2023 Australasian Database Conference
ADC2023
Melbourne, AU

MONASH University

Griffith UNIVERSITY

Towards Data-centric Graph Machine Learning

Xin Zheng[1], Shirui Pan[2]

[1] Monash University

[2] Griffith University

Tutorial Outline

Time	Topic	Part	Presenters
15 min	**Introduction & Overview** • Why Data-centric Graph Machine Learning (DC-GML) • Framework of DC-GML	Part 1	*Shirui Pan*
30 min	**Frontiers of Graph Data Enhancement** • Graph Structure Enhancement • Graph Feature Enhancement • Graph Label Enhancement • Graph Size Enhancement	Part 2	*Shirui Pan*
20 min	**Frontiers of Graph Data Exploitation** • Overview of Graph Data Exploitation Strategies • Graph Self-supervised Learning	Part 3	*Xin Zheng*
10 min	**Frontiers of Graph Data-centric MLOps** • Overview of Graph MLOps • GNN Evaluation On OOD Graph Data	Part 4	*Xin Zheng*
10 min	**Future Directions & Conclusion**	Part 5	*Xin Zheng*

Part 1: Introduction & Overview

- Why Data-centric Graph Machine Learning (DC-GML)
- Overview of DC-GML Framework

AI System

AI system = Code + Data

(model/algorithm)

What is data-centric AI?

“ Data-centric AI (DCAI) is the discipline of systematically engineering the data used to build an AI system.” – –Andrew Ng

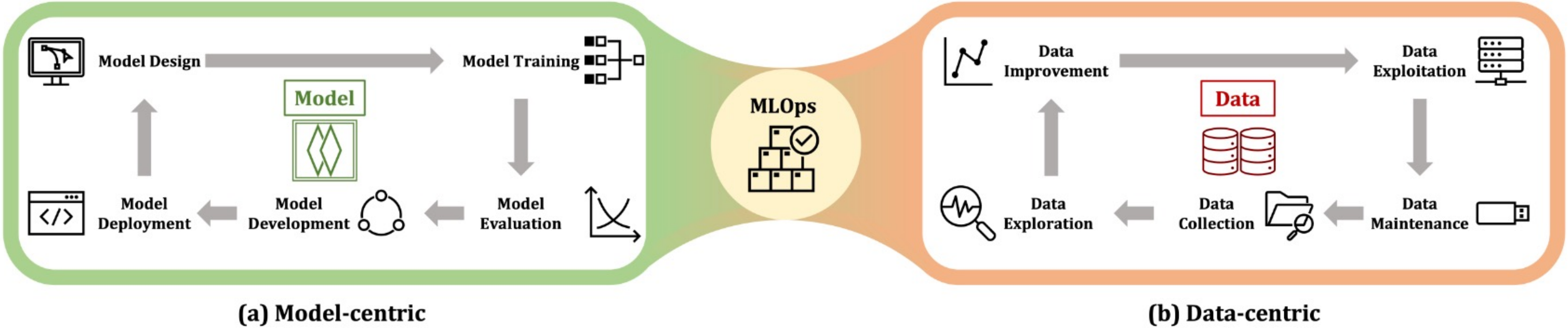
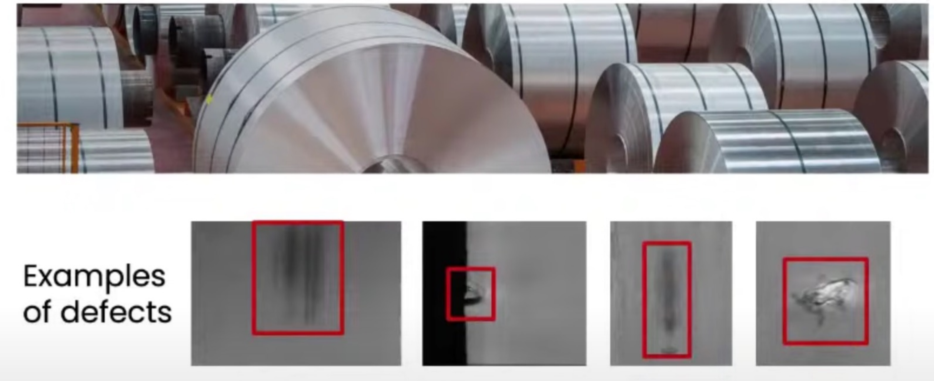


Fig. 1. General comparison between (a) model-centric AI and (b) data-centric AI.

Why data-centric AI matters

An example:

Inspecting steel sheets for defects



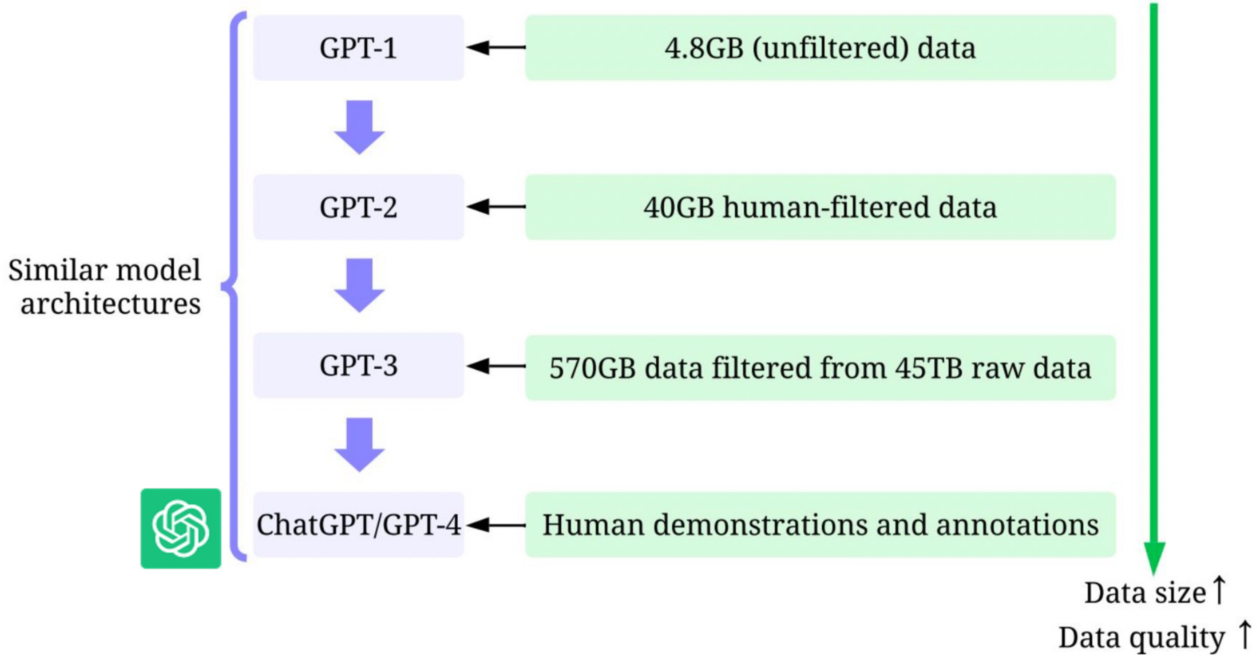
	Steel defect detection	Solar panel	Surface inspection
Baseline	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

Data-centric improves more than model-centric!

[1] A Chat with Andrew on MLOps: From Model-centric to Data-centric AI: https://www.youtube.com/watch?v=06-AZXmwHjo

Why data-centric AI matters

When model design becomes mature, the significance of both the size and quality of the data increases.



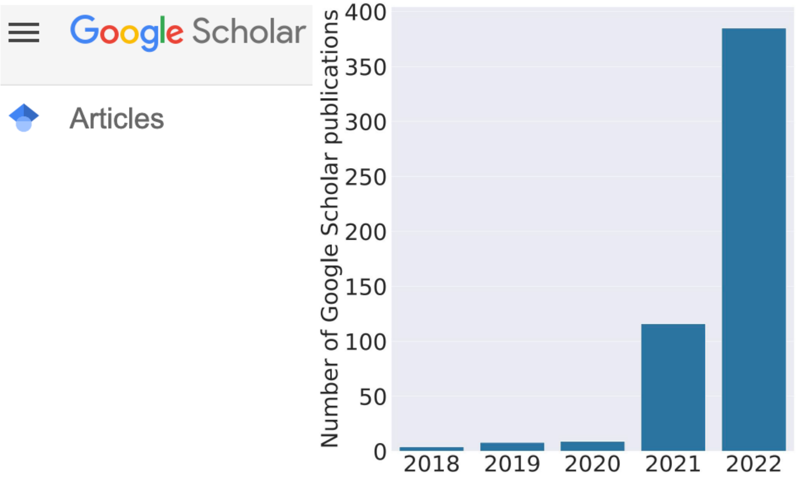
❖ **Core idea:**

Engineering data to enable great **“availability and quality”** for serving and promoting model-related ML tasks.

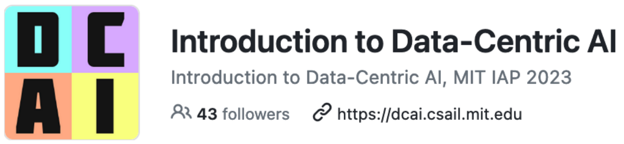
[1] Zha, Daochen, et al. Data-centric Artificial Intelligence: A Survey. arXiv, 2023.

Data-centric AI is attracting attentions…

- **Exponentially growing DCAI research papers**



- **DCAI Courses, Workshops, Competitions**



- **AI Startups**

LANDING AI Snorkel scale

[1] Zha, Daochen, et al. Data-centric AI: Perspectives and Challenges. SDM, 2023.

Graphs: A typical & vital instantiation in DCAI

Example: A Social Network Graph

A Graph has nodes/vertices and edges:

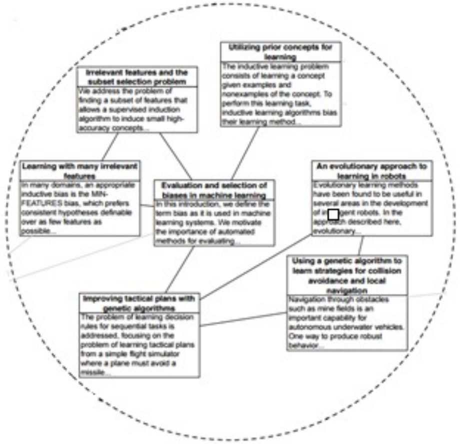
- Nodes/vertices → a person in the social network
- Edges → Connection between people

Graphs have the ability of:

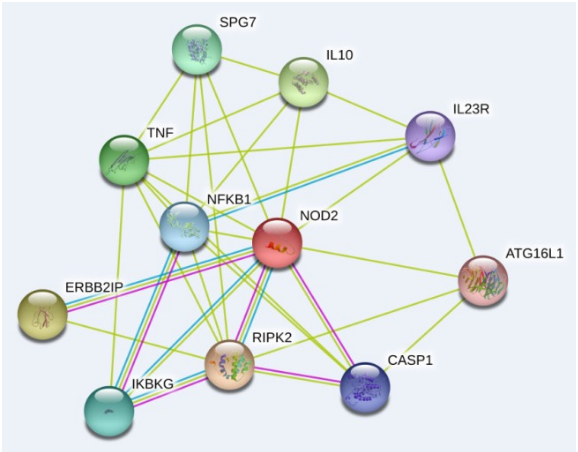
- Representing complex structural relationships among massive diverse entities in the real world

Graphs in real-world applications

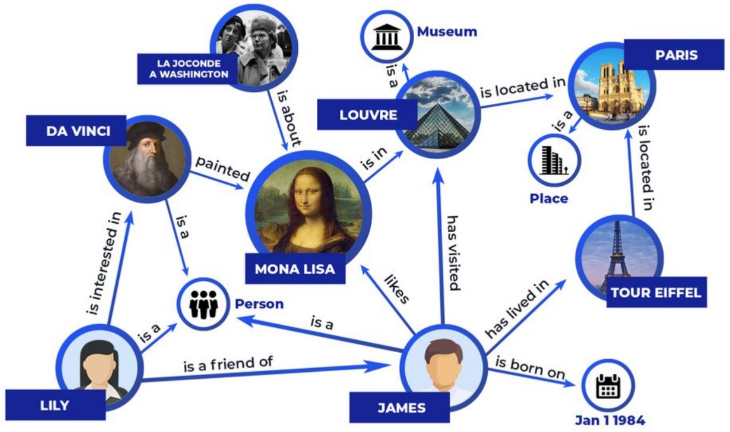
Social Networks



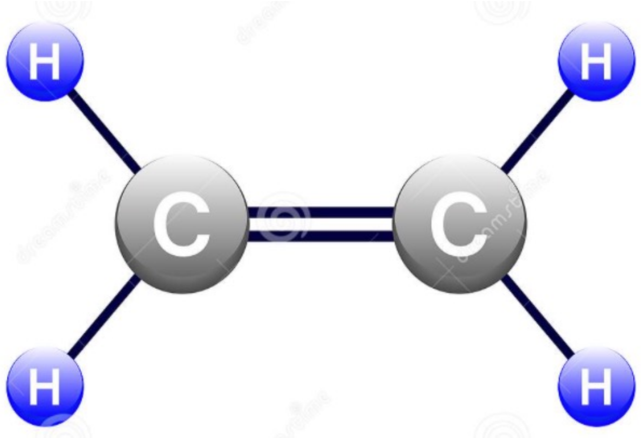
Bibliography Networks



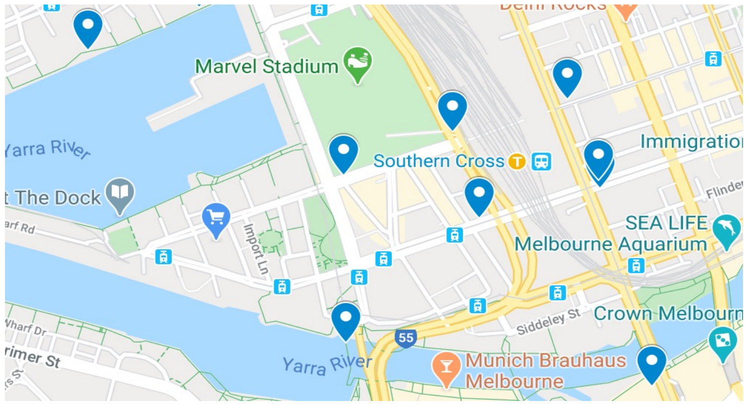
Protein Interaction Networks



Knowledge Graphs



Chemical Compounds

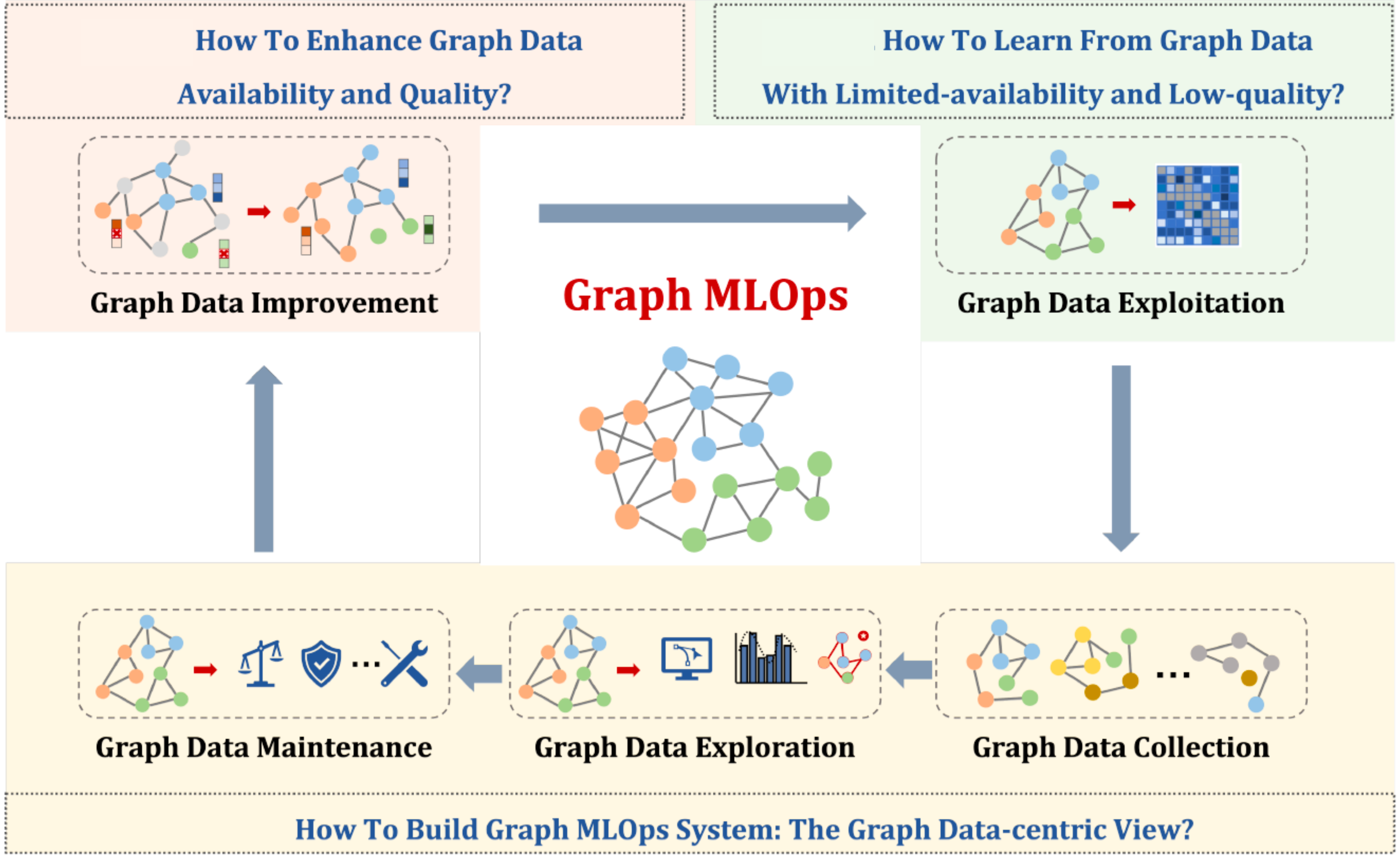


Traffic Networks

Towards data-centric **graph** machine learning

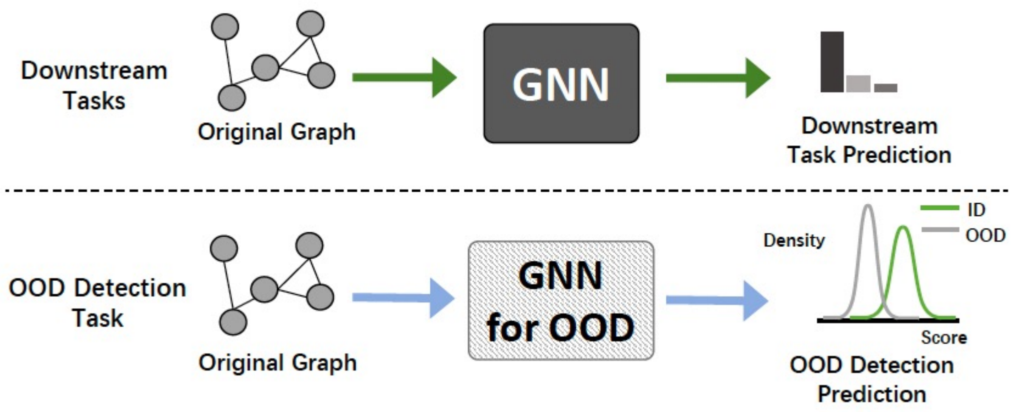
Data-centric graph machine learning (DC-GML) aims to:

- Process, analyze, and understand graph data in entire lifecycle
- Enhancing the quality
- Uncovering the insights
- Developing comprehensive representations
- Working collaboratively with graph ML models under graph MLOps

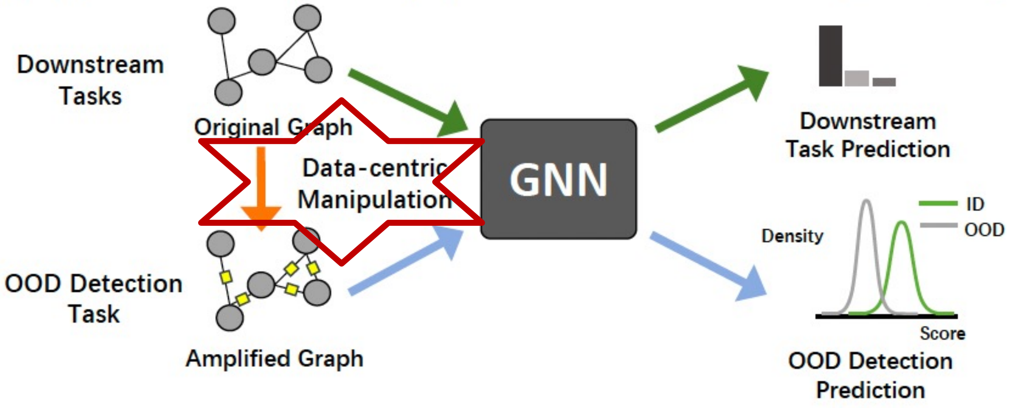


Why data-centric GML matters

❖ **Taking graph OOD detection as example:**



(a) Typical retraining-based graph OOD detection methods



(b) Our proposed data-centric framework for graph OOD detection.

