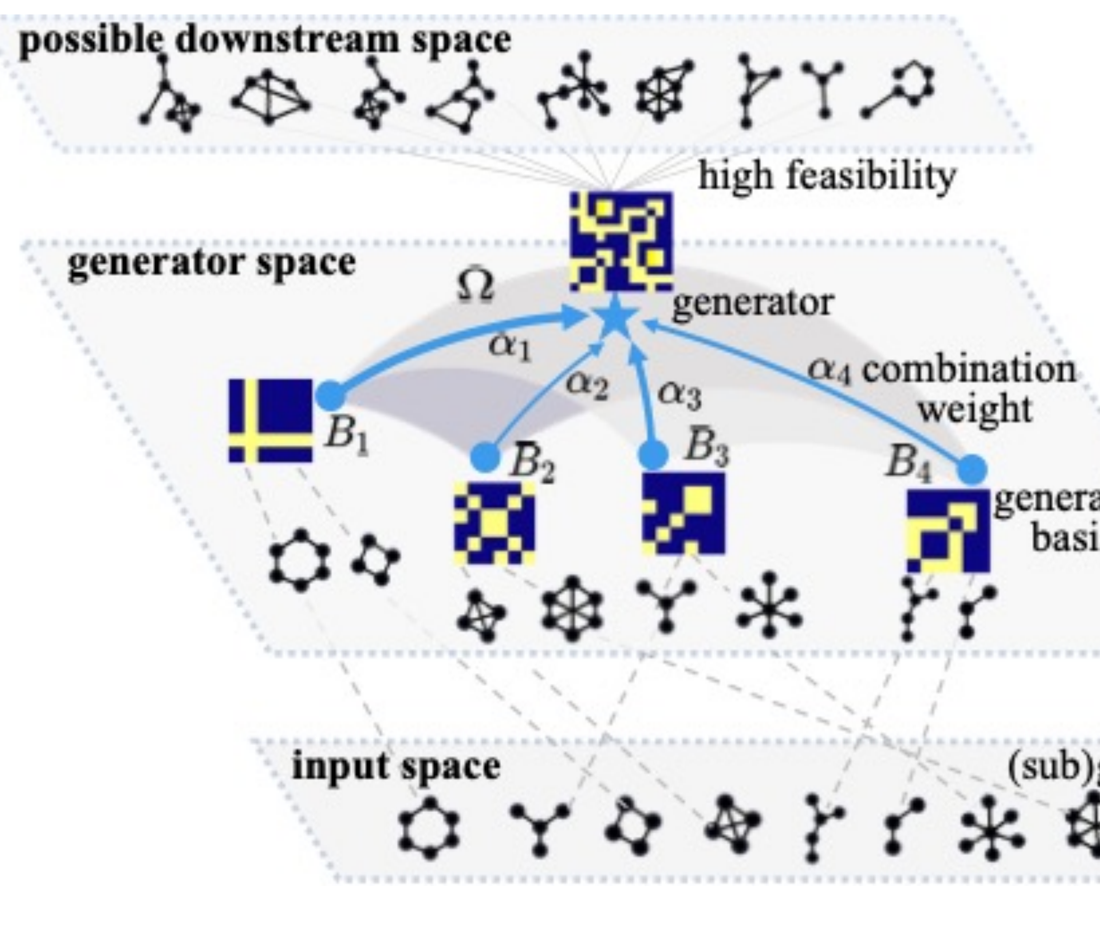
■ Provide the **application scope** of a graph pre-trained model.

■ Estimate the **feasibility** of performing pre-training for a downstream.

■ Pre-training **data selection** to benefit the downstream.

Key Insight: Downstream data can benefit from pre-training data (i.e., has high feasibility of performing pre-training), if it can be generated with high probability by a graph generator that summarizes the transferable patterns of pre-training data.

Figure : Illustration of our proposed framework W2PGNN to answer when to pre-train GNNs.



