AnyGraph: Graph Foundation Model in the Wild

Lianghao Xia and Chao Huang
The University of Hong Kong

Aug 3–Aug 7, 2025

Abstract

The growing ubiquity of relational data structured as graphs has underscored the need for graph learning models with exceptional generalization capabilities. However, current approaches often struggle to effectively extract generalizable insights, frequently requiring extensive fine-tuning and limiting their versatility. Graph foundation models offer a transformative solution, with the potential to learn robust, generalizable representations from graph data. This enables more effective and adaptable applications across a wide spectrum of tasks and domains. In this work, we investigate a unified graph model, AnyGraph, designed to handle key challenges: i) Structure Heterogeneity, ii) Feature Heterogeneity, iii) Fast Adaptation, and iv) Scaling Law Emergence. **Keywords:** Graph Neural Networks, Generalization, Mixture-of-Experts.

Keywords: Graph Foundation Model, Mixture-of-Experts (MoE).

1 Introduction

Graph-structured data is ubiquitous in various domains, including social networks, biological systems, and academic research. However, existing graph learning models, particularly Graph Neural Networks (GNNs), often lack generalization capabilities and require extensive fine-tuning when applied across diverse datasets. Inspired by advancements in foundation models for vision and language, this paper introduces AnyGraph, a novel graph foundation model built upon a Mixture-of-Experts (MoE) architecture. AnyGraph addresses critical challenges, including structural and feature heterogeneity, fast adaptation to new datasets, and scaling law emergence, achieving exceptional zero-shot learning performance across 38 graph datasets.

2 Related works

Traditional GNNs, such as Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs), have achieved notable success in tasks like node classification and link prediction. However, these models struggle with cross-domain generalization due to their dependence on domain-specific training. Recent approaches, such as Graph Contrastive Learning (GraphCL) and pre-trained models like GraphGPT, attempt to bridge this gap but fail to address heterogeneity effectively. AnyGraph builds upon these foundations by introducing an MoE-based framework, offering superior generalization and adaptation capabilities.

3 Method

3.1 Overview

AnyGraph [1] utilizes a Mixture-of-Experts architecture with a dynamic expert routing mechanism. This enables the model to select the most suitable expert sub-network for handling specific graph properties, ensuring both cross-domain generalization and computational efficiency. As shown in Figure 1

3.2 Methodology of AnyGraph

AnyGraph is designed to handle graph heterogeneity and enable fast adaptation. It consists of the following key components:

1. **MoE Architecture**: AnyGraph employs a Mixture-of-Experts (MoE) architecture to address cross-domain graph heterogeneity. It consists of multiple graph expert models, each handling graphs with specific characteristics. An automated routing algorithm assigns input graph data to the most competent expert model for training and prediction. The routing mechanism measures the competence of expert models using self-supervised learning loss values and incorporates training frequency regularization to prevent a winner-takes-all situation.

2. **Graph Expert Routing Mechanism**: Inspired by graph self-supervised learning tasks, the routing mechanism calculates relatedness scores for positive and negative edges to determine the competence of expert models. Training frequency regularization is introduced to ensure all experts are utilized and to avoid suboptimal training outcomes.

3. **Fast Adaptation Capabilities**: With the MoE architecture and routing mechanism, AnyGraph's training and inference process is conducted by only one expert model, consuming only of the computational and memory resources required by other non-MoE graph foundation models. This enables fast adaptation when dealing with new datasets.

4. **Adaptive and Efficient Graph Experts**:

    (a) **Addressing In-domain Graph Heterogeneity**: Expert models in AnyGraph use a structure and feature unification process to handle graph data with different adjacency and feature dimensionalities. Singular value decomposition (SVD) is utilized for unified mapping, and high-order connectivity injection is applied to preserve multi-hop connection information.

    (b) **Efficient and Strong Feature Encoder**: Graph experts are configured with deep multi-layer perceptron (MLP) networks for efficient feature encoding. Each expert adopts a simple learnable network, and multiple experts work together to accelerate training and inference.