ID	OOD	Metric	GCL_S	GCL_S+	Improv.
ENZYMES	PROTEIN	AUC ↑	62.97	**73.76**	+17.14%
		AUPR ↑	62.47	**75.27**	+20.49%
		FPR95 ↓	93.33	**88.33**	-5.36%
IMDBM	IMDBB	AUC ↑	80.52	**83.84**	+4.12%
		AUPR ↑	74.43	**80.16**	+7.70%
		FPR95 ↓	38.67	**38.33**	-0.88%
BZR	COX2	AUC ↑	75.00	**97.31**	+29.75%
		AUPR ↑	62.41	**97.17**	+55.70%
		FPR95 ↓	47.50	**15.00**	-68.42%

Model-centric GML method

✩ **Data-centric GML method and improvements**

Guo, Y., Yang, C., Chen, Y., Liu, J., Shi, C., & Du, J. (2023). A Data-centric Framework to Endow Graph Neural Networks with Out-Of-Distribution Detection Ability.

Overview of DC-GML Framework

Graph Data Improvement

- **Graph Structure Enhancement**
  - *Graph Structure Learning || Graph Sparsification || Graph Diffusion*
- **Graph Feature Enhancement**
  - *Graph Feature Completion || Graph Feature Denoising*
- **Graph Label Enhancement**
  - *Graph Pseudo-labelling || Graph Label Denoising || Graph Class-imbalanced Sampling*
- **Graph Size Enhancement**
  - *Graph Size Reduction || Graph Data Augmentation*

Graph Data Exploitation

- **Graph Self-supervised Learning**
- **Graph Semi-supervised Learning**
- **Graph Active Learning**
- **Graph Transfer Learning**

Graph Data Collection

- **Graph Data Crowdsourcing & Synthesis**

Graph Data Exploration

- **Graph Data Understanding, Visualization & Valuation**

Graph Data Maintenance

- **Graph Data Privacy & Security**

Resources

❖ **More resources and details in our work**

- Survey paper: Towards Data-centric Graph Machine Learning: Review and Outlook
- Github collection: https://github.com/Data-Centric-GraphML/awesome-papers

Data-centric Graph ML
Review & Outlook

DC-GML GitHub Collection

Part 2: Frontiers of Graph Data Enhancement

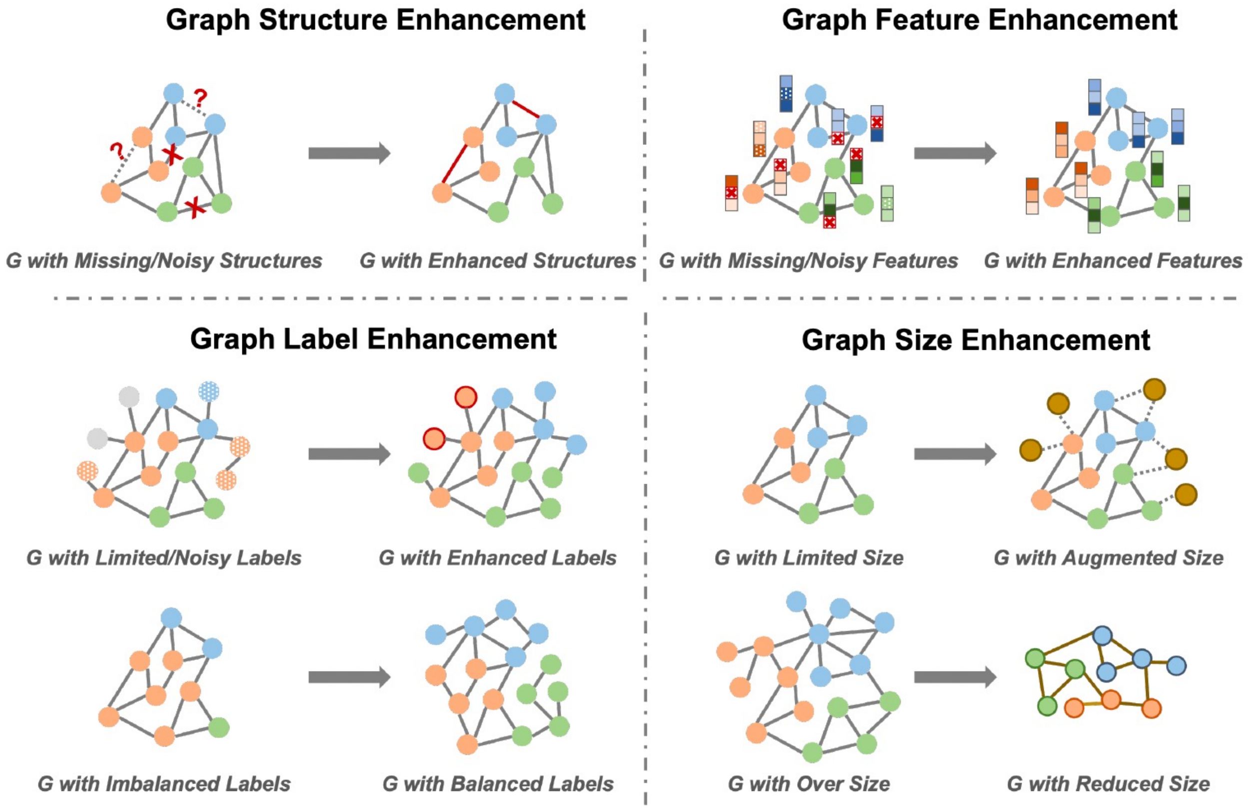
Overview of Graph Data Enhancement

❖ **Core Strategy**

aim to synthesize or modify graph data itself to improve availability and quality by comprehensively fixing potential issues of graph data.

Given a graph $\mathbf{G} = (\mathbf{A}, \mathbf{X}, \mathbf{Y})$, with several essential components of :

1) graph structure $\mathbf{A}$;
2) node/edge attribute features $\mathbf{X}$;
3) node/graph annotated labels $\mathbf{Y}$;
4) the holistic graph $\mathbf{G}$ related scale



Outline for Graph Data Enhancement

- ❖ **Overview of Graph Data Enhancement**
- ❖ **Techniques with Case Studies :**
  - **Graph Structure Enhancement**
  - Graph Feature Enhancement
  - Graph Label Enhancement
  - Graph Size Enhancement

Graph Structure Enhancement

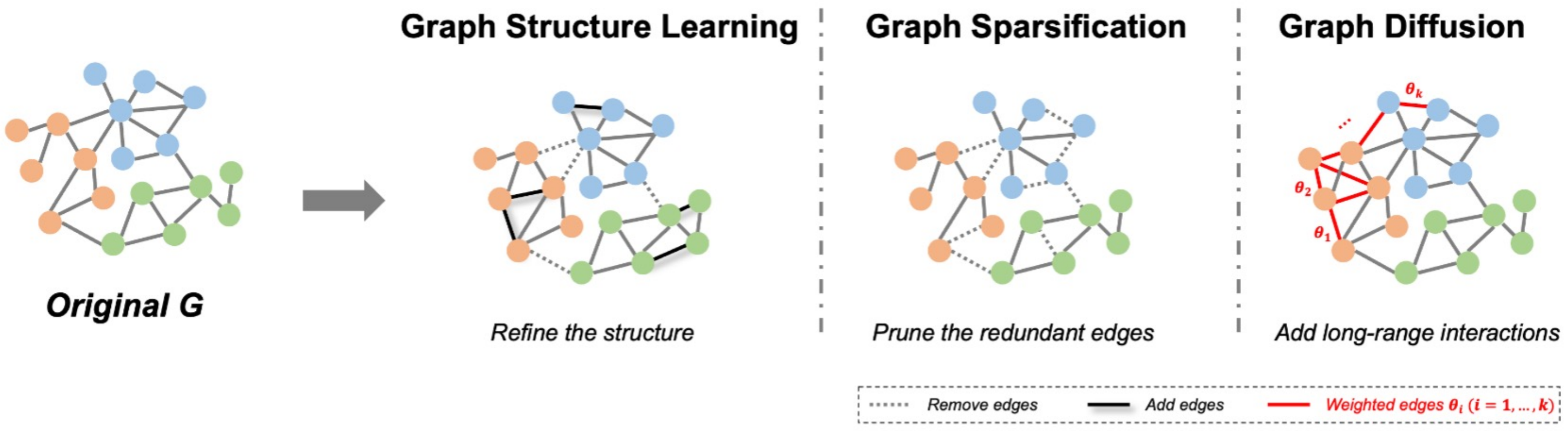


Fig. 5. Illustration of graph structure enhancement methods.

- **Graph Structure Learning:** add, remove, and reweight the edges on noisy or incomplete structures
- **Graph Sparsification:** prune the redundant edges to avoid over-dense structures
- **Graph Diffusion:** establish links with global and long-range structural interactions

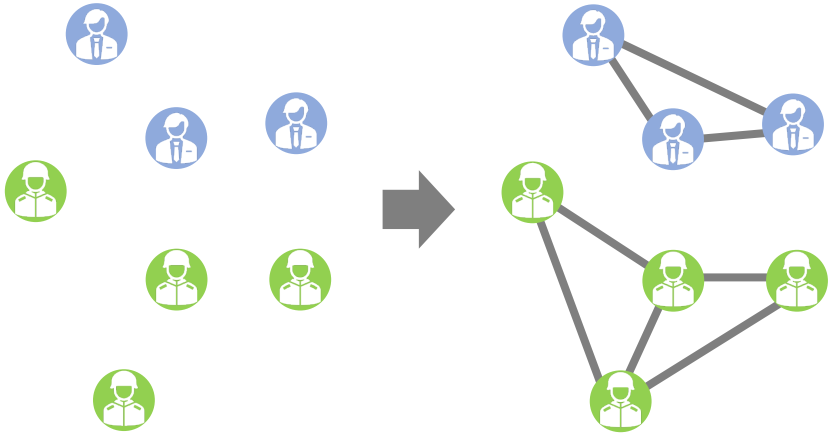
Towards Unsupervised Deep Graph Structure Learning

--Case Study on Graph Structure Enhancement

❖ **Graph structure learning (GSL)**: learning graph structure from data when structure is

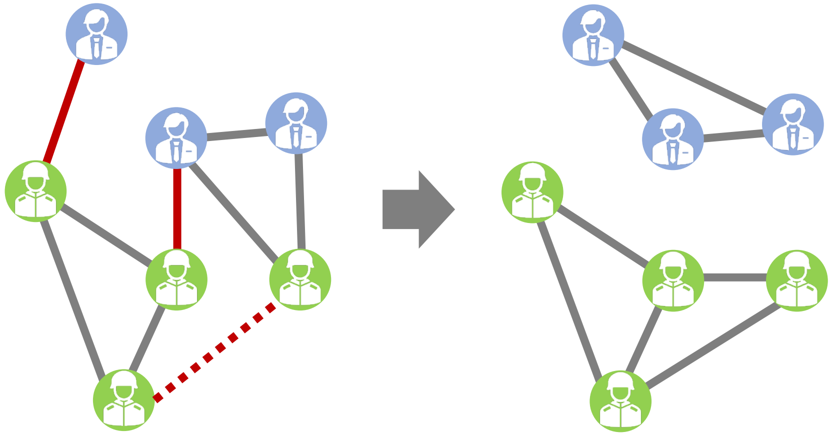
missing or unreliable

Structure inference:
Learning from non-structured data



Non-structured data   Learned structure

Structure refinement:
Learning with structure-noisy graph data



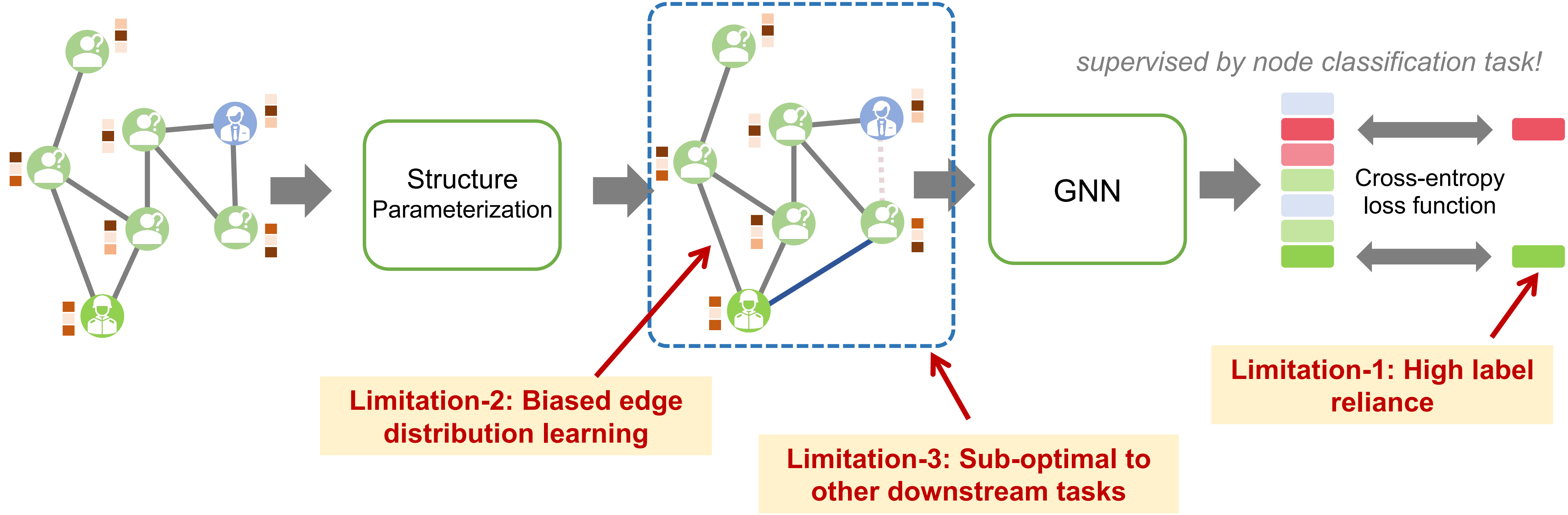
Noisy graph structure   Learned structure

Liu, Y., Zheng, Y., Zhang, D., Chen, H., Peng, H., & Pan, S. (2022, April). Towards unsupervised deep graph structure learning. In Proceedings of the ACM Web Conference (WWW). 2022 (pp. 1392-1403).

Towards Unsupervised Deep Graph Structure Learning

--Case Study on Graph Structure Enhancement

❖ **Existing methods: Supervised graph structure learning**



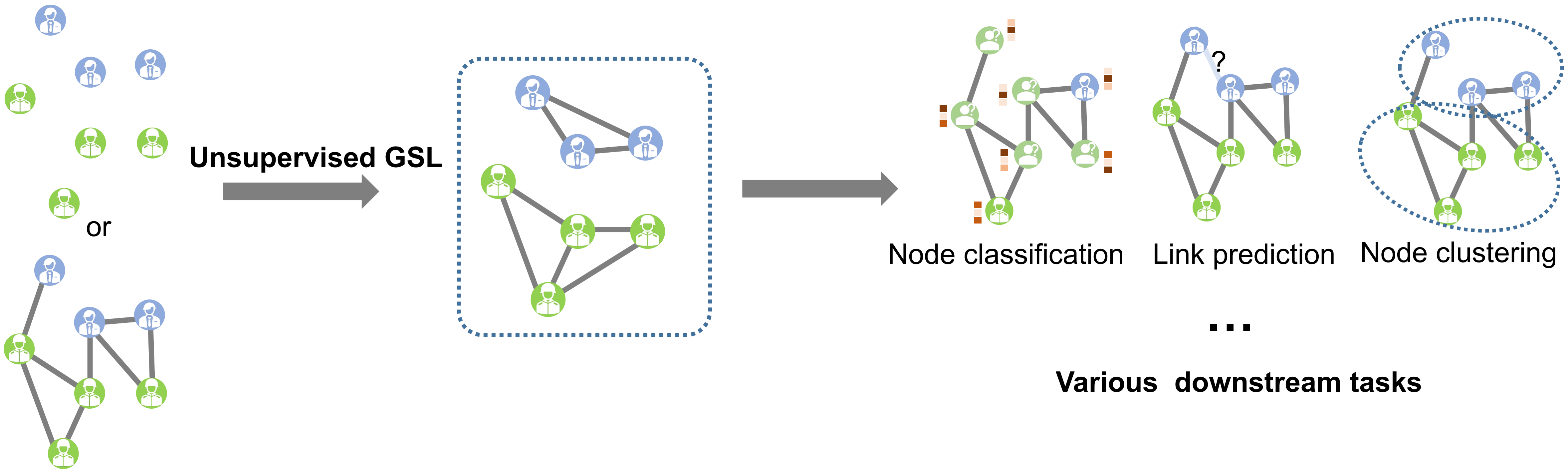
Liu, Y., Zheng, Y., Zhang, D., Chen, H., Peng, H., & Pan, S. (2022, April). Towards unsupervised deep graph structure learning. In Proceedings of the ACM Web Conference (WWW). 2022 (pp. 1392-1403).

Towards Unsupervised Deep Graph Structure Learning

--Case Study on Graph Structure Enhancement

❖ **More practical scenario: Unsupervised graph structure learning**

aim to optimize the graph structure as an independent task and without label-based supervision.

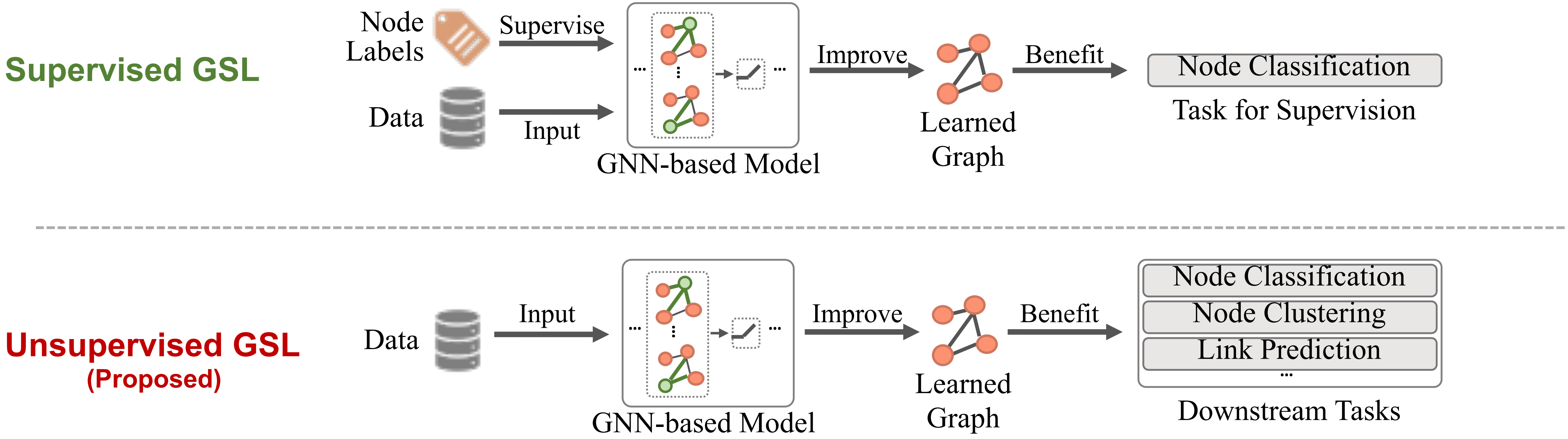


Liu, Y., Zheng, Y., Zhang, D., Chen, H., Peng, H., & Pan, S. (2022, April). Towards unsupervised deep graph structure learning. In Proceedings of the ACM Web Conference (WWW). 2022 (pp. 1392-1403).