■ **input space:**
ego-networks (node-level) & graphs (graph-level)

■ **generator space:**

- a graphon basis B_i (i.e., generator) fitted from a set of (sub)graphs with similar patterns.
- each B_i is assigned with a corresponding weight α_i .
- weighted combination of generator basis $f(\{\alpha_i\},\{B_i\}) = \sum_{i=1}^{K} \alpha_i B_i$
- generator space: all weighted combinations $\Omega = \{f(\{\alpha_i\},\{B_i\}) | \forall \{\alpha_i\},\{B_i\}\}$

■ **possible downstream space:**
all the graphs produced by the generators in the generator space $D = \{G \leftarrow f | f \in \Omega\}$.

Feasibility Definition & Approximation

■ **Definition[feasibility of performing pre-training]:**

$$\zeta(\mathcal{G}_{\text{train}} \to \mathcal{G}_{\text{down}}) = \sup_{\{\alpha_i\},\{B_i\}} \Pr(\mathcal{G}_{\text{down}} \mid f(\{\alpha_i\},\{B_i\}))$$

highest probability of the downstream data generated from a generator in the generator space

Problem: Exhausting all possible $\{\alpha_i\}, \{B_i\}$ is impractical.

■ **Approximated feasibility :**

$$\zeta \leftarrow -\text{MIN}\left(\left\{\inf_{\{\alpha_i\}} \text{dist}(f(\{\alpha_i\},\{B_i\}), B_{\text{down}}), \forall \{B_i\} \in \mathcal{B}\right\}\right),$$

☐ **reduced generator basis space**

$$\mathcal{B} = \left\{\{B_i\}_{\text{topo}}, \{B_i\}_{\text{domain}}, \{B_i\}_{\text{integr}}\right\}$$

- integrated basis $\{B_i\}_{\text{integr}}$
- domain basis $\{B_i\}_{\text{domain}}$
- topological basis $\{B_i\}_{\text{topo}}$

☐ **$\{\alpha_i\}$ is learnable parameter**

Theoretical Analysis

■ **An illustrative example**

Assume a collection of pre-training graphs fit into a generator basis $\{B_1, B_2, B_3\}$, with corresponding key transferable patterns , and , respectively. Their convex combination gives rise to a mixed generator $f(\{\alpha_i\},\{B_i\}) = \sum_{i=1}^{K} \alpha_i B_i$.

☐ **Theoretical Justification of Generator Space.**

The following theory proves that all these three transferable patterns (, and) and their mixtures can occur frequently in the mixed generator with high probability.

Theorem 5.2. *Assume a graphon basis $\{B_1, \cdots, B_k\}$ and their convex combination $f(\{\alpha_i\},\{B_i\}) = \sum_{i=1}^{k} \alpha_i B_i$. The a-th element of graphon basis B_a corresponds to a motif set. For each motif F_a in the motif set, the difference between the homomorphism density of F_a in $f(\{\alpha_i\},\{B_i\})$ and that in basis element B_a is upper bounded by*

$$|t(F_a, f(\{\alpha_i\},\{B_i\})) - t(F_a, B_a)| \le \sum_{b=1, b\neq a}^{k} |F_a| \alpha_b ||B_b - B_a||_\square \quad (8)$$

where $|F_a|$ represents the number of nodes in motif F_a, $||\cdot||_\square$ is the cut norm.

☐ **Theoretical Justification of Possible downstream Space.**

The following theory proves that all graphs generated from generator space preserve a mixture of key transferable patterns in mixed generator, e.g., a mixture of and : .

Theorem 5.3. *Given a graph generator $f(\{\alpha_i\},\{B_i\})$, we can obtain sufficient number of random graphs $\mathbb{G} = \mathbb{G}(n, f(\{\alpha_i\},\{B_i\}))$ with n nodes generated from $f(\{\alpha_i\},\{B_i\})$. The homomorphism density of graph motif F in $\mathbb{G}$ can be considered approximately equal to that in $f(\{\alpha_i\},\{B_i\})$ with high probability and can be represented as*

$$\mathrm{P}(|t(F,\mathbb{G}) - t(F, f(\{\alpha_i\},\{B_i\}))| > \varepsilon) \le 2\exp\left(-\frac{\varepsilon^2 n}{8v(F)^2}\right), \quad (9)$$

where $v(F)$ denotes the number of nodes in F, and $0 \le \epsilon \le 1$.