5. **Efficient Cross-domain Model Training**:

    (a) **Training Samples and Task**: Training samples from different datasets are mixed and shuffled for model training. Link prediction is used as the training task, and the loss function maximizes the prediction scores for positive samples and minimizes those for negative samples.

(b) **Feature and Structure Augmentation**: The training of AnyGraph involves periodic reprocessing of initial graph embeddings and graph routing results. SVD and simplified GCN processes are periodically reconducted for the embeddings, and the graph routing results are recalculated to perform structure augmentation. This enriches the training data and enhances the model's generalizability and robustness.

(c) **Complexity Analysis**: The complexity of AnyGraph's training and inference process is lower than that of other graph foundation models. The expert routing performs computations with a complexity comparable to simple GNNs, making AnyGraph more efficient in both training and inference.

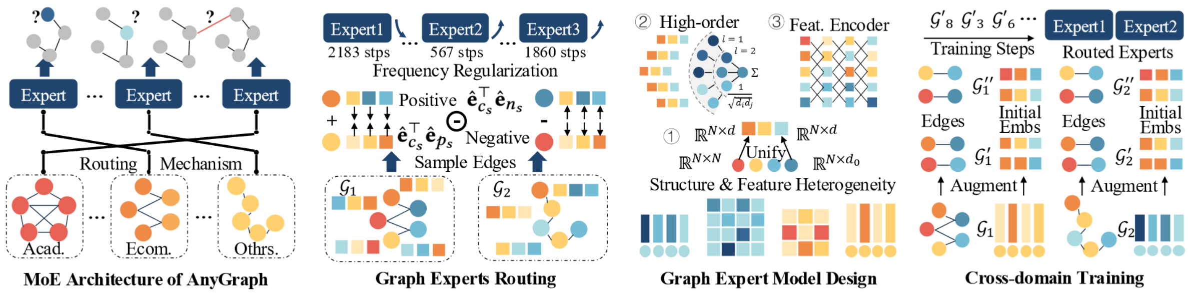

Figure 1. Overview of the AnyGraph

3.3 Feature extraction

The model employs a unified representation approach, where node features and adjacency matrices are processed using Singular Value Decomposition (SVD). This step generates fixed-dimensional embeddings that capture high-order structural and feature relationships.

3.4 Loss

A link prediction-based loss function is adopted to optimize the model, ensuring that positive relationships between nodes are reinforced while reducing the likelihood of false positives. To enhance generalization, AnyGraph incorporates feature augmentation and training frequency regularization.

$$\mathcal{L} = \sum_{S} \sum_{b \in B} -\frac{1}{B} \log \frac{\exp(\hat{y}_{c_b,p_b} - \hat{y}_{\max})}{\sum_{v_n \in \mathcal{V}_{G_s}} \exp(\hat{y}_{c_b,n} - \hat{y}_{\max})} \tag{1}$$

4 Implementation details

4.1 Comparing with the Released Models

AnyGraph significantly outperforms baseline methods, such as GraphCL, GPF, and traditional GNNs, in both zero-shot and few-shot settings. Unlike these models, AnyGraph's MoE-based architecture enables it to handle diverse graph structures and feature spaces effectively.

4.2 Experimental environment setup

Experiments were conducted across 38 datasets spanning domains such as e-commerce, academic networks, and biological systems. The datasets were divided into training and testing groups (e.g., Link1 and Link2. As show in Figure 2) to evaluate cross-domain generalization.And the result is shown in Figure 3