Towards Unsupervised Deep Graph Structure Learning

--Case Study on Graph Structure Enhancement

❖ Comparison: Supervised GSL *vs.* Unsupervised GSL



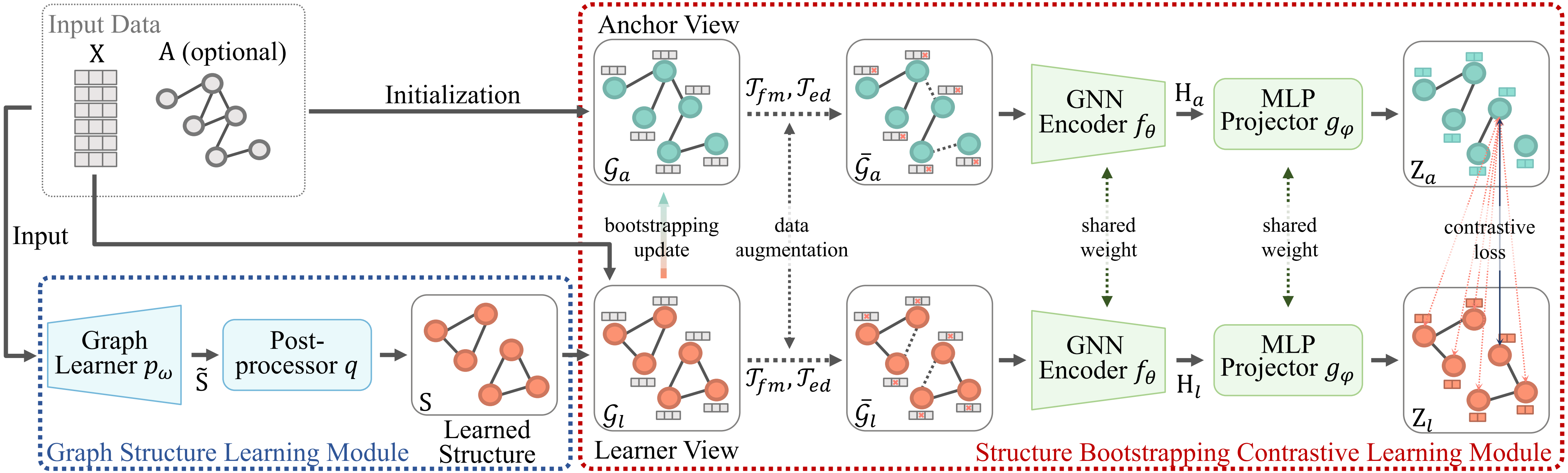
Advantages of UGSL: ☑ Does not rely on labels ☑ Unbiased learning ☑ Task-agnostic

Liu, Y., Zheng, Y., Zhang, D., Chen, H., Peng, H., & Pan, S. (2022, April). Towards unsupervised deep graph structure learning. In Proceedings of the ACM Web Conference (WWW). 2022 (pp. 1392-1403).

Towards Unsupervised Deep Graph Structure Learning

--Case Study on Graph Structure Enhancement

❖ **Proposed framework - SUBLIME**



To model and regularize the learned graph topology.

To provide a self-optimized supervision signal for GSL.

Liu, Y., Zheng, Y., Zhang, D., Chen, H., Peng, H., & Pan, S. (2022, April). Towards unsupervised deep graph structure learning. In Proceedings of the ACM Web Conference (WWW). 2022 (pp. 1392-1403).

Towards Unsupervised Deep Graph Structure Learning

--Case Study on Graph Structure Enhancement

❖ **SUBLIME Performance on Node classification @ Structure Inference**

Available Data for GSL	Method	Dataset							
		Cora	Citeseer	Pubmed	ogbn-arxiv	Wine	Cancer	Digits	20news
-	LR	60.8±0.0	62.2±0.0	72.4±0.0	52.5±0.0	92.1±1.3	93.3±0.5	85.5±1.5	42.7±1.7
-	Linear SVM	58.9±0.0	58.3±0.0	72.7±0.1	51.8±0.0	93.9±1.6	90.6±4.5	87.1±1.8	40.3±1.4
-	MLP	56.1±1.6	56.7±1.7	71.4±0.0	54.7±0.1	89.7±1.9	92.9±1.2	36.3±0.3	38.6±1.4
-	GCN_{knn} [22]	66.5±0.4	68.3±1.3	70.4±0.4	54.1±0.3	93.2±3.1	83.8±1.4	91.3±0.5	41.3±0.6
-	GAT_{knn} [40]	66.2±0.5	70.0±0.6	69.6±0.5	OOM	91.5±2.4	95.1±0.8	91.4±0.1	45.0±1.2
-	$SAGE_{knn}$ [15]	66.1±0.7	68.0±1.6	68.7±0.2	55.2±0.4	87.4±0.8	93.7±0.3	91.6±0.7	45.4±0.4
X, Y	LDS [12]	71.5±0.8	71.5±1.1	OOM	OOM	97.3±0.4	94.4±1.9	92.5±0.7	46.4±1.6
X, Y, A_{knn}	GRCN [53]	69.6±0.2	70.4±0.3	70.6±0.1	OOM	96.6±0.4	95.4±0.6	92.8±0.2	41.8±0.2
X, Y, A_{knn}	Pro-GNN [20]	69.2±1.4	69.8±1.7	OOM	OOM	95.1±1.5	96.5±0.1	93.9±1.9	45.7±1.4
X, Y, A_{knn}	GEN [45]	69.1±0.7	70.7±1.1	70.7±0.9	OOM	96.9±1.0	96.8±0.4	94.1±0.4	47.1±0.3
X, Y	IDGL [7]	70.9±0.6	68.2±0.6	70.1±1.3	55.0±0.2	98.1±1.1	95.1±1.0	93.2±0.9	48.5±0.6
X, Y	SLAPS [11]	**73.4±0.3**	72.6±0.6	**74.4±0.6**	**56.6±0.1**	96.6±0.4	96.6±0.2	**94.4±0.7**	**50.4±0.7**
A_{knn}	GDC [23]	68.1±1.2	68.8±0.8	68.4±0.4	OOM	96.1±1.0	95.9±0.4	92.6±0.5	46.4±0.9
X	SLAPS-2s [11]	72.1±0.4	69.4±1.4	71.1±0.5	54.2±0.2	96.2±2.1	95.9±1.2	93.6±0.8	47.7±0.7
X	**SUBLIME**	73.0±0.6	**73.1±0.3**	73.8±0.6	55.5±0.1	**98.2±1.6**	**97.2±0.2**	94.3±0.4	49.2±0.6

Liu, Y., Zheng, Y., Zhang, D., Chen, H., Peng, H., & Pan, S. (2022, April). Towards unsupervised deep graph structure learning. In Proceedings of the ACM Web Conference (WWW). 2022 (pp. 1392-1403).

Towards Unsupervised Deep Graph Structure Learning

--Case Study on Graph Structure Enhancement

❖ SUBLIME Performance

- Node classification @ structure refinement

Available Data for GSL	Method	Cora	Citeseer	Pubmed	ogbn-arxiv
-	GCN	81.5	70.3	79.0	71.7±0.3
-	GAT	83.0±0.7	72.5±0.7	79.0±0.3	OOM
-	SAGE	77.4±1.0	67.0±1.0	76.6±0.8	71.5±0.3
X, Y, A	LDS	83.9±0.6	**74.8±0.3**	OOM	OOM
X, Y, A	GRCN	84.0±0.2	73.0±0.3	78.9±0.2	OOM
X, Y, A	Pro-GNN	82.1±0.4	71.3±0.4	OOM	OOM
X, Y, A	GEN	82.3±0.4	73.5±1.5	80.9±0.8	OOM
X, Y, A	IDGL	84.0±0.5	73.1±0.7	**83.0±0.2**	**72.0±0.3**
A	GDC	83.6±0.2	73.4±0.3	78.7±0.4	OOM
X, A	SUBLIME	**84.2±0.5**	73.5±0.6	81.0±0.6	71.8±0.3

- Node clustering @ structure refinement

Method	Cora C-ACC	Cora NMI	Cora F1	Cora ARI	Citeseer C-ACC	Citeseer NMI	Citeseer F1	Citeseer ARI
K-means	50.0	31.7	37.6	23.9	54.4	31.2	41.3	28.5
SC	39.8	29.7	33.2	17.4	30.8	9.0	25.7	8.2
GE	30.1	5.9	23.0	4.6	29.3	5.7	21.3	4.3
DW	52.9	38.4	43.5	29.1	39.0	13.1	30.5	13.7
DNGR	41.9	31.8	34.0	14.2	32.6	18.0	30.0	4.3
M-NMF	42.3	25.6	32.0	16.1	33.6	9.9	25.5	7.0
RMSC	46.6	32.0	34.7	20.3	51.6	30.8	40.4	26.6
TADW	53.6	36.6	40.1	24.0	52.9	32.0	43.6	28.6
VGAE	59.2	40.8	45.6	34.7	39.2	16.3	27.8	10.1
ARGA	64.0	44.9	61.9	35.2	57.3	35.0	54.6	34.1
MGAE	68.1	48.9	53.1	**56.5**	66.9	41.6	52.6	42.5
AGC	68.9	53.7	65.6	44.8	67.0	41.1	62.5	41.5
DAEGC	70.4	52.8	**68.2**	49.6	67.2	39.7	**63.6**	41.0
SUBLIME	**71.3**	**54.2**	63.5	50.3	**68.5**	**44.1**	63.2	**43.9**

Liu, Y., Zheng, Y., Zhang, D., Chen, H., Peng, H., & Pan, S. (2022, April). Towards unsupervised deep graph structure learning. In Proceedings of the ACM Web Conference (WWW). 2022 (pp. 1392-1403).

Outline for Graph Data Enhancement

- ❖ **Overview of Graph Data Enhancement**
- ❖ **Techniques with Case Studies :**
  - Graph Structure Enhancement
  - **Graph Feature Enhancement**
  - Graph Label Enhancement
  - Graph Size Enhancement

Graph Feature Enhancement

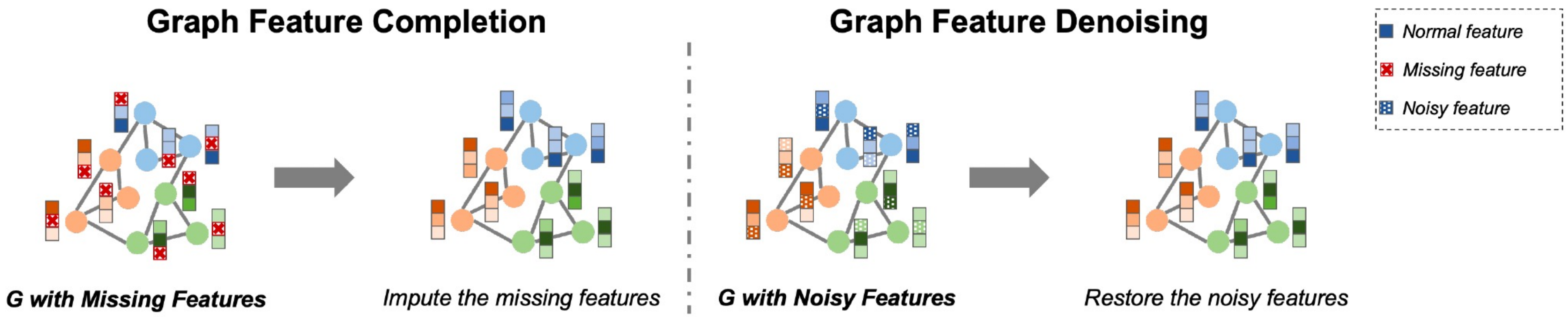


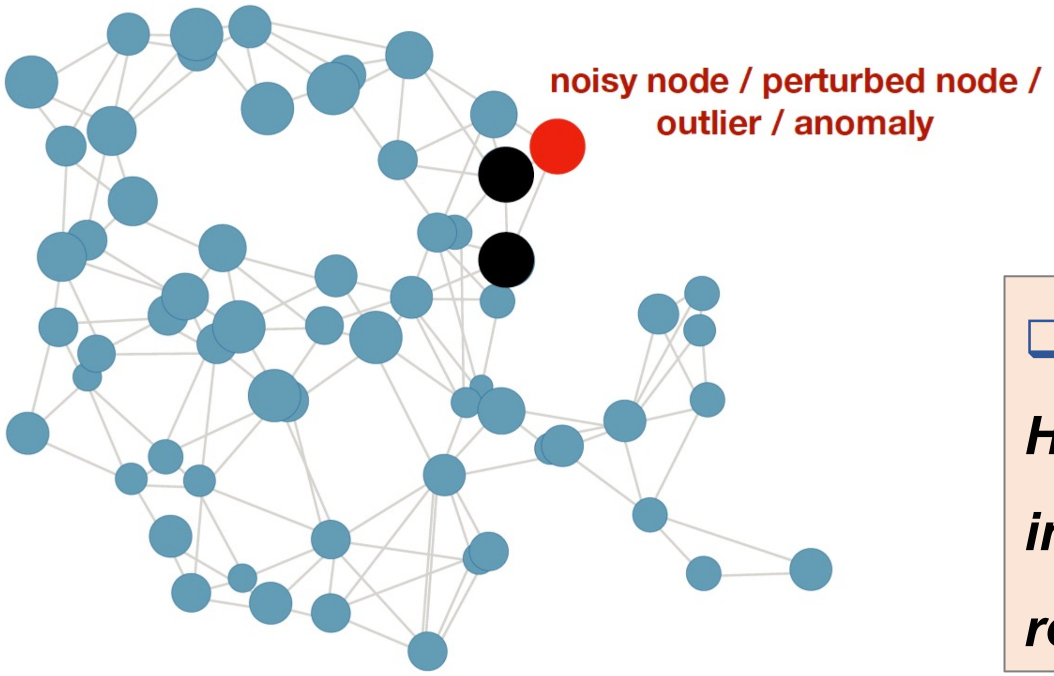
Fig. 6. Illustration of graph feature enhancement methods.

- **Graph Feature Completion:** focuses on imputing the missing features
- **Graph Feature Denoising:** refining the noisy features.

Robust Graph Representation Learning for Local Corruption Recovery

-- Case study on Graph Feature Completion

❖ **Graph node noise exists widely**



❑ **The question is:**

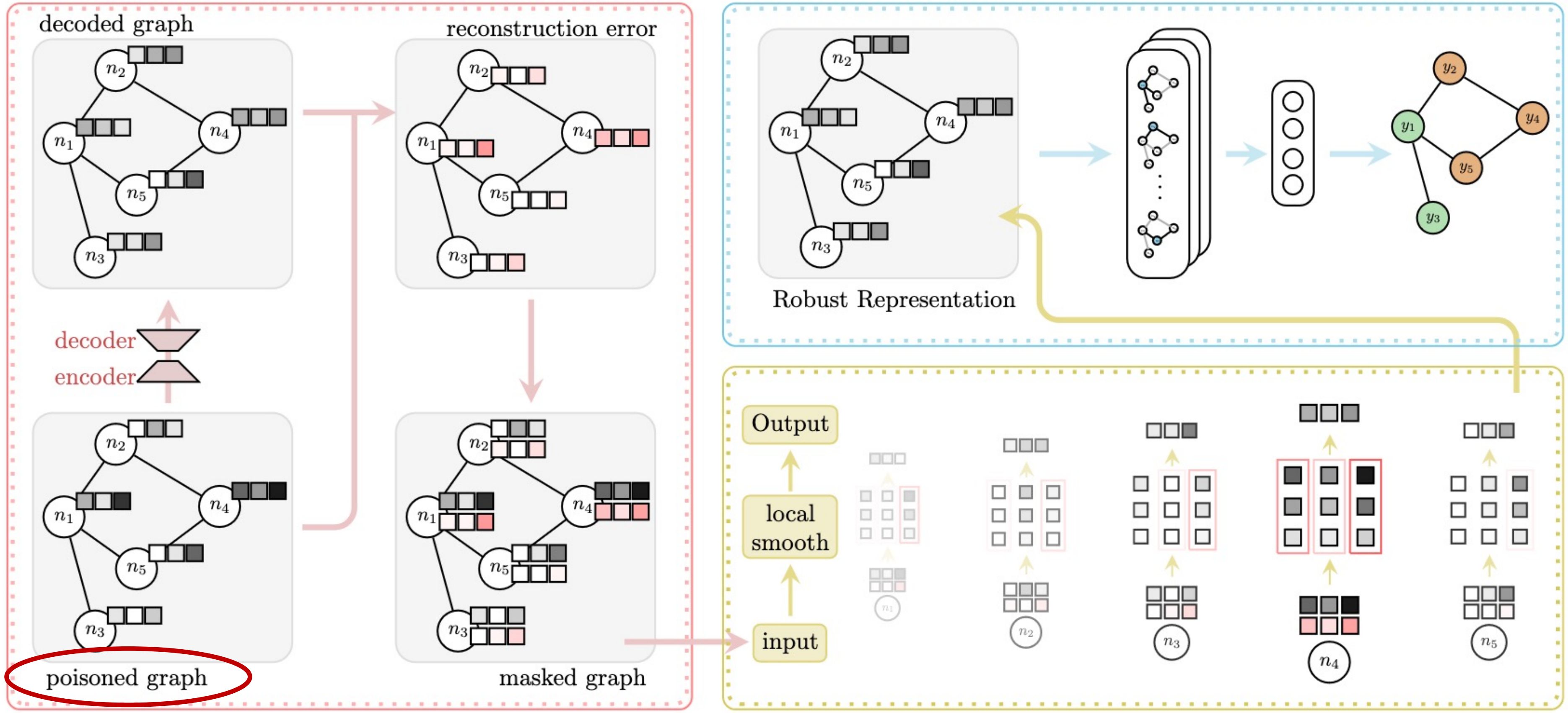
How to eliminate undesirable corruptions the input node attributes to enhance graph representation learning?

Zhou, B., Jiang, Y., Wang, Y., Liang, J., Gao, J., Pan, S., & Zhang, X. (2023, April). Robust graph representation learning for local corruption recovery. In Proceedings of the ACM Web Conference (WWW). 2023 (pp. 438-448).

Robust Graph Representation Learning for Local Corruption Recovery

-- Case study on Graph Feature Completion

❖ Framework of the proposed MAGNET

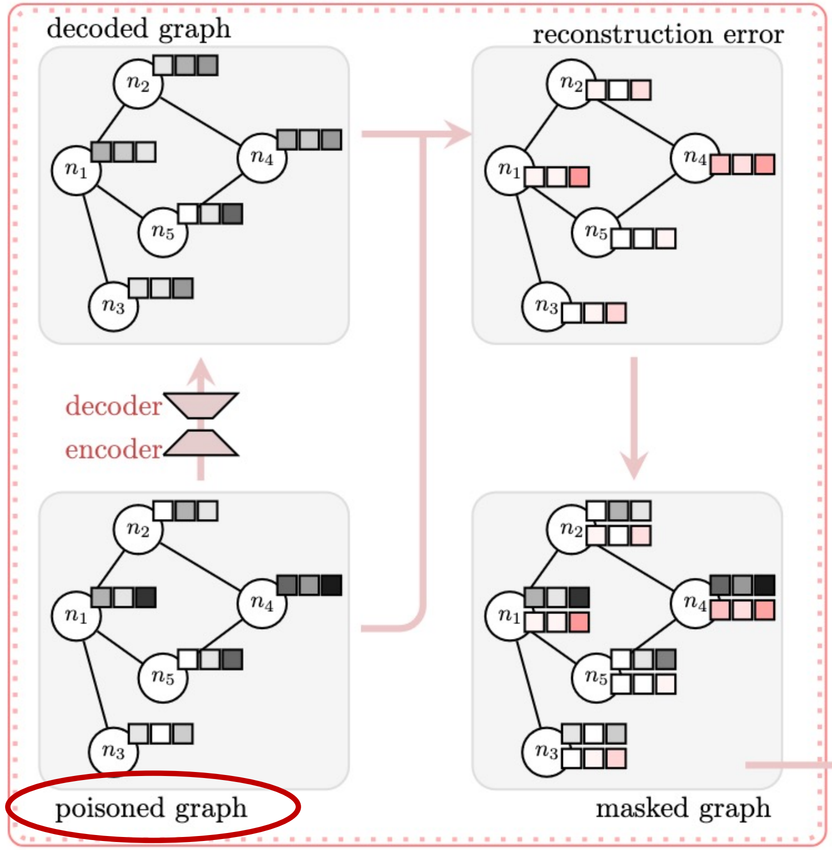


Zhou, B., Jiang, Y., Wang, Y., Liang, J., Gao, J., Pan, S., & Zhang, X. (2023, April). Robust graph representation learning for local corruption recovery. In Proceedings of the ACM Web Conference (WWW). 2023 (pp. 438-448).