Experiment Results

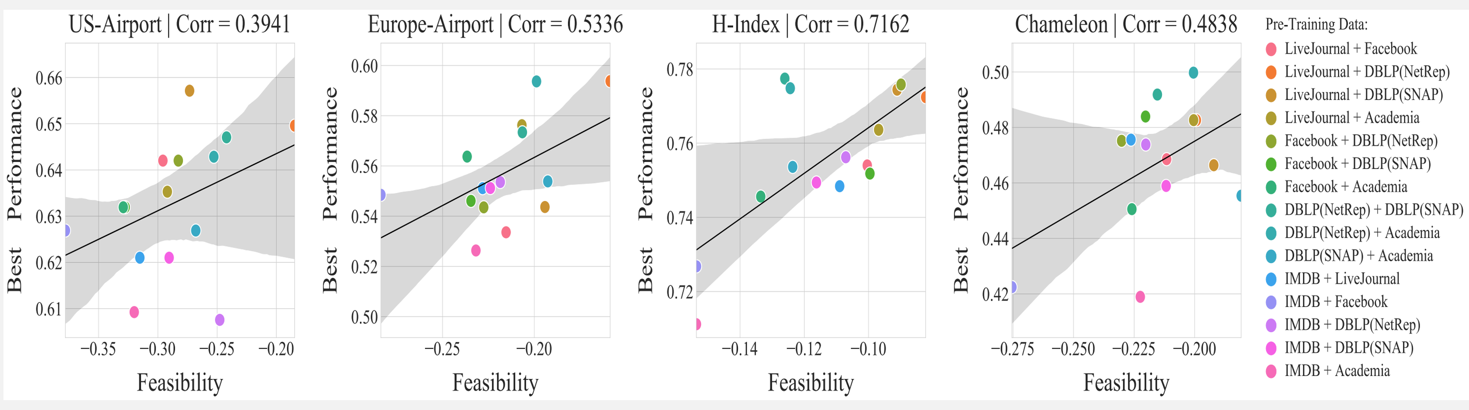
■ **Q1**: Is the feasibility of pre-training estimated by W2PGNN positively correlated with the downstream performance (application case of **feasibility**)?

☐ Table: Pearson correlation coefficient between the estimated feasibility and the best downstream performance on node classification. *N* denotes the number of candidate pre-training datasets (i.e., select budget) that form the pre-training data.