Group	Included Datasets
Link1	Products-tech, Yelp2018, Yelp-textfeat, Products-home, Steam-text, Amazon-text, Amazon-book, Citation-2019, Citation-20Century, Pubmed-link, Citeseer, OGB-PPA, P2P-Gnutella06, Soc-Epinions1, Email-Enron
Link2	Photo, Goodreads, Fitness, Movielens-1M, Movielens10M, Gowalla, Arxiv, Arxiv-t, Cora, CS, OGB-Collab, Proteins-0, Proteins- 1, Proteins-2, Proteins-3, OGB-DDI, Web-Stanford, RoadNet-PA
Ecommerce	Products-tech, Yelp2018, Yelp-textfeat, Products-home, Steam-text, Amazon-text, Amazon-book, Photo, Goodreads, Fitness, Movielens-1M, Movielens10M, Gowalla
Academic	Citation-2019, Citation-20Century, Pubmed-link, Citeseer, OGB-PPA, Arxiv, Arxiv-t, Cora, CS, OGB-Collab
Others	P2P-Gnutella06, Soc-Epinions1, Email-Enron, Proteins-0, Proteins- 1, Proteins-2, Proteins-3, OGB-DDI, Web-Stanford, RoadNet-PA
Node	Cora, Arxiv, Pubmed, Home, Tech

Figure 2. Interface

4.3 Interface design

AnyGraph's modular design simplifies integration into existing workflows. The routing mechanism automatically selects the most appropriate expert for each dataset, requiring minimal user intervention.

4.4 Main contributions

- **Mixture-of-Experts Architecture**: Enables dynamic expert selection to handle heterogeneous graph data.
- **Scalability**: Demonstrates scaling law properties, improving performance with increased data and parameters.
- **Fast Adaptation**: Adapts efficiently to new datasets without extensive fine-tuning.
- **Zero-Shot Generalization**: Achieves state-of-the-art results on unseen datasets.

Table 1: We evaluate the AnyGraph model (in zero-shot settings) and baseline models (with 5% and 10% training data) on link prediction (Recall@20, NDCG@20), node classification (Accuracy, Macro F1), and graph classification (Accuracy, Macro F1).

Data	GIN				GAT				GPF				GraphPrompt				GraphCL				**AnyGraph**		复现结果	
	Train 5%		Train 10%		Train 5%		Train 10%		Tune 5%		Tune 10%		Tune 5%		Tune 10%		Tune 5%		Tune 10%		0-shot			
Metric	Rec	NDCG	Rec	NDCG	Rec	NDCG	Rec	NDCG	Rec	NDCG	Rec	NDCG	Rec	NDCG	Rec	NDCG	Rec	NDCG	Rec	NDCG	Rec	NDCG	Rec	NDCG
Link1	6.46	3.06	11.80	5.45	13.52	6.65	13.45	6.78	6.04	2.92	6.80	3.27	4.33	2.24	5.42	3.11	17.23	9.00	20.55	10.76	**23.94**	**12.68**	24.23	12.91
Link2	6.72	4.50	21.62	13.41	9.83	5.91	15.30	8.84	7.44	4.25	16.58	9.84	6.06	3.36	6.10	3.62	29.18	17.62	31.42	19.91	**46.42**	**27.21**	45.88	27.25
Ecom.	3.36	2.58	13.41	8.06	3.79	2.94	9.64	5.78	7.25	3.84	18.72	10.94	4.90	2.59	6.06	3.36	22.13	13.19	26.05	14.59	**26.92**	**15.05**	22.86	13.97
Acad.	10.82	4.70	20.61	9.04	14.95	6.29	11.17	4.67	13.22	5.80	14.83	6.41	6.73	3.05	7.72	3.40	24.86	12.50	28.69	14.31	**32.74**	**15.31**	33.74	16.08
Othrs.	6.92	4.46	18.43	11.85	16.34	9.22	16.17	20.88	2.40	2.12	4.51	3.44	2.93	2.36	3.42	2.72	24.54	14.93	24.62	15.90	**46.83**	**28.97**		
Metric	Acc	MacF1	Acc	MacF1	Acc	MacF1	Acc	MacF1	Acc	MacF1	Acc	MacF1	Acc	MacF1	Acc	MacF1	Acc	MacF1	Acc	MacF1	Acc	MacF1		
Node	20.79	19.46	36.04	30.60	53.76	40.14	54.83	41.61	12.77	11.45	16.29	16.00	18.01	20.59	23.15	22.89	43.70	33.72	48.75	36.15	**64.31**	**43.24**		