Robust Graph Representation Learning for Local Corruption Recovery

-- Case study on Graph Feature Completion

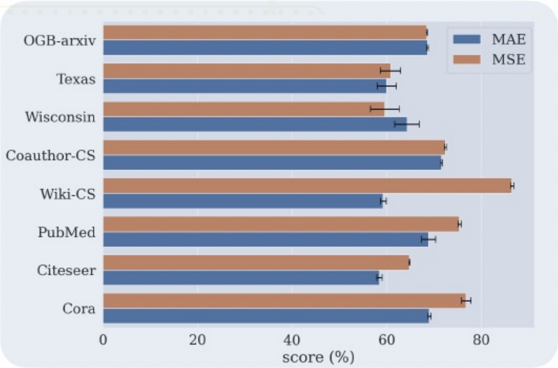
- **First, mask matrix (M) generation**



$$\min_{\mathbf{U},\mathbf{Z}} \|\nu \mathbf{Z}\|_{p,G} + \frac{1}{2}\|\mathbf{M} \odot (\mathbf{U}-\mathbf{X})\|_{q,G}^{q}, \quad \text{s.t. } \mathbf{Z} = \mathcal{W}\mathbf{U}$$

$$\text{where } \mathbf{M} = 1 - \text{threshold}(\|\mathbf{X}-\mathbf{X}'\|_1, \tau)$$

remark. $\mathbf{X}'$ is the reconstructed signal with a trainable **GAE** with MSE loss



Zhou, B., Jiang, Y., Wang, Y., Liang, J., Gao, J., Pan, S., & Zhang, X. (2023, April). Robust graph representation learning for local corruption recovery. In Proceedings of the ACM Web Conference (WWW). 2023 (pp. 438-448).

Robust Graph Representation Learning for Local Corruption Recovery

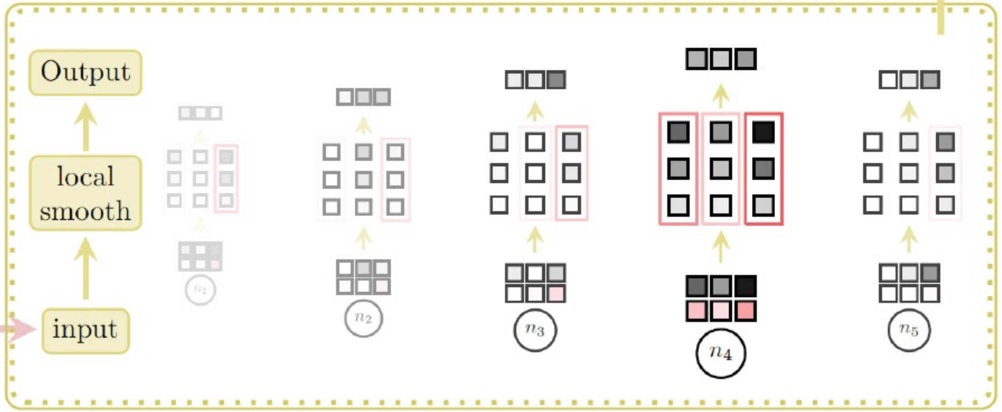
-- Case study on Graph Feature Completion

- **Next, find a robust signal representation**

This forms a standard formulation of problems that can be solved by ADMM. The associated augmented Lagrangian with respect to the objective function reads

$$\mathcal{L}(\mathbf{U},\mathbf{Z};\mathbf{Y}) := \|\nu\mathbf{Z}\|_{p,G} + \frac{1}{2}\|\mathbf{M}\odot(\mathbf{U}-\mathbf{X})\|_{q,G}^{q} + \langle \mathbf{Y}, \mathcal{W}\mathbf{U}-\mathbf{Z}\rangle + \frac{\gamma}{2}\|\mathcal{W}\mathbf{U}-\mathbf{Z}\|^{2}$$

where $\gamma > 0$.

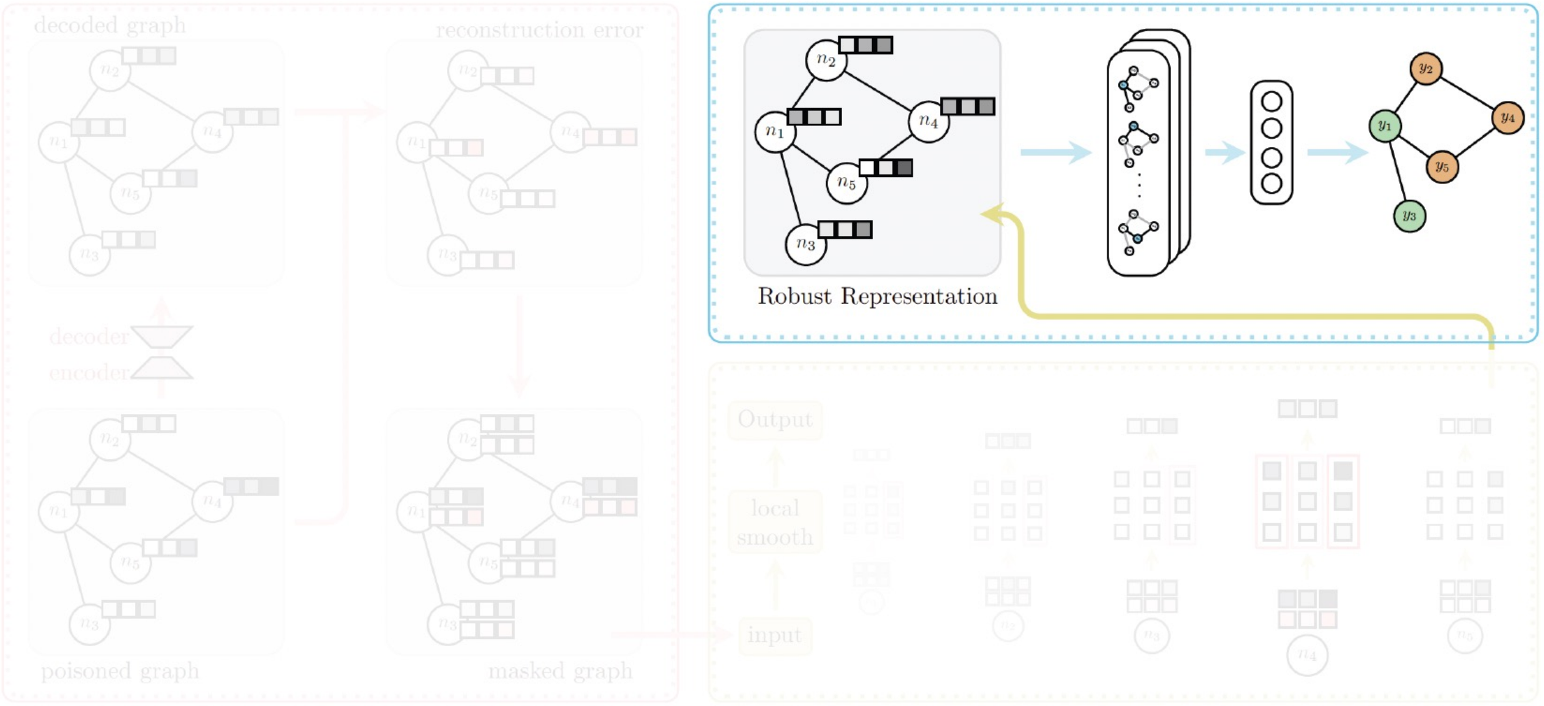


Zhou, B., Jiang, Y., Wang, Y., Liang, J., Gao, J., Pan, S., & Zhang, X. (2023, April). Robust graph representation learning for local corruption recovery. In Proceedings of the ACM Web Conference (WWW). 2023 (pp. 438-448).

Robust Graph Representation Learning for Local Corruption Recovery

-- Case study on Graph Feature Completion

- **Finally, learning robust GRL**



Zhou, B., Jiang, Y., Wang, Y., Liang, J., Gao, J., Pan, S., & Zhang, X. (2023, April). Robust graph representation learning for local corruption recovery. In Proceedings of the ACM Web Conference (WWW). 2023 (pp. 438-448).

Robust Graph Representation Learning for Local Corruption Recovery

-- Case study on Graph Feature Completion

❖ **Test the performance with node classification tasks**

Module	attribute injection								meta attack		
	Cora	Citeseer	PubMed	Coauthor-CS	Wiki-CS	Wisconsin	Texas	OGB-arxiv	Cora	Citeseer	PubMed
clean	81.26±0.65	71.77±0.29	79.01±0.44	90.19±0.48	77.62±0.26	56.47±5.26	65.14±1.46	71.10±0.21	81.26±0.65	71.77±0.29	79.01±0.44
GCN	69.06±0.74	57.58±0.71	67.69±0.40	82.41±0.23	65.44±0.23	48.24±3.19	58.92±2.02	68.42±0.15	75.07±0.64	55.32±2.22	72.88±0.30
APPNP	68.46±0.81	60.04±0.59	68.70±0.47	71.14±0.54	56.53±0.72	61.76±5.21	59.46±0.43	OOM	73.49±0.59	55.67±0.28	70.63±1.07
GNNGuard	61.96±0.30	54.94±1.00	68.50±0.38	80.67±0.88	65.69±0.32	46.86±1.06	59.19±0.81	65.75±0.32	72.02±0.61	57.64±1.31	71.10±0.32
ElasticGNN	77.74±0.79	64.61±0.85	71.23±0.21	79.91±1.39	64.18±0.53	53.33±2.45	59.77±3.24	41.34±0.38	**79.25±0.50**	67.29±1.17	71.95±0.52
AirGNN	76.22±3.75	62.14±0.82	74.73±0.43	80.18±0.31	71.36±0.20	61.56±0.72	59.46±1.24	52.32±0.58	78.94±0.45	65.58±0.63	78.58±0.71
MAGNETone	75.88±0.42	59.22±0.34	68.97±0.21	84.04±0.56	70.83±0.29	55.49±1.53	60.27±1.73	68.24±0.30	77.11±0.45	62.49±1.70	75.83±2.05
MAGNETgae	**79.07±0.56**	**64.79±0.73**	**75.41±0.35**	**86.50±0.37**	**72.40±0.21**	**64.31±2.60**	**60.81±2.18**	**68.68±0.03**	79.04±0.50	**67.40±0.73**	**78.63±0.32**
MAGNETtrue	78.48±0.67	68.55±0.74	75.63±0.56	89.23±0.40	75.50±0.20	65.69±1.57	60.54±2.16	69.57±0.23	80.88±0.37	67.46±0.95	79.16±0.41

- The three baseline graph smoothing methods fail to denoise local corruption within the input.
- MAGNET-gae outperforms its competitors and recovers at most 94% prediction accuracy from the perturbed attributes.
- An accurate mask approximation can push the prediction performance of graph representation up to MAGNET true's scores.

Zhou, B., Jiang, Y., Wang, Y., Liang, J., Gao, J., Pan, S., & Zhang, X. (2023, April). Robust graph representation learning for local corruption recovery. In Proceedings of the ACM Web Conference (WWW). 2023 (pp. 438-448).

Outline for Graph Data Enhancement

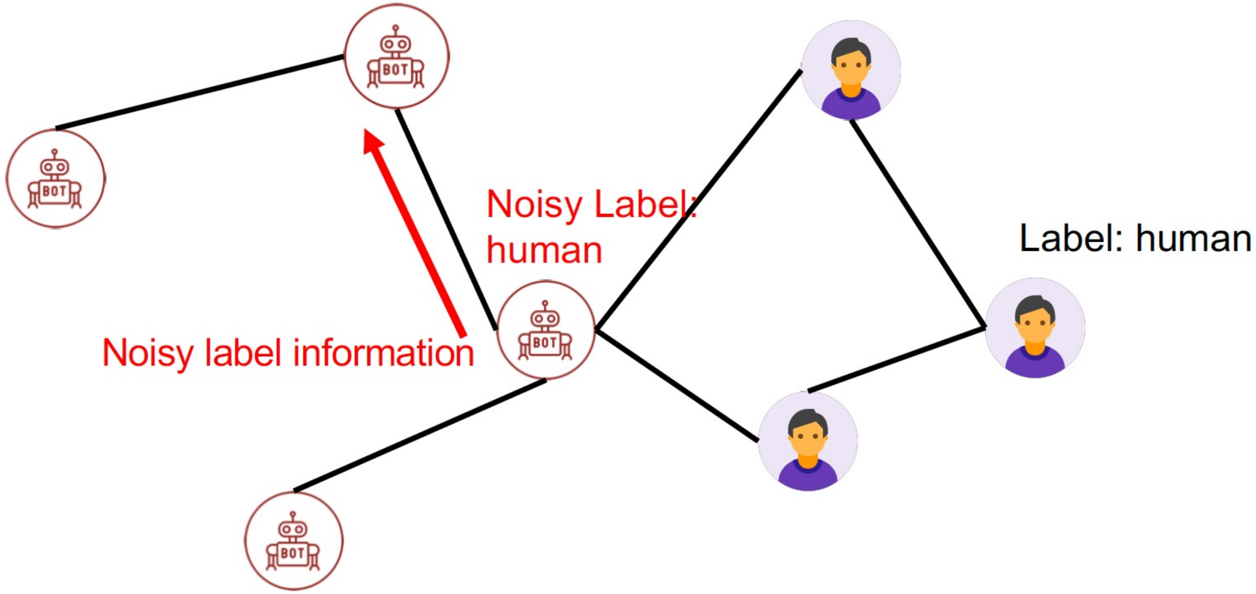
- ❖ **Overview of Graph Data Enhancement**
- ❖ **Techniques with Case Studies :**
  - Graph Structure Enhancement
  - Graph Feature Enhancement
  - **Graph Label Enhancement**
  - Graph Size Enhancement

Background of Graph Label Enhancement

Real-world graphs are generally **sparsely and noisily labeled**

Noise in sparsely labeled graphs can **degrade** the performance of GNN:

- Χ The size of labels is limited and GNN will overfit to noisy labels
- Χ Noisy label information propagates to their unlabeled neighbors



Dai, E., Aggarwal, C., & Wang, S. (2021, August). NRGNN: Learning a label noise resistant graph neural network on sparsely and noisily labeled graphs. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 227-236).

Overview Graph Label Enhancement

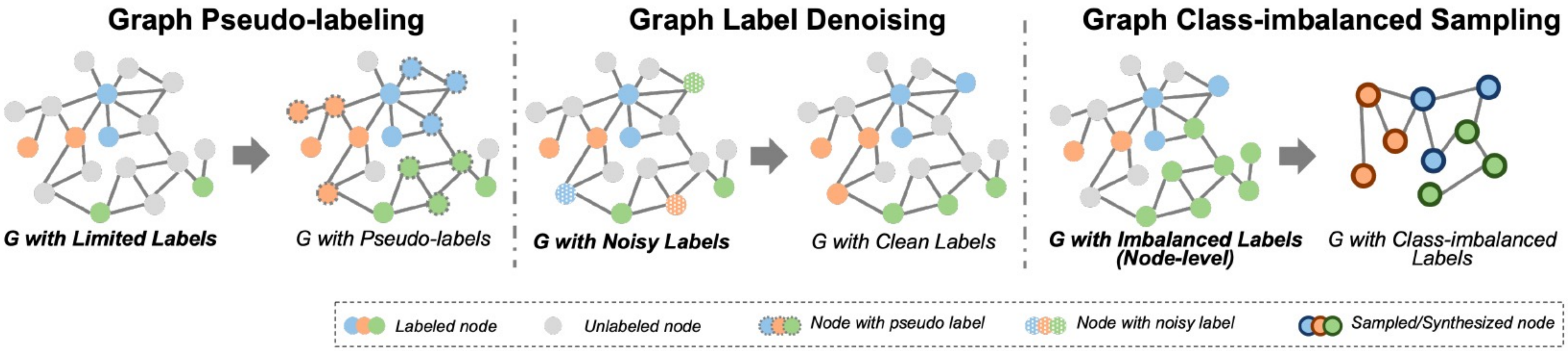


Fig. 7. Illustration of graph label enhancement methods.

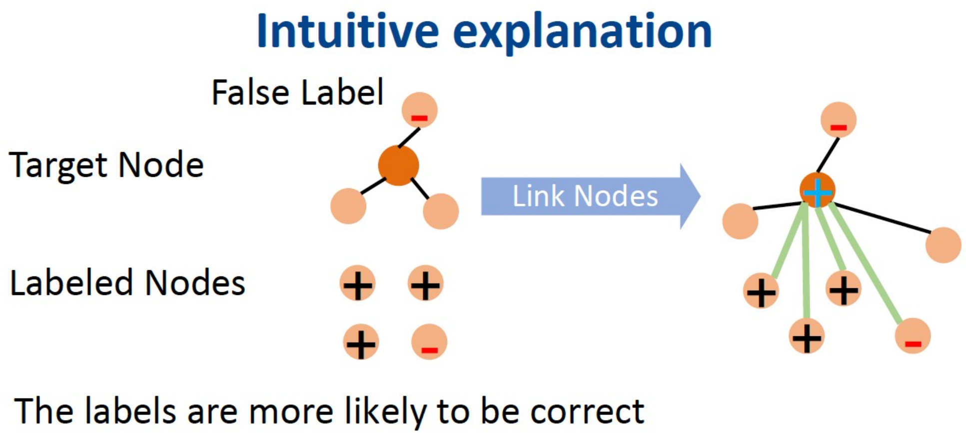
- **Graph Pseudo-labelling:** enriching the label information to alleviate the scarce label issue
- **Graph Label Denoising:** removing the redundant noisy label information to clean the noisy labels
- **Graph Class-imbalanced Sampling:** downsampling majority and/or synthesizing minority class labels to tackle the class-imbalanced label issue

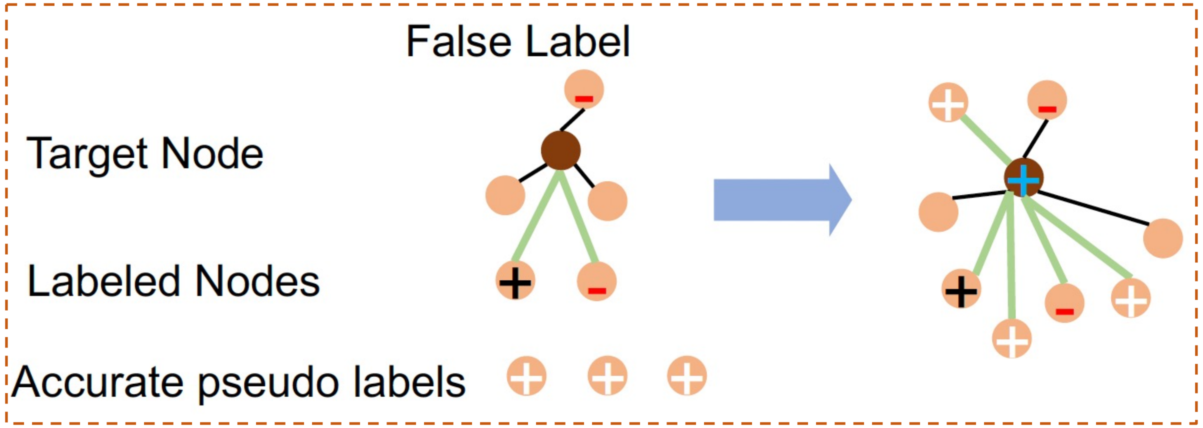
NRGNN: Learning on Sparsely and Noisily Labeled Graphs

--Case Study on Graph Label Enhancement

❖ **Preliminary Analysis**

- Linking **an unlabeled node** with **similar labeled nodes** belonging to the same class can increase the robustness against label noise.
- **Strategy: Extend the label set with accurate pseudo labels by selecting the predictions with high confidence score**



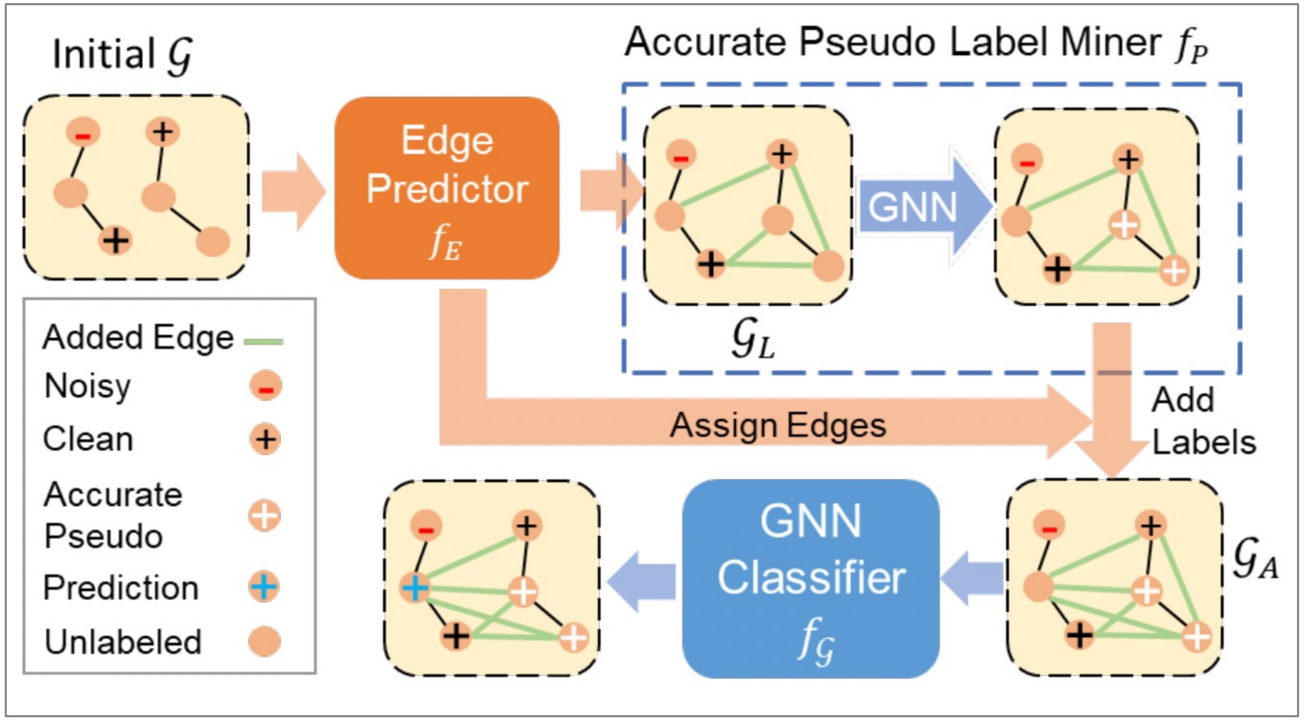


Dai, E., Aggarwal, C., & Wang, S. (2021, August). NRGNN: Learning a label noise resistant graph neural network on sparsely and noisily labeled graphs. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 227-236).