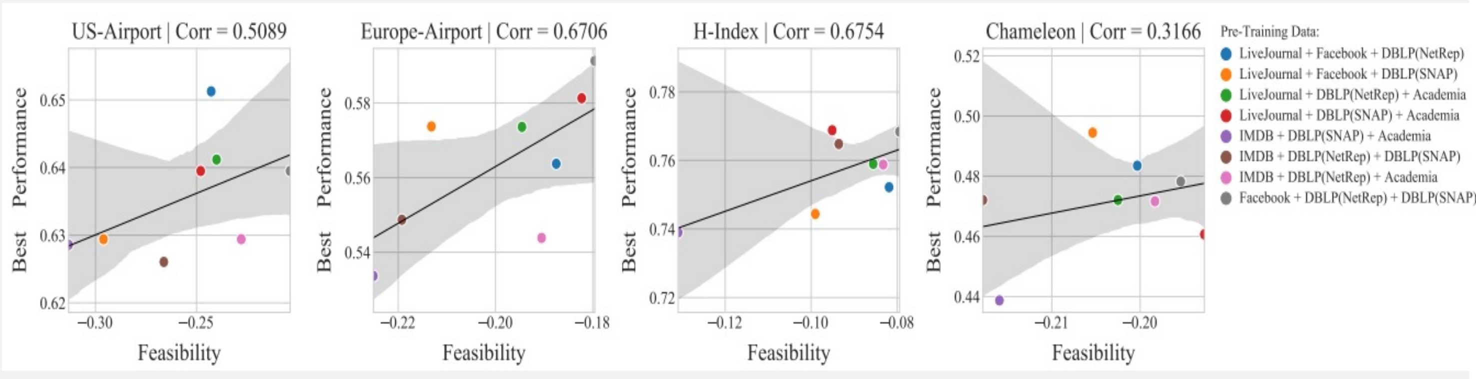
	N = 2					N = 3				
	US-Airport	Europe-Airport	H-index	Chameleon	Rank	US-Airport	Europe-Airport	H-index	Chameleon	Rank
Graph Statistics	-0.6068	0.3571	-0.6220	-0.2930	10	-0.7096	-0.5052	-0.2930	-0.8173	10
EGI	0.0672	-0.6077	-0.2152	-0.2680	9	-0.2358	-0.5540	-0.2822	-0.6511	9
Clustering Coefficient	-0.0273	0.1519	0.3622	0.3130	5	-0.0039	0.2069	0.4829	0.2279	4
Spectrum of Graph Laplacian	-0.2023	0.1467	0.0794	0.0095	8	-0.7648	-0.4311	0.2611	-0.2300	8
Betweenness Centrality	-0.2739	-0.2554	0.2051	0.2241	7	-0.3421	-0.5903	0.1364	0.0849	7
W2PGNN (intergr)	0.3579	0.1224	0.3313	0.1072	6	0.0841	0.5310	0.4213	-0.0916	6
W2PGNN (domain)	**0.4774**	0.4666	0.6775	0.3460	3	**0.7132**	0.5523	**0.7381**	0.1857	3
W2PGNN (topo)	0.2059	0.3908	0.3745	0.4464	4	0.4900	0.5061	0.4072	0.1497	5
W2PGNN ($\alpha = 1$)	0.4172	0.5206	0.6829	0.4391	2	0.5282	0.6663	0.7240	**0.3246**	1
W2PGNN	0.3941	**0.5336**	**0.7162**	**0.4838**	1	0.5089	**0.6706**	0.6754	0.3166	2

■ The feasibility estimated by W2PGNN achieve the highest overall ranking in most cases!

☐ Figure: Estimated feasibility (in x-axis) versus the best downstream performance (in y-axis) of all <pre-training data, downstream data> pairs on node classification when select budget is 2.



☐ Figure: Estimated feasibility (in x-axis) versus the best downstream performance (in y-axis) of all <pre-training data, downstream data> pairs on node classification when select budget is 3.



■ A strong positive correlation between estimated pre-training feasibility and the best downstream performance!

■ **Q2**: Does the pre-training data selected by W2PGNN actually help improve the downstream performance (application case of **data selection**)?

☐ Table: Node classification results when performing pre-training on different selected pre-training data. "All Datasets" refers to the results of using all pre-training data without selection.

		N = 2					N = 3				
		US-Airport	Europe-Airport	H-index	Chameleon	Rank	US-Airport	Europe-Airport	H-index	Chameleon	Rank
①	All Datasets	65.62	55.65	75.22	46.81	-	65.62	55.65	75.22	46.81	-
	Graph Statistics	64.20	53.36	74.30	44.31	4	62.27	54.58	72.88	43.87	5
	EGI	**64.96**	57.37	74.30	43.21	2	62.27	57.36	72.88	45.93	3
	Clustering Coefficient	62.61	52.87	**77.74**	43.21	3	62.94	54.58	75.18	44.66	4
	Spectrum of Graph Laplacian	61.76	**57.88**	73.14	42.20	5	**63.95**	54.87	73.90	44.66	2
	Betweenness Centrality	**64.96**	52.87	73.50	41.63	6	62.27	54.87	75.18	43.87	6
②	W2PGNN	**64.96**	**57.88**	77.24	**45.54**	1	**63.95**	**57.59**	**75.68**	**46.07**	1

① Using **all pre-training data** for pre-training is not always a reliable choice.

② Pre-training data selected by W2PGNN ranks first.

https://github.com/zjunet/W2PGNN