Figure 3. Experimental results

5 Results and analysis

5.1 Zero-Shot Prediction

Superior Generalizability

- **Prediction Accuracy**: AnyGraph exhibits excellent zero-shot prediction accuracy in link prediction and node classification tasks when compared to existing GNN models, pre-training techniques, and foundation models.
- **Handling Heterogeneity**: It effectively manages structure-level and feature-level data heterogeneity through unified representations in expert models, thereby enabling comprehensive modeling across diverse graph data.
- **Comprehensive Training**: The extensive training on a variety of large-scale datasets endows AnyGraph with strong graph modeling and prediction capabilities.

Limitation of Existing Pre-training GNNs

- **Cross-Domain Transfer Challenges**: Existing pre-training and tuning methods such as GPF, GraphPrompt, and GraphCL face significant difficulties in dealing with distribution disparities in data domains, which restricts their effectiveness in knowledge transfer.
- **AnyGraph's Robust Adaptability**: AnyGraph's MoE architecture, which incorporates multiple graph expert models tailored to different sub-domains, effectively manages diverse datasets, thus demonstrating robust adaptability.

5.2 Scaling Law of AnyGraph Framework

Generalizability Follows the Scaling Law

As the model size and the amount of training data increase, AnyGraph's zero-shot prediction accuracy improves, while the full-shot performance saturates. This indicates that scaling up enhances the model's capabilities, likely due to the MoE architecture's ability to handle distribution disparities.

Emergent Abilities

The overall zero-shot performance curve reveals that with an increase in model size, AnyGraph's performance sometimes stagnates but then significantly improves, demonstrating emergent abilities and the effectiveness of scaling in generalization.

Impact of Training Data

Insufficient training data can initially have a negative impact on performance due to dataset differences. However, expanding the training data mitigates this issue, preventing overfitting and reducing bias.

5.3 Ablation Study

MoE's Role

The -MoE variant (a single expert model without the MoE architecture) shows a decline in performance in zero-shot prediction, highlighting the crucial role of the MoE architecture in enhancing AnyGraph's generalization.

Feature Modeling Importance

Omitting node features (-Feat variant) leads to a significant degradation in both zero-shot and full-shot performance, emphasizing the effectiveness of AnyGraph's unified structure and feature representation in handling in-domain graph heterogeneity.

Effectiveness of Other Components

Removing frequency regularization (-FreqReg) and graph augmentation (-Aug) negatively affects model training, confirming their beneficial impact on AnyGraph's robustness in handling diverse datasets.

5.4 Investigation on Expert Routing

Datasets with common characteristics are routed to similar expert models by AnyGraph's routing mechanism, as observed in datasets sharing feature construction methods or sources. This demonstrates the mechanism's efficacy in identifying appropriate experts and reveals graph-wise relatedness.

5.5 Efficiency Study

Tuning Curve Comparison

Pre-trained AnyGraph rapidly reaches a high performance saturation when fine-tuned on new datasets, outperforming GraphCL and GCN. This is attributed to its strong cross-domain generalization and the efficiency of the MoE architecture.

Training Time Comparison

AnyGraph has comparable or even lower training times than GCN and GraphCL, despite having more parameters. It avoids full-graph propagation and utilizes structure-aware embeddings, reducing time and memory requirements. The MoE architecture further reduces computational costs.

6 Conclusion and future work

AnyGraph represents a significant advancement in graph foundation models, addressing heterogeneity and scaling challenges while achieving exceptional generalization across domains. Future work will focus on integrating multimodal data, exploring even larger datasets, and refining the model's efficiency to further reduce computational requirements.

References

[1] Lianghao Xia and Chao Huang. Anygraph: Graph foundation model in the wild. *arXiv preprint arXiv:2408.10700*, 2024.