NRGNN: Learning on Sparsely and Noisily Labeled Graphs

--Case Study on Graph Label Enhancement

❖ NRGNN Framework



The Proposed NRGNN contains:

1) **Edge predictor**

Link unlabeled nodes with similar nodes having noisy/pseudo labels

2) **Accurate pseudo label miner**

Obtain accurate pseudo labels with high confidence score

3) **GNN classifier**

provide robust predictions

Dai, E., Aggarwal, C., & Wang, S. (2021, August). NRGNN: Learning a label noise resistant graph neural network on sparsely and noisily labeled graphs. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 227-236).

Outline for Graph Data Enhancement

- **Overview of Graph Data Enhancement**
- **Techniques with Case Studies :**
  - Graph Structure Enhancement
  - Graph Feature Enhancement
  - Graph Label Enhancement
  - **Graph Size Enhancement**

Graph Size Enhancement

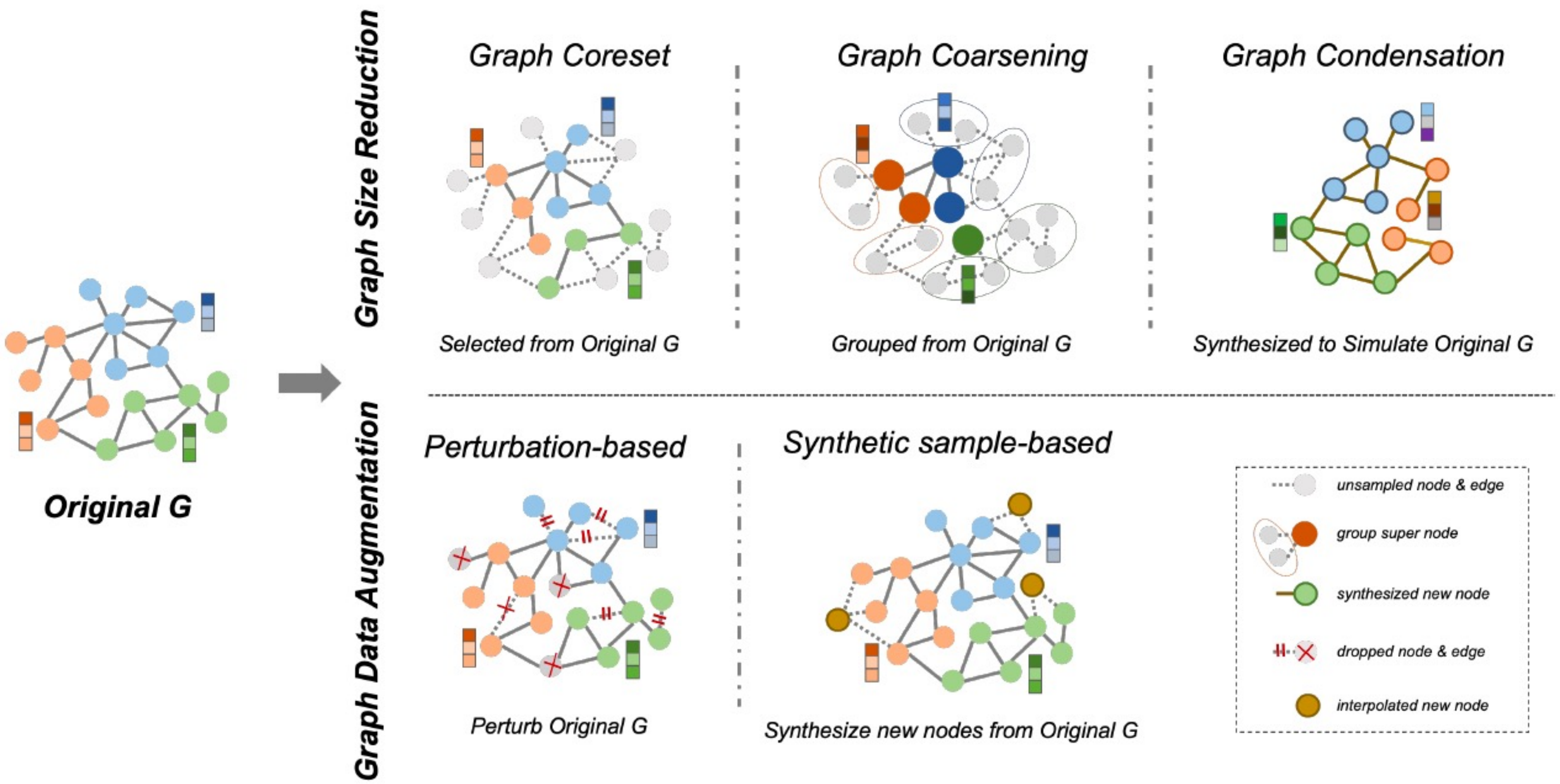


Fig. 8. Illustration of graph data-centric size enhancement methods.

- **Graph Size Reduction:** the oversized large-scale graphs with redundant information
- **Graph Data Augmentation:** small-scale graphs with limited data sources and insufficient information

Graph Size Enhancement

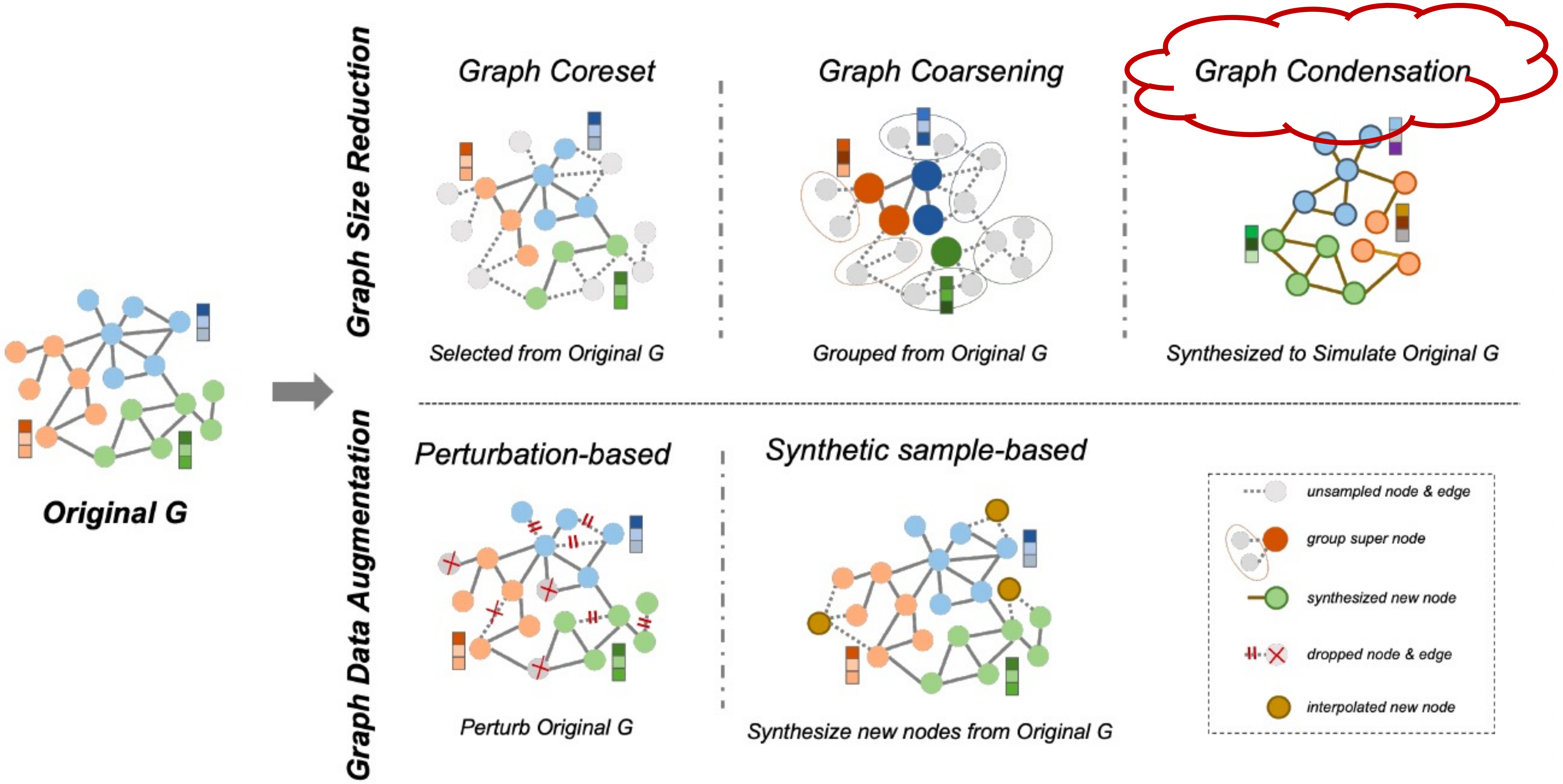


Fig. 8. Illustration of graph data-centric size enhancement methods.

- **Graph Size Reduction:** the oversized large-scale graphs with redundant information
- **Graph Data Augmentation:** small-scale graphs with limited data sources and insufficient information

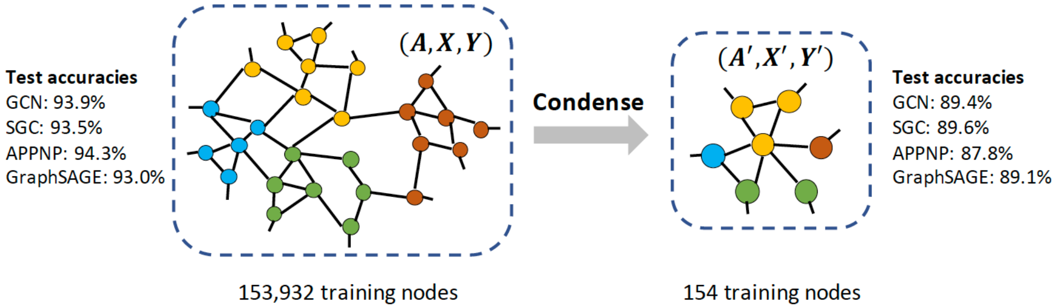
Background of Graph Condensation

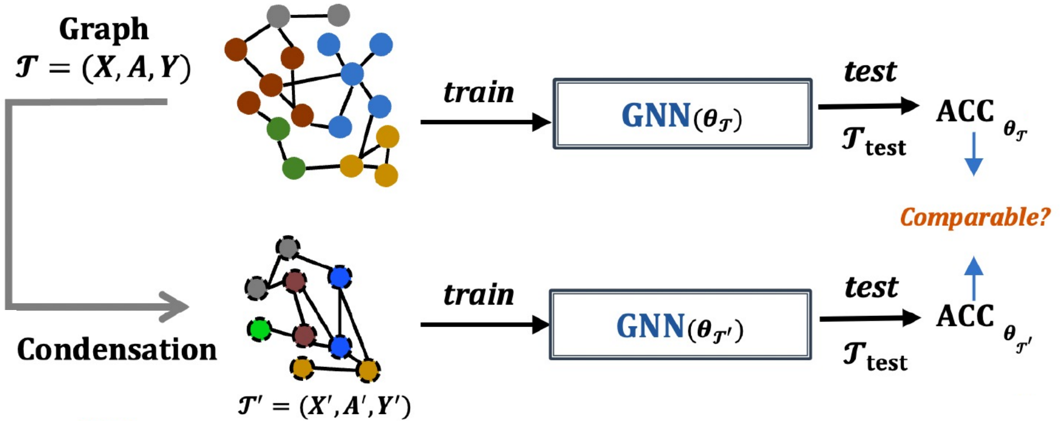
--Case Study on Graph Size Enhancement

❖What is graph condensation?

*aim to **reduce the size of a large-scale graph** by synthesizing a small-scale condensed graph*

→ → **the small-scale condensed graph** achieves **comparable test performance** as the large-scale graph when training the same GNN model.





Zheng, X., Zhang, M., Chen, C., Nguyen, Q. V. H., Zhu, X., & Pan, S. (2023). Structure-free Graph Condensation: From Large-scale Graphs to Condensed Graph-free Data. Advances in Neural Information Processing Systems (NeurIPS), 2023.

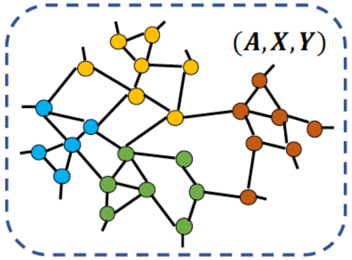
Background of Graph Condensation

--Case Study on Graph Size Enhancement

❖ Requirements, Advantages, & Applications

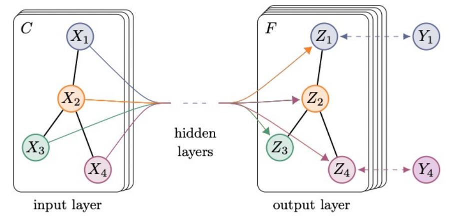
1) Why need GC [Requirements]?

Modelling large-scale graphs hinders GNN development with heavy costs



153,932 training nodes

repeat training & tuning.



Graph Convolutional Networks (GCN)

Table 1: Model serving space

Datasets	Model size	Training graph size	Training feature size	Total serving size
Arxiv	1.4MB	5.9MB	46.5MB	53.8MB
Reddit	7.6MB	86.0MB	370.7MB	464.3MB
Product	4.8MB	87.2MB	78.6MB	170.6MB
Amazon2M	3.0MB	485.4MB	684.0MB	1.17GB

X **Heavy costs on: graph data storage, computation, and memory**

Zheng, X., Zhang, M., Chen, C., Nguyen, Q. V. H., Zhu, X., & Pan, S. (2023). Structure-free Graph Condensation: From Large-scale Graphs to Condensed Graph-free Data. Advances in Neural Information Processing Systems (NeurIPS), 2023.

Background of Graph Condensation

--Case Study on Graph Size Enhancement

❖Requirements, Advantages, & Applications

2) How GC benefit [Advantages]?

Using condensed graph as substitution to facilitate GNN training:

- *Alleviated graph data storage/computation/memory costs*

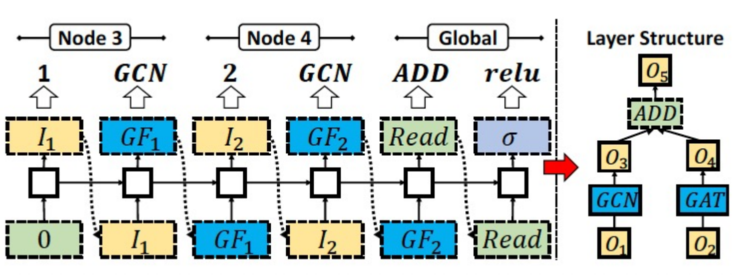
3) What practical applications of GC [Applications]?

- Graph Neural Architecture Search (GraphNAS)

  By searching on a small-scale condensed graph, accelerating new GNN architecture development in GraphNAS

  ...

- Privacy Protection
- Adversarial Robustness

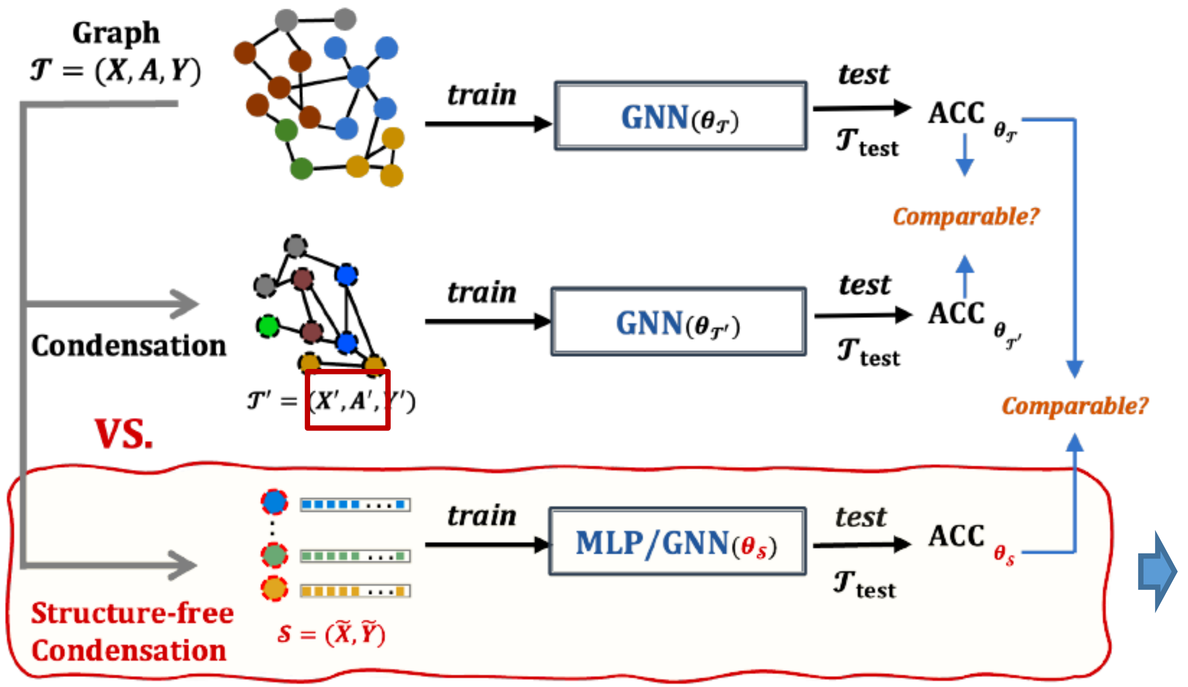


Graph NAS

Zheng, X., Zhang, M., Chen, C., Nguyen, Q. V. H., Zhu, X., & Pan, S. (2023). Structure-free Graph Condensation: From Large-scale Graphs to Condensed Graph-free Data. Advances in Neural Information Processing Systems (NeurIPS), 2023.

Our Solution: Structure-free Graph Condensation

--Case Study on Graph Size Enhancement



- **Existing works :**

$$\mathcal{T} = (\mathbf{X}, \mathbf{A}, \mathbf{Y}) \rightarrow \mathcal{T}' = (\mathbf{X}', \mathbf{A}', \mathbf{Y}'), \quad \text{GC}.$$

- **Our SFGC:**

$$\mathcal{T} = (\mathbf{X}, \mathbf{A}, \mathbf{Y}) \rightarrow \mathcal{S} = (\widetilde{\mathbf{X}}, \mathbf{I}, \widetilde{\mathbf{Y}}) = \mathcal{S} = (\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}), \quad \text{SFGC}.$$

- **Our Solution:**
  - ✓ **Structure-free paradigm** → • **Only synthesizes a small scaled node set to train a GNN/MLP**
  - ✓ **Long-range parameter matching schema** → • **Implicitly encodes topology structure into node attributes**

Zheng, X., Zhang, M., Chen, C., Nguyen, Q. V. H., Zhu, X., & Pan, S. (2023). Structure-free Graph Condensation: From Large-scale Graphs to Condensed Graph-free Data. Advances in Neural Information Processing Systems (NeurIPS), 2023.

Structure-free Graph Condensation

--Case Study on Graph Size Enhancement

Condensing large-scale graph into only node set without structures!

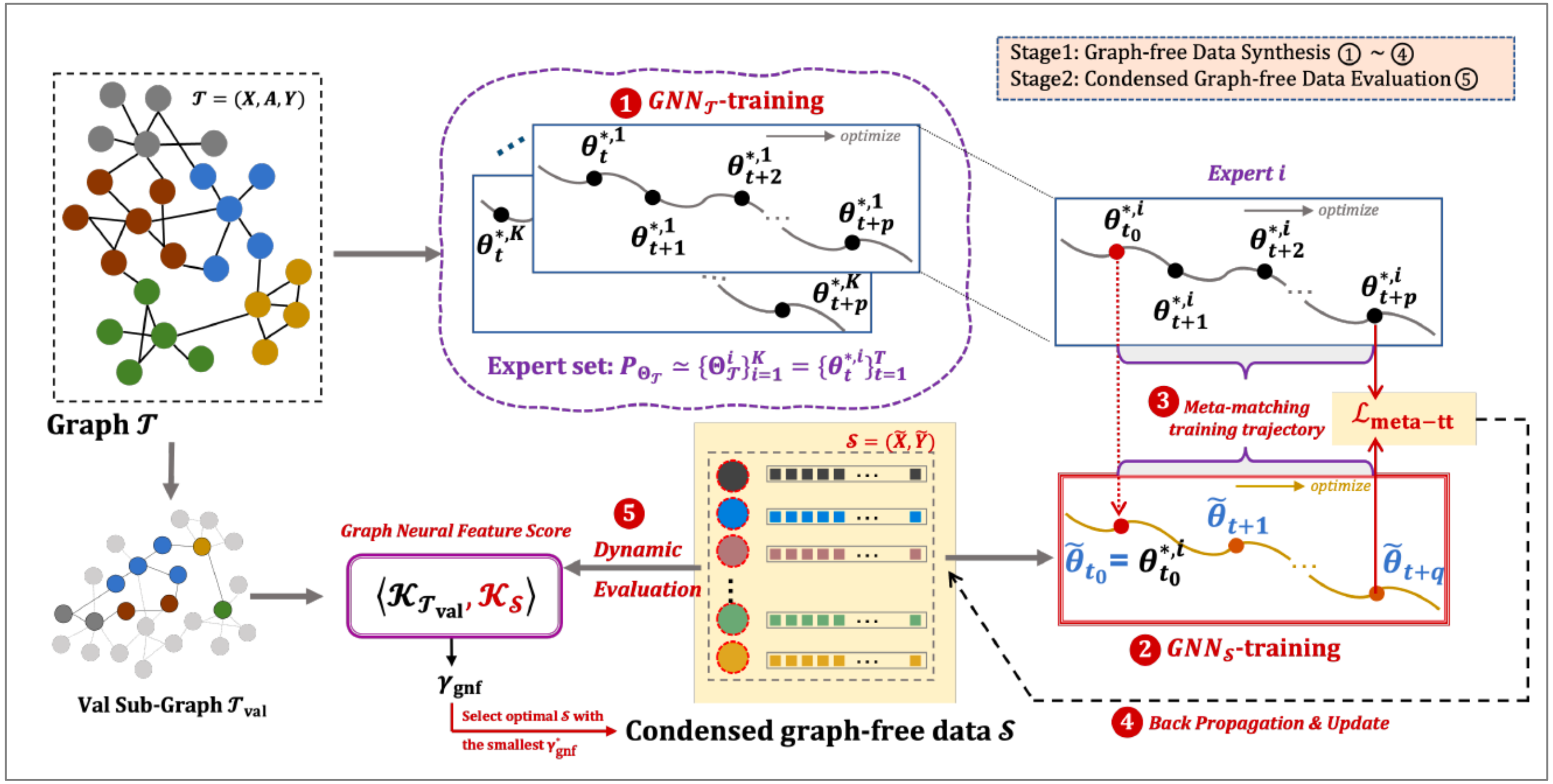


Figure 1. Overall pipeline of the proposed Structure-Free Graph Condensation (SFGC) framework

Input: large-scale T, GNN(T)

Output: small-scale condensed S

- **S1: train expert GNN on large-scale T**
- **S2-3: <u>long-term meta training trajectory matching</u> with condensed S**
- **S4: update S**
- **S5: dynamically evaluates S with a <u>GNTK-based score</u>**

Zheng, X., Zhang, M., Chen, C., Nguyen, Q. V. H., Zhu, X., & Pan, S. (2023). Structure-free Graph Condensation: From Large-scale Graphs to Condensed Graph-free Data. Advances in Neural Information Processing Systems (NeurIPS), 2023.

Experiments of SFGC

--Case Study on Graph Size Enhancement

Table 1: Node classification performance (ACC%±std) comparison between condensation methods and other graph size reduction methods with different condensation ratios. (Best results are in bold, and the second-bests are underlined.)

Datasets	Ratio (r)	*Other Graph Size Reduction Baselines*				*Condensation Methods*				Whole Dataset
		Coarsening [13]	Random [31]	Herding [31]	K-Center [28]	DC-Graph [42]	GCOND-X [18]	GCOND [18]	SFGC (**ours**)	
Citeseer	0.9%	52.2±0.4	54.4±4.4	57.1±1.5	52.4±2.8	66.8±1.5	71.4±0.8	70.5±1.2	**71.4**±0.5	71.7±0.1
	1.8%	59.0±0.5	64.2±1.7	66.7±1.0	64.3±1.0	66.9±0.9	69.8±1.1	70.6±0.9	**72.4**±0.4	
	3.6%	65.3±0.5	69.1±0.1	69.0±0.1	69.1±0.1	66.3±1.5	69.4±1.4	69.8±1.4	**70.6**±0.7	
Cora	1.3%	31.2±0.2	63.6±3.7	67.0±1.3	64.0±2.3	67.3±1.9	75.9±1.2	79.8±1.3	**80.1**±0.4	81.2±0.2
	2.6%	65.2±0.6	72.8±1.1	73.4±1.0	73.2±1.2	67.6±3.5	75.7±0.9	80.1±0.6	**81.7**±0.5	
	5.2%	70.6±0.1	76.8±0.1	76.8±0.1	76.7±0.1	67.7±2.2	76.0±0.9	79.3±0.3	**81.6**±0.8	
Ogbn-arxiv	0.05%	35.4±0.3	47.1±3.9	52.4±1.8	47.2±3.0	58.6±0.4	61.3±0.5	59.2±1.1	**65.5**±0.7	71.4±0.1
	0.25%	43.5±0.2	57.3±1.1	58.6±1.2	56.8±0.8	59.9±0.3	64.2±0.4	63.2±0.3	**66.1**±0.4	
	0.5%	50.4±0.1	60.0±0.9	60.4±0.8	60.3±0.4	59.5±0.3	63.1±0.5	64.0±0.4	**66.8**±0.4	
Flickr	0.1%	41.9±0.2	41.8±2.0	42.5±1.8	42.0±0.7	46.3±0.2	45.9±0.1	46.5±0.4	**46.6**±0.2	47.2±0.1
	0.5%	44.5±0.1	44.0±0.4	43.9±0.9	43.2±0.1	45.9±0.1	45.0±0.2	**47.1**±0.1	47.0±0.1	
	1%	44.6±0.1	44.6±0.2	44.4±0.6	44.1±0.4	45.8±0.1	45.0±0.1	**47.1**±0.1	**47.1**±0.1	
Reddit	0.05%	40.9±0.5	46.1±4.4	53.1±2.5	46.6±2.3	88.2±0.2	88.4±0.4	88.0±1.8	**89.7**±0.2	93.9±0.0
	0.1%	42.8±0.8	58.0±2.2	62.7±1.0	53.0±3.3	89.5±0.1	89.3±0.1	89.6±0.7	**90.0**±0.3	
	0.2%	47.4±0.9	66.3±1.9	71.0±1.6	58.5±2.1	**90.5**±1.2	88.8±0.4	90.1±0.5	90.3±0.3	

- Generally, SFGC achieves the best performance on the node classification task with 13 of 15 cases (five datasets and three condensation ratios for each of them), illustrating the high quality and expressiveness of the condensed graph-free data synthesized by our SFGC

Zheng, X., Zhang, M., Chen, C., Nguyen, Q. V. H., Zhu, X., & Pan, S. (2023). Structure-free Graph Condensation: From Large-scale Graphs to Condensed Graph-free Data. Advances in Neural Information Processing Systems (NeurIPS), 2023.

Part 3: Frontiers of Graph Data Exploitation

Outline for Graph Data Exploitation

- **Overview of Graph Data Exploitation**
- **Techniques with Case Studies :**
  - Graph Self-supervised Learning
  - Graph Semi-supervised Learning
  - Graph Active Learning
  - Graph Transfer Learning

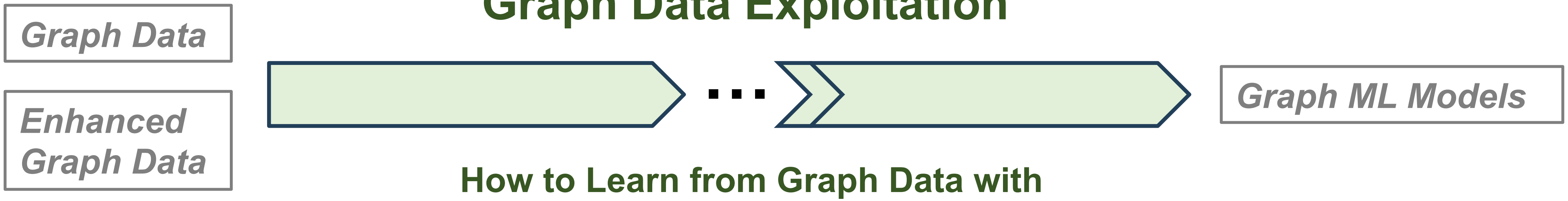
Overview of Graph Data Exploitation

Despite much effort on improving graph data quality, new graph data with high dynamics, complexity, diversity comes every day…

❖ **Core Question:**

- ❑ ***What if directly graph data enhancement not feasible?***
- ❑ ***What if after enhancement, it's still not enough to instruct the graph model development?***

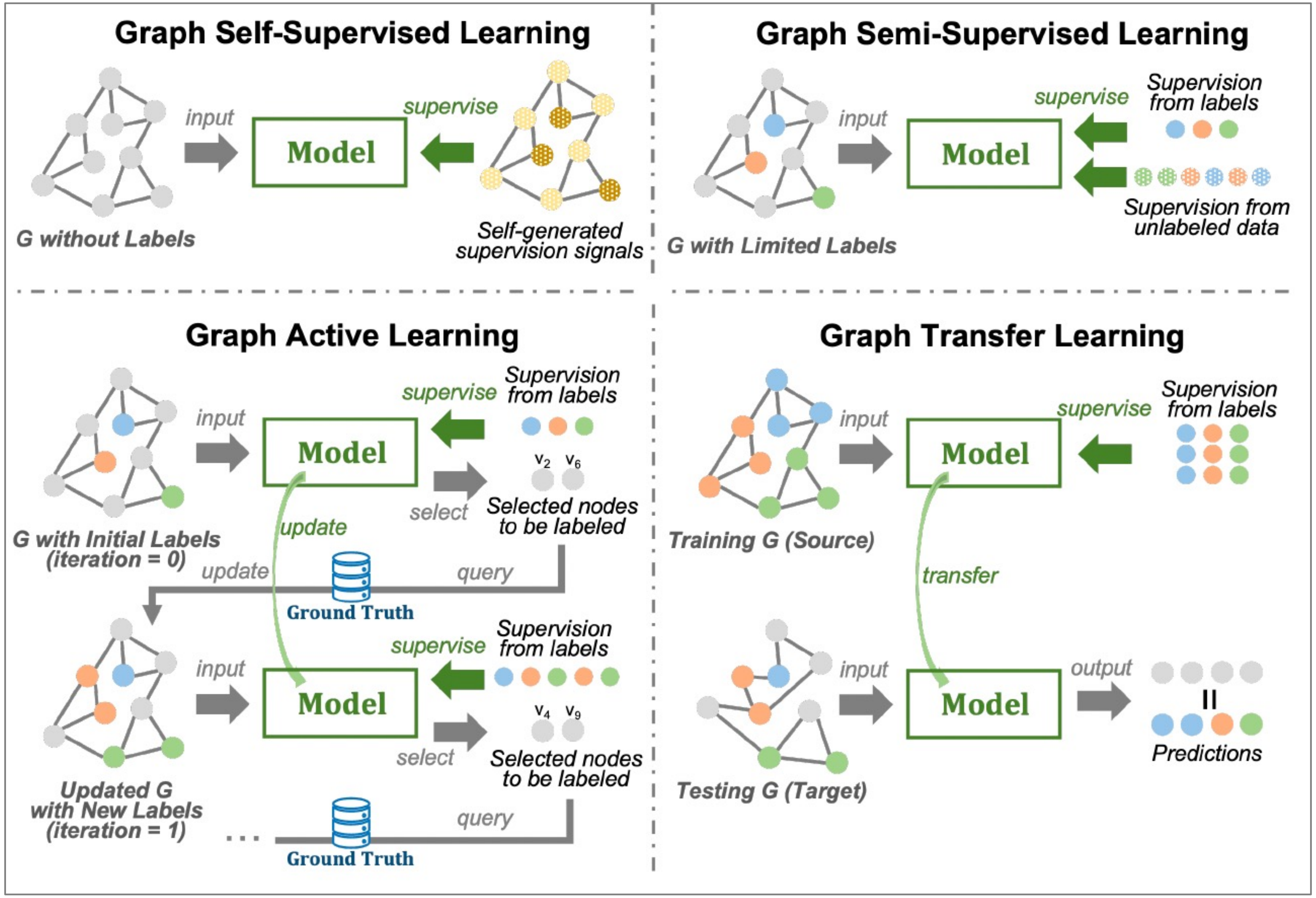
Graph Data Exploitation

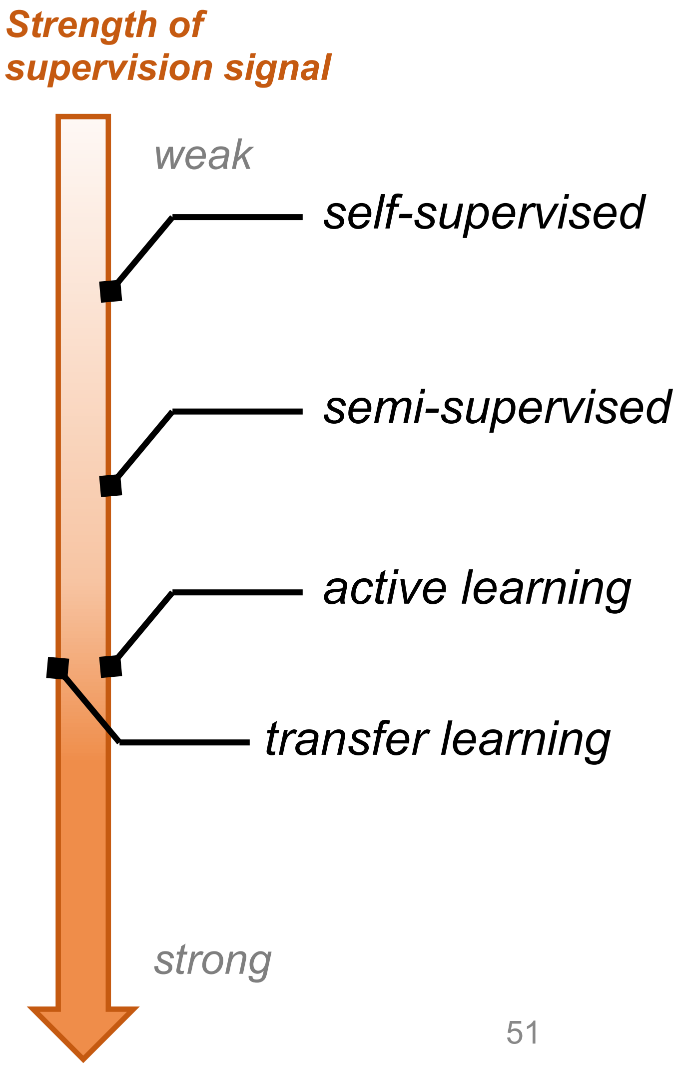


How to Learn from Graph Data with Limited-availability & Low-quality?

Overview of Graph Data Exploitation

❖ **Category of Graph Data Exploitation :**





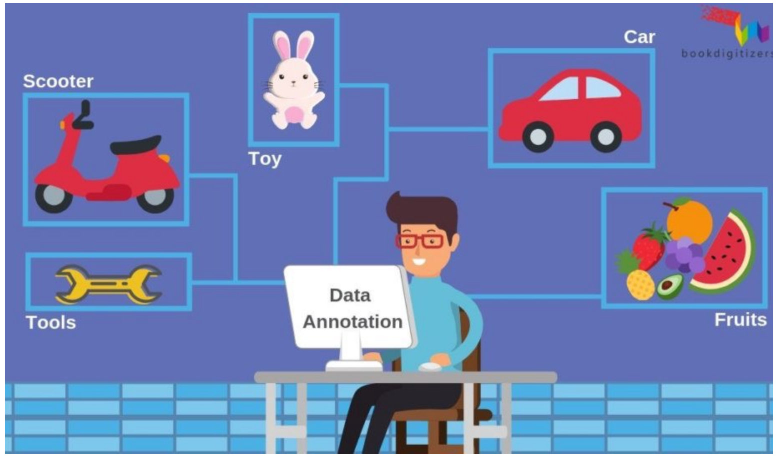
Outline for Graph Data Exploitation

- **Overview of Graph Data Exploitation**
- **Techniques with Case Studies :**
  - **Graph Self-supervised Learning**
  - Graph Semi-supervised Learning
  - Graph Active Learning
  - Graph Transfer Learning

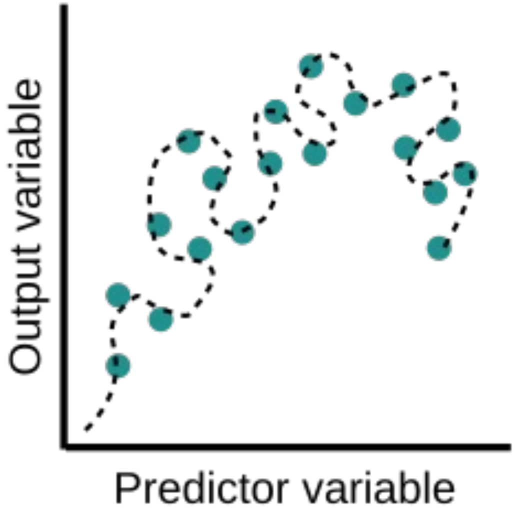
Motivation of Graph Self-supervised Learning

When lacking of sufficient supervision signals, the potential problems are…

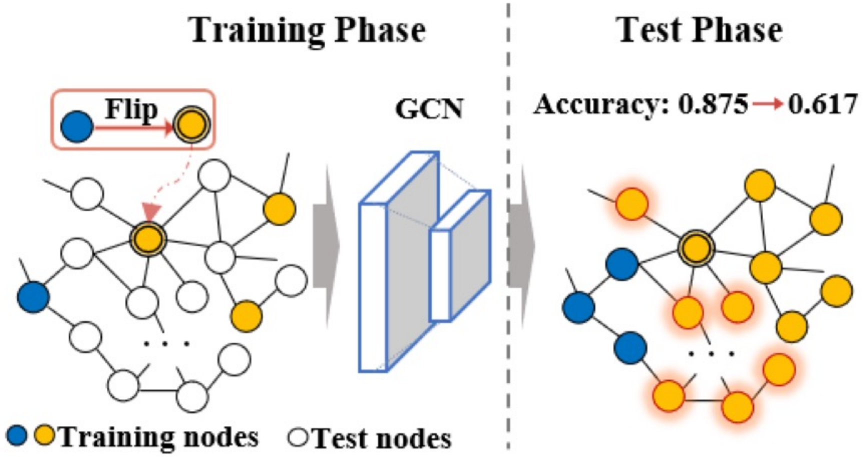
- Expensive cost of data collection and annotation



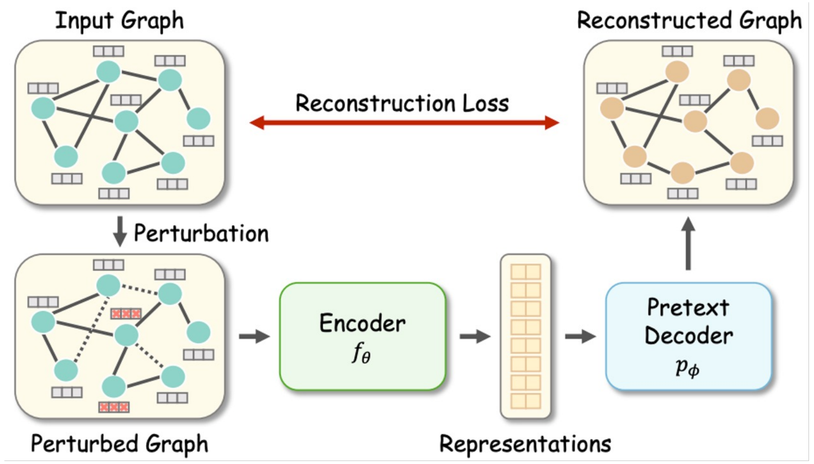
- Poor generalization



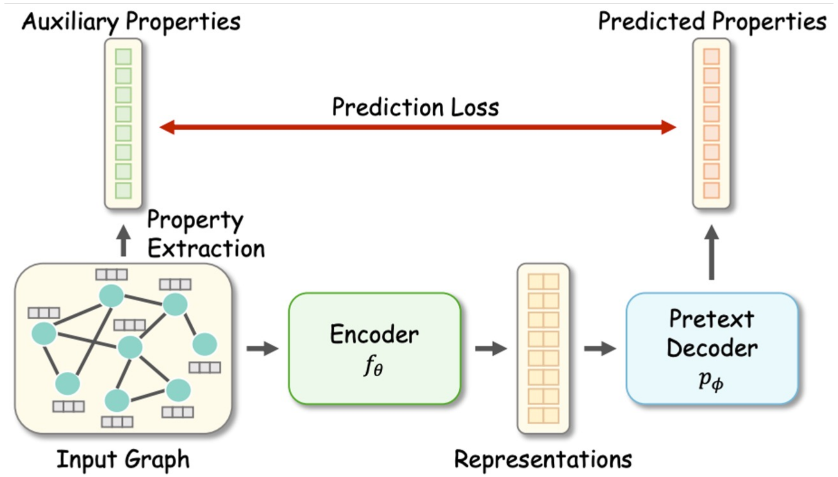
- Vulnerable to label-related adversarial attacks



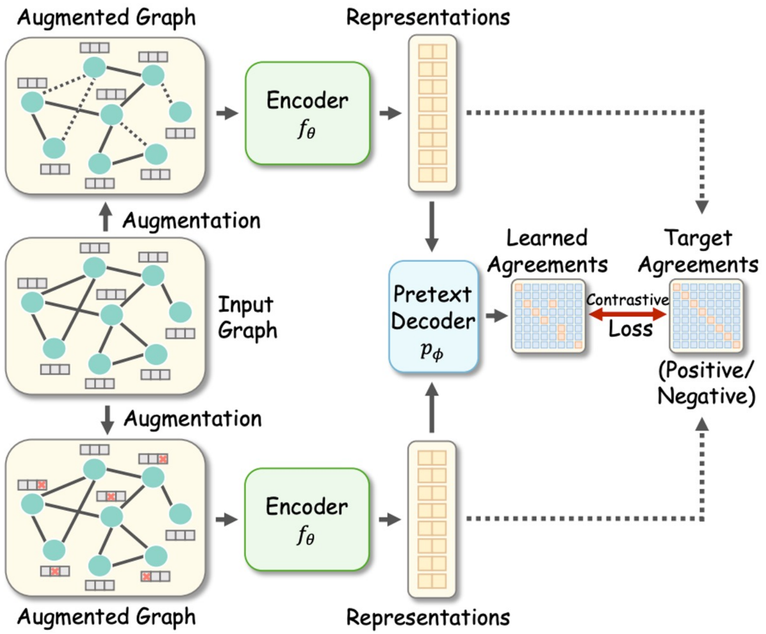
Typical Categories of GSSL



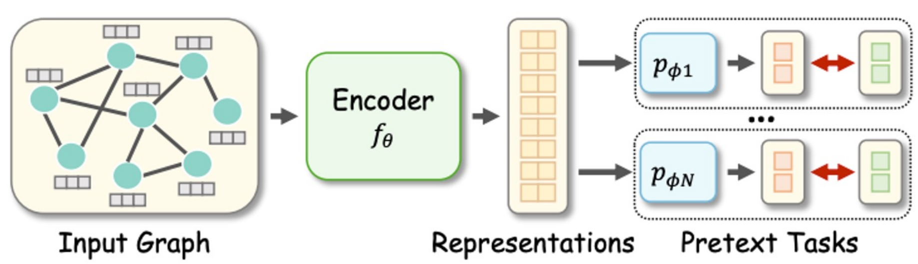
(1) Generation-based



(2) Auxiliary Property-based

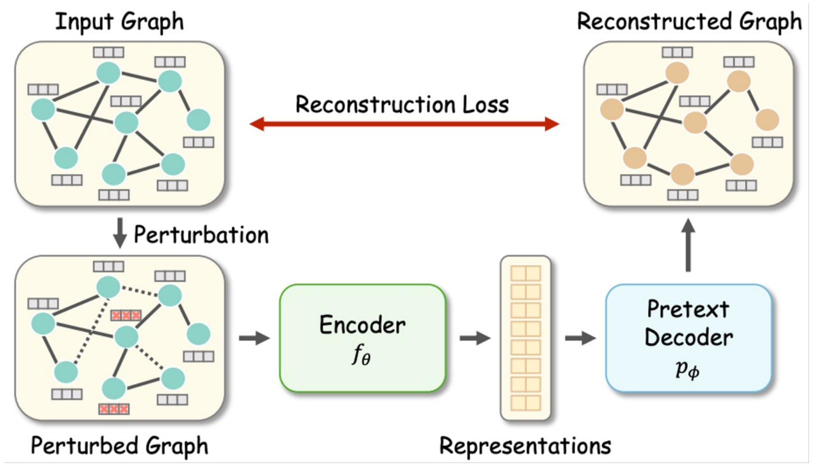


(3) Contrast-based

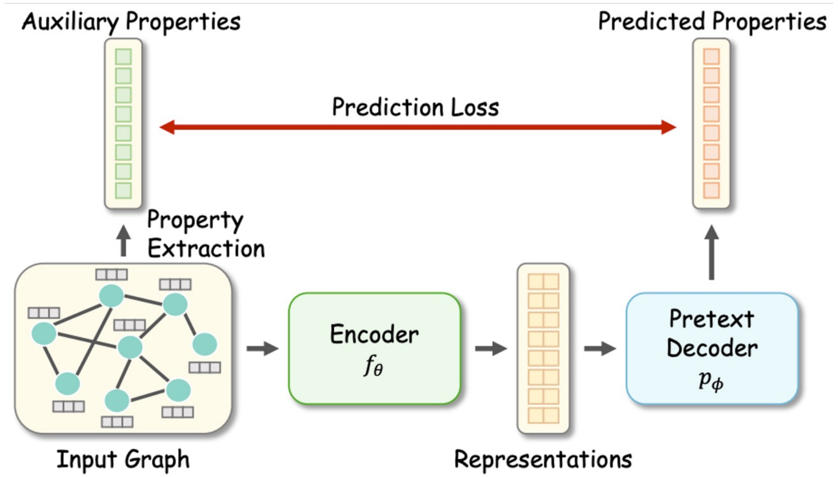


(4) Hybrid

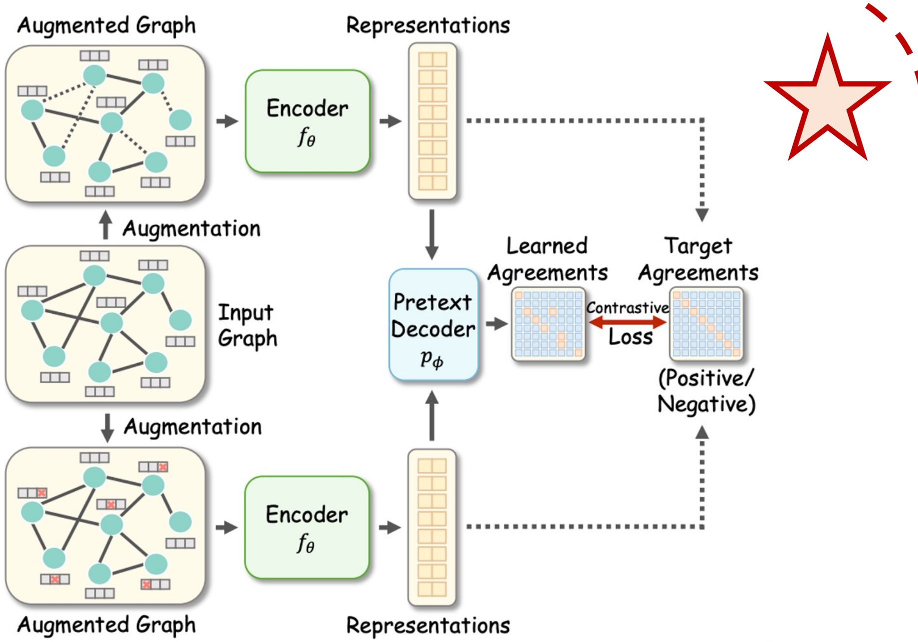
Typical Categories of GSSL



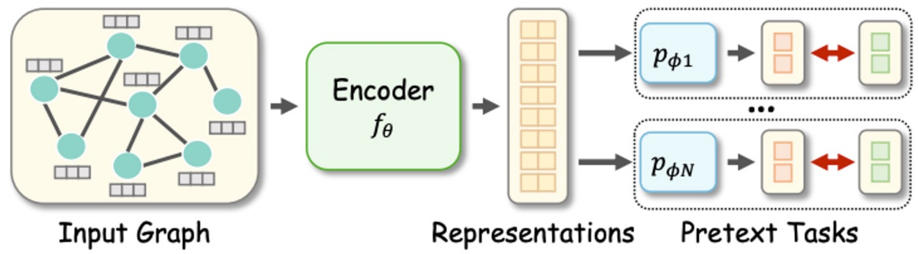
(1) Generation-based



(2) Auxiliary Property-based



(3) Contrast-based

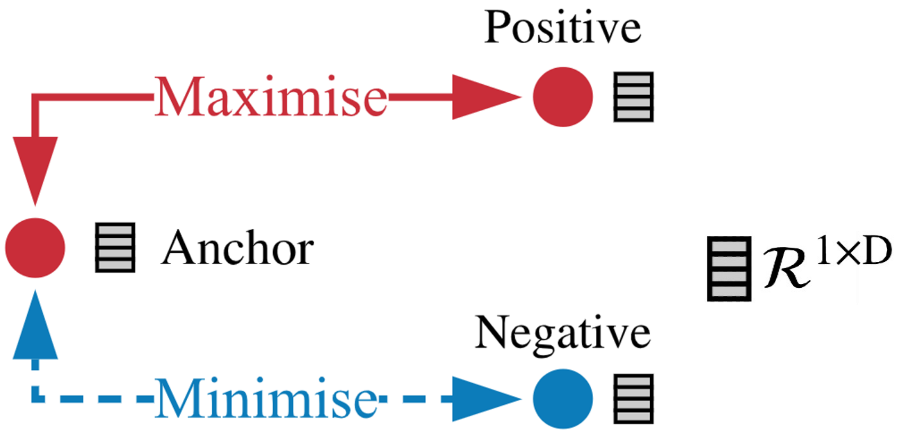


(4) Hybrid

Existing Problems - Slow Computation with Node Comparison

Most contrastive-learning approaches

- rely on **node-to-node comparison**
- require **heavy gradient computation**



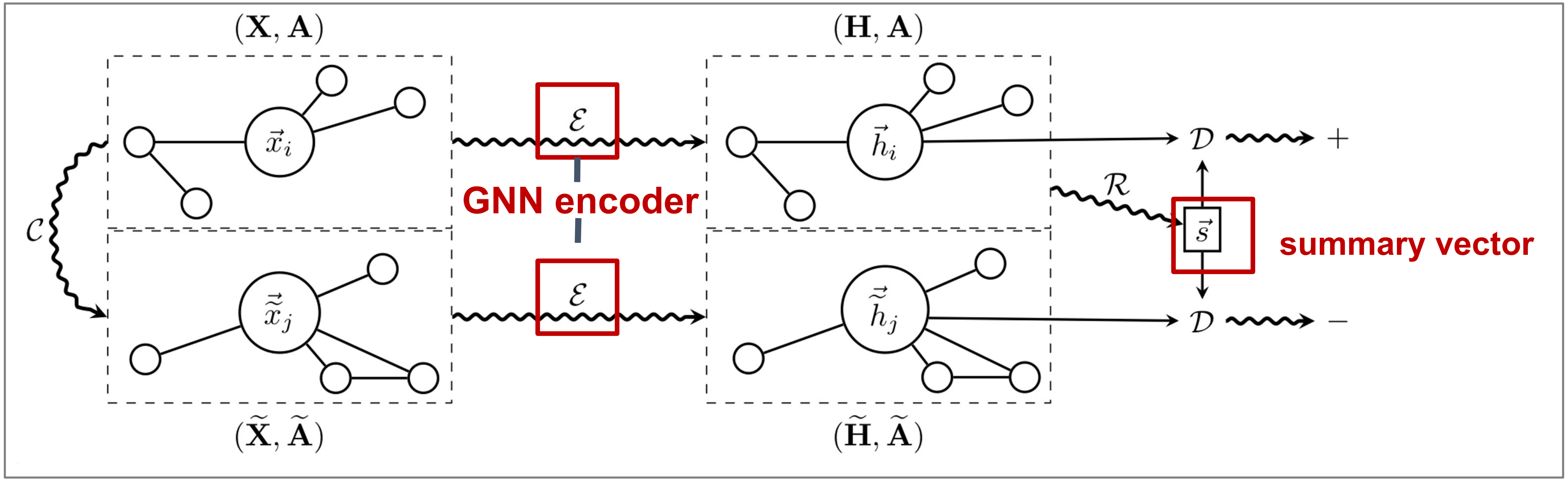
(a)Node-to-node Comparison

Zheng, Y., Pan, S., Lee, V., Zheng, Y., & Yu, P. S. (2022). Rethinking and scaling up graph contrastive learning: An extremely efficient approach with group discrimination. Advances in Neural Information Processing Systems (NeurIPS), *35*, 10809-10820.

Existing Problems - Slow Computation with Node Comparison

❖ **Existing typical Deep Graph Infomax (DGI) framework**

MI maximization between nodes and **summary vector**



$$\mathcal{L}_{\mathrm{DGI}} = \frac{1}{2N}\left(\sum_{i=1}^{N} \log \mathcal{D}(\vec{z_i}, s) + \log(1 - \mathcal{D}(\vec{\tilde{z}_i}, s))\right),$$

Zheng, Y., Pan, S., Lee, V., Zheng, Y., & Yu, P. S. (2022). Rethinking and scaling up graph contrastive learning: An extremely efficient approach with group discrimination. Advances in Neural Information Processing Systems (NeurIPS), *35*, 10809-10820.

Rethinking Existing DGI

--Case Study on Graph Self-supervised Learning

❖ Our important findings:

- Value in **summary vector s** almost becomes constant vector with **no variance**
- **DGI loss can be further simplified as BCE loss**

Activation	Statistics	Cora	CiteSeer	PubMed
ReLU/LReLU/PReLU	Mean	0.50	0.50	0.50
	Std	1.3e-03	1.0e-04	4.0e-04
	Range	1.4e-03	8.0e-04	1.5e-03
Sigmoid	Mean	0.62	0.62	0.62
	Std	5.4e-05	2.9e-05	6.6e-05
	Range	3.6e-03	3.0e-03	3.2e-03

Dataset	0	0.2	0.4	0.6	0.8	1.0
Cora	70.3±0.7	82.4±0.2	82.3±0.3	82.5±0.4	82.3±0.3	82.5±0.1
CiteSeer	61.8±0.8	71.7±0.6	71.9±0.7	71.6±0.9	71.7±1.0	71.6±0.8
PubMed	68.3±1.5	77.8±0.5	77.9±0.8	77.7±0.9	77.4±1.1	77.2±0.9

Set ϵ to 1 for $s = \epsilon I = I$, and remove w in $\mathcal{D}(z_i, \vec{s}) = z_i \cdot w \cdot \vec{s}$,

$$\mathcal{L}_{DGI} = \frac{1}{2N}\Big(\sum_{i=1}^{N} \log \mathcal{D}(z_i, \boxed{s}) + \log(1 - \mathcal{D}(\tilde{z}_i, s))\Big),$$

1

$$= \frac{1}{2N}\Big(\sum_{i=1}^{N} \log(z_i \cdot s) + \log(1 - \tilde{z}_i \cdot s)\Big),$$

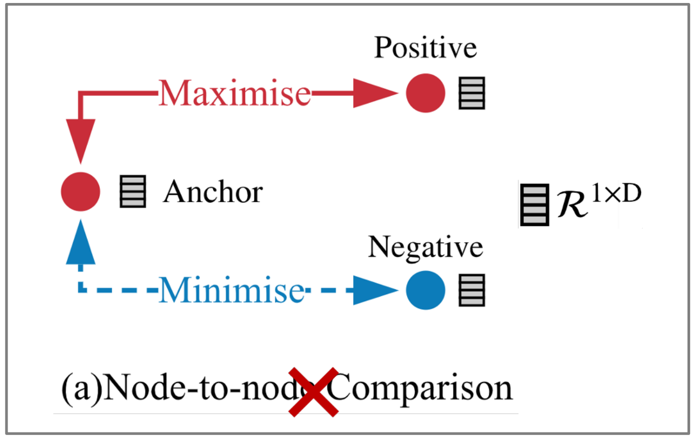
$$= \frac{1}{2N}\Big(\sum_{i=1}^{N} \log(\boxed{\mathrm{sum}}(z_i)) + \log(1 - \mathrm{sum}(\tilde{z}_i))\Big),$$

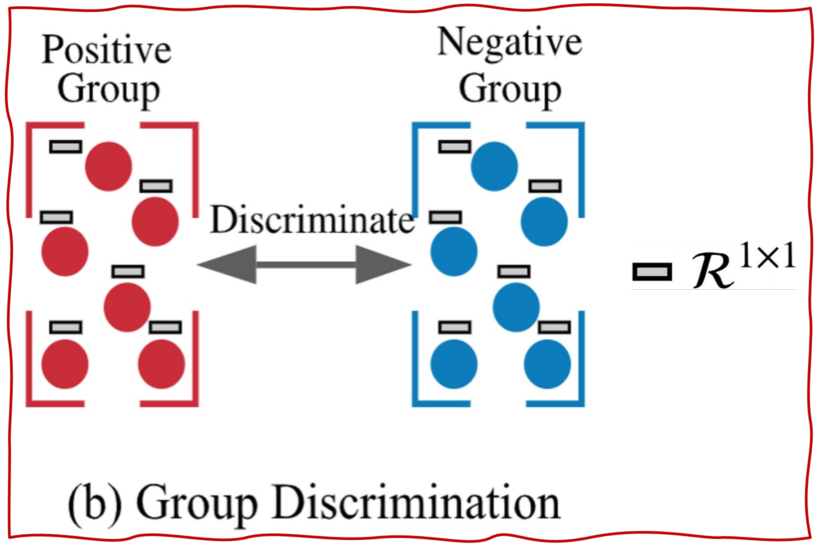
2

Zheng, Y., Pan, S., Lee, V., Zheng, Y., & Yu, P. S. (2022). Rethinking and scaling up graph contrastive learning: An extremely efficient approach with group discrimination. Advances in Neural Information Processing Systems (NeurIPS), *35*, 10809-10820.

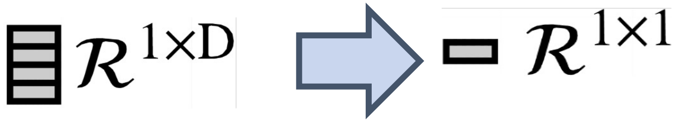
Our Solution: Group Discrimination (GD)

--Case Study on Graph Self-supervised Learning





Summarisation (e.g., sum):



- **Positive Group:**

  Summarised Node representations generated with original or augmented graph.

- **Negative Group:**

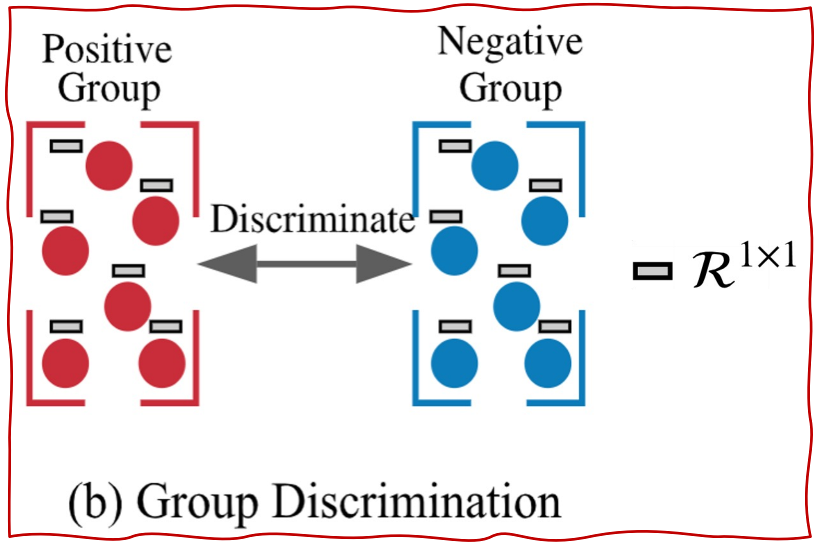
  Summarised Node representations generated with corrupted graph.

Zheng, Y., Pan, S., Lee, V., Zheng, Y., & Yu, P. S. (2022). Rethinking and scaling up graph contrastive learning: An extremely efficient approach with group discrimination. Advances in Neural Information Processing Systems (NeurIPS), *35*, 10809-10820.

Our Solution: Group Discrimination (GD)

--Case Study on Graph Self-supervised Learning

Use a very simple BCE loss to conduct discrimination



$$\mathcal{L}_{BCE} = -\frac{1}{2N}\left(\sum_{i=1}^{2N} y_i \log h_i + (1 - y_i)\log(1 - h_i)\right)$$

If positive → y = 1, else → y = 0

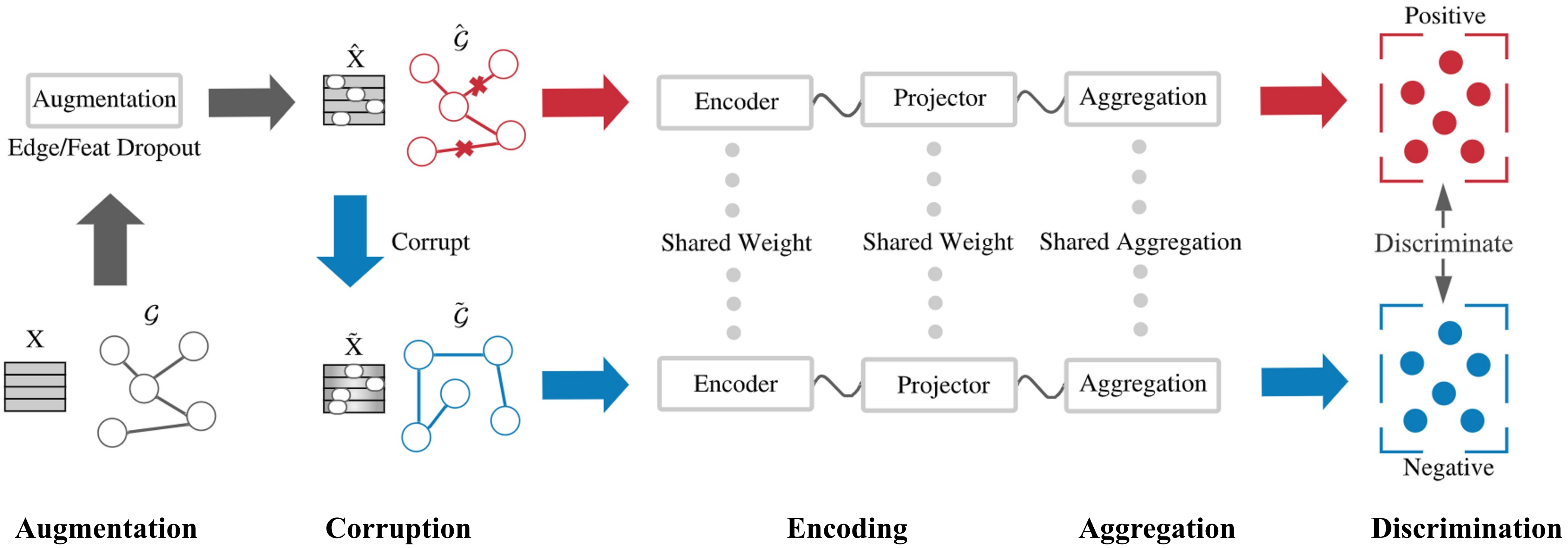
where $h_i \in \mathcal{R}^{1\times1}$ is the summarised node embedding/binary prediction for a node i

A very simple binary classification task: discriminating positive/negative samples

Zheng, Y., Pan, S., Lee, V., Zheng, Y., & Yu, P. S. (2022). Rethinking and scaling up graph contrastive learning: An extremely efficient approach with group discrimination. Advances in Neural Information Processing Systems (NeurIPS), *35*, 10809-10820.

Proposed Framework: Graph Group Discrimination (GGD)

--Case Study on Graph Self-supervised Learning



Zheng, Y., Pan, S., Lee, V., Zheng, Y., & Yu, P. S. (2022). Rethinking and scaling up graph contrastive learning: An extremely efficient approach with group discrimination. Advances in Neural Information Processing Systems (NeurIPS), *35*, 10809-10820.

Performance of Graph Group Discrimination (GGD)

--Case Study on Graph Self-supervised Learning

Small-to-Medium scale Dataset

Data	Method	Cora	CiteSeer	PubMed	Comp	Photo
X, A, Y	GCN	81.5	70.3	79.0	76.3±0.5	87.3±1.0
X, A, Y	GAT	83.0±0.7	72.5±0.7	79.0±0.3	79.3±1.1	86.2±1.5
X, A, Y	SGC	81.0±0.0	71.9±0.1	78.9±0.0	74.4±0.1	86.4±0.0
X, A, Y	CG3	83.4±0.7	**73.6**±0.8	80.2±0.8	79.9±0.6	89.4±0.5
X, A	DGI	81.7±0.6	71.5±0.7	77.3±0.6	75.9±0.6	83.1±0.5
X, A	GMI	82.7±0.2	73.0±0.3	80.1±0.2	76.8±0.1	85.1±0.1
X, A	MVGRL	82.9±0.7	72.6±0.7	79.4±0.3	79.0±0.6	87.3±0.3
X, A	GRACE	80.0±0.4	71.7±0.6	79.5±1.1	71.8±0.4	81.8±1.0
X, A	BGRL	80.5±1.0	71.0±1.2	79.5±0.6	89.2±0.9	91.2±0.8
X, A	GBT	81.0±0.5	70.8±0.2	79.0±0.1	88.5±1.0	91.1±0.7
X, A	**GGD**	**84.1**±0.4	73.0±0.6	**81.3**±0.8	**90.1**±0.9	**92.5**±0.6

Time Consumption Improvement (epoch per second)

Method	Cora	CiteSeer	PubMed	Comp	Photo
DGI	0.085	0.134	0.158	0.171	0.059
GMI	0.394	0.497	2.285	1.297	0.637
MVGRL	0.123	0.171	0.488	0.663	0.468
GRACE	0.056	0.092	0.893	0.546	0.203
BGRL	0.085	0.094	0.147	0.337	0.273
GBT	0.073	0.072	0.103	0.492	0.173
GGD	0.010	0.021	0.015	0.016	0.009
Improve	7.3-39.4×	3.4-23.7×	6.9-152.3×	10.7-15.3×	19.2-70.8×

Memory Consumption Improvement (MB)

Method	Cora	CiteSeer	PubMed	Comp	Photo
DGI	4,189	8,199	11,471	7,991	4,946
GMI	4,527	5,467	14,697	10,655	5,219
MVGRL	5,381	5,429	6,619	6,645	6,645
GRACE	1,913	2,043	12,597	8,129	4,881
BGRL	1,627	1,749	2,299	5,069	3,303
GBT	1,651	1,799	2,461	5,037	2,641
GGD	1,475	1,587	1,629	1,787	1,637
Improve	10.7-72.6%	11.8-80.6%	27.2-85.8%	64.5-83.2%	38.0-75.4%

Zheng, Y., Pan, S., Lee, V., Zheng, Y., & Yu, P. S. (2022). Rethinking and scaling up graph contrastive learning: An extremely efficient approach with group discrimination. Advances in Neural Information Processing Systems (NeurIPS), *35*, 10809-10820.

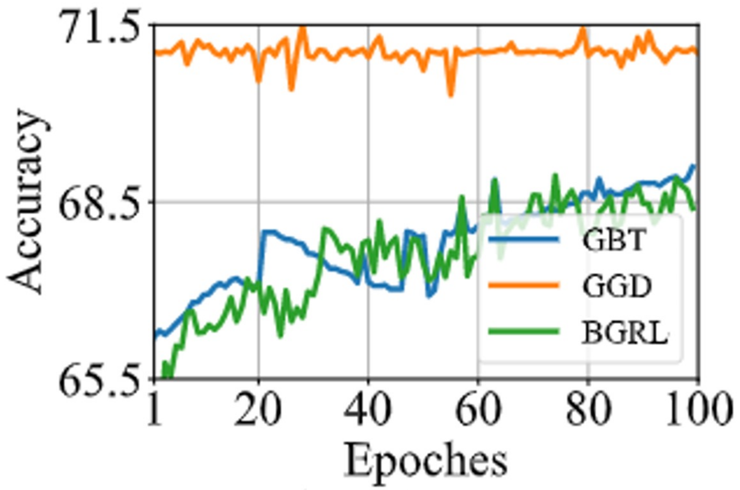
Performance of Graph Group Discrimination (GGD)

--Case Study on Graph Self-supervised Learning

Using only 0.18 seconds and 69.8% less memory to reach SOTA.

10783 faster than existing methods.

Method	Valid	Test	Memory	Time	Total
Supervised GCN	73.0±0.2	71.7±0.3	-	-	-
MLP	57.7±0.4	55.5±0.2	-	-	-
Node2vec	71.3±0.1	70.1±0.1	-	-	-
DGI	71.3±0.1	70.3±0.2	-	-	-
GRACE(10k epos)	72.6±0.2	71.5±0.1	-	-	-
BGRL(10k epos)	72.5±0.1	71.6±0.1	OOM (Full-graph)	/	/
GBT(300 epos)	71.0±0.1	70.1±0.2	14,959MB	6.47	1,941.00
GGD(1 epo)	72.7±0.3	71.6±0.5	4,513MB|69.8%	0.18	0.18|10,783×



(a) ogbn-arxiv

Fast convergence →
converge with only 1 epoch

Zheng, Y., Pan, S., Lee, V., Zheng, Y., & Yu, P. S. (2022). Rethinking and scaling up graph contrastive learning: An extremely efficient approach with group discrimination. *Advances in Neural Information Processing Systems*, *35*, 10809-10820.

Outline for Graph Data Exploitation

- **Overview of Graph Data Exploitation**
- **Techniques with Case Studies :**
  - Graph Self-supervised Learning
  - **Graph Semi-supervised Learning**
  - Graph Active Learning
  - Graph Transfer Learning

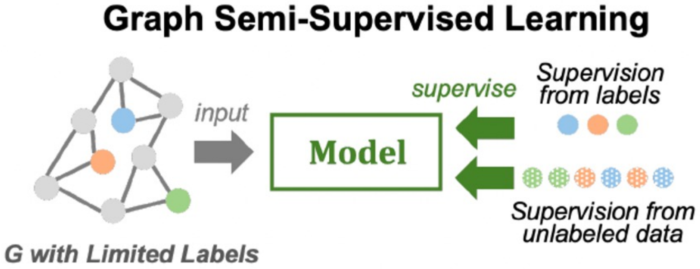
Background of Graph Semi-supervised Learning

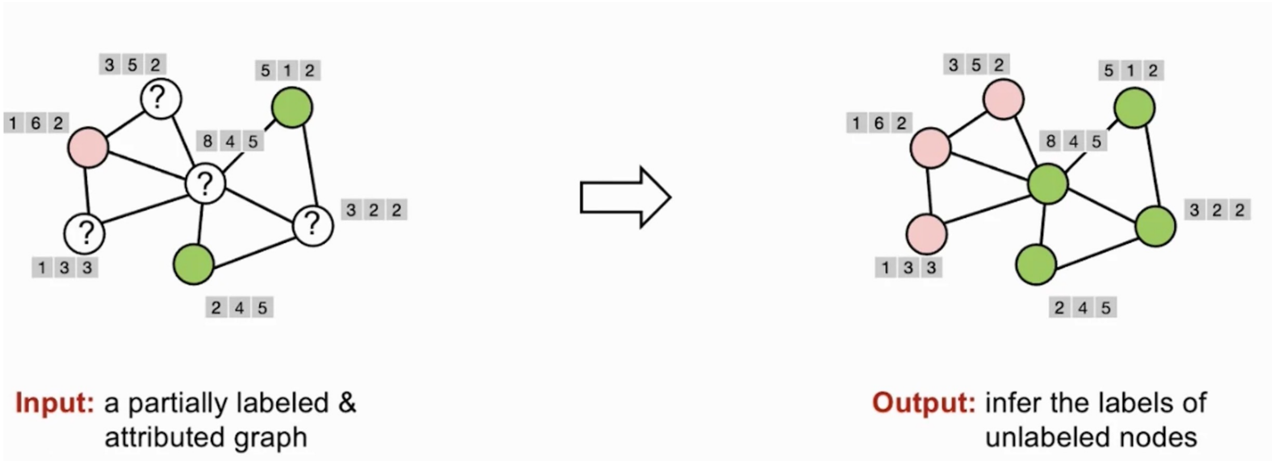
❖ **Graph Semi-supervised Learning: Only limited labels are provided**

- **Core idea from DC-GML view:**

Learn to fully leverage/exploit the unlabeled part and collaborate with the labeled part

- **Methodology: Regularization & Pseudo Labelling**





Category of Graph Semi-supervised Learning

❖ **Category from DC-GML view**

Table 6. Summary of methods in graph semi-supervised learning.

Graph Semi-supervised Learning	Techniques	Categories
Zhu et al. [223]	Graph Laplacian regularization	Regularization-based
Zhou et al. [216]	Graph Laplacian regularization	Regularization-based
Zhou et al. [217]	Local smoothness under homophily	Regularization-based
Li et al. [82]	Self-training with training set extension	Pseudo-labelling
NodeAug [171]	KL divergence-based consistency	Regularization-based
GRAND [39]	L2 distance-based consistency	Regularization-based
M3S [148]	Clustering-based pseudo label generation	Pseudo-labelling
SimP-GCN [68]	Feature-level similarity in pairwise distance	Regularization-based
GCN-LPA [166]	Edge weights with graph structure regularization	Regularization-based
CG^3 [162]	Self-supervised objective based regularization	Regularization-based
GCPN [163]	Contrastive and possion learning based regularization	Regularization-based
Meta-PN [32]	Adaptive label propagator based on label propagation	Pseudo-labelling
CycProp [88]	High-quality contextual node selection	Pseudo-labelling

Outline for Graph Data Exploitation

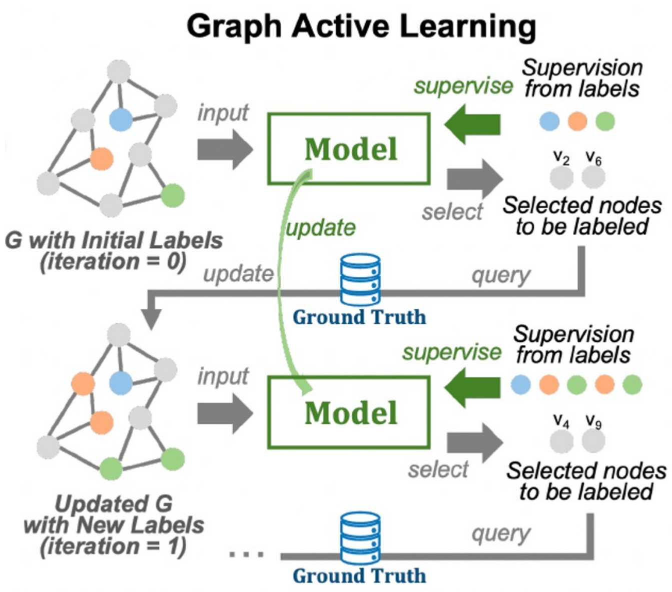
- **Overview of Graph Data Exploitation**
- **Techniques with Case Studies :**
  - Graph Self-supervised Learning
  - Graph Semi-supervised Learning
  - **Graph Active Learning**
  - Graph Transfer Learning

Background of Graph Active Learning

Given the fixed cost (e.g., human labour and expert knowledge) for label, how can we fully make the best use of such labelling budget?

❖ **Graph Active Learning: dynamically select the samples to label during the training procedure**

In the practical active learning process, the nodes to label are selected automatically by the models following several ***selection criteria****.*



❖ **Category from DC-GML view:**

- *Rule-based*
- *Reinforcement learning-based,*
- *Influence function-based,*
- *Other hybrid methods*

Category of Graph Active Learning

❖ **Category from DC-GML view**

Table 7. Summary of methods in graph active learning.

Graph Active Learning	Techniques	Categories
AGE [13]	Information entropy, density, and centrality rules	Rule-based
ANRMAB [42]	Multi-armed bandit mechanism	Rule-based
ActiveHNE [24]	Multi-armed bandit mechanism on heterogeneous graphs	Rule-based
FeatProp [183]	Closest cluster center based labelling	Clustering-based
ATNE [65]	Active transfer learning based node selection	Rule-based
ASGN [50]	Sample diversity based node selection	Rule-based
GPA [54]	GCN-based policy network	RL-based
MetAL [103]	Meta-gradients estimation	Meta Learning-based
SEAL [85]	Adversarial learning with divergence value	Adversarial-based
GRAIN [205]	Diversified influence maximization objective	Influence-based
RIM [204]	Label reliability based influence score scaling	Influence-based
Attent [219]	Active graph alignment	Influence-based
ALG [202]	Clustering-based density & Attention-based score	Metric-based
ALLIE [27]	Integrated graph coarsening and focal loss	RL-based
BIGENE [207]	Q-value decomposition with batch sampling selection	RL-based
IGP [203]	Information gain propagation for soft labelling	Influence-based
JuryGCN [75]	Jackknife uncertainty estimation	Influence-based

Outline for Graph Data Exploitation

- **Overview of Graph Data Exploitation**
- **Techniques with Case Studies :**
  - Graph Self-supervised Learning
  - Graph Semi-supervised Learning
  - Graph Active Learning
  - **Graph Transfer Learning**

Background of Graph Transfer Learning

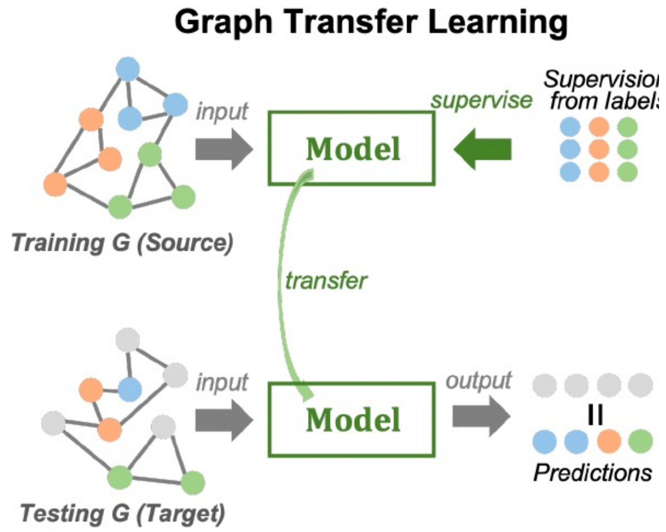
❖ **Graph Transfer Learning**

- Graph data distribution shift between the training and test graph data widely exits.
- Shifts might encompass attributes like node features, graph structures, and label distributions.

According to whether label spaces of graphs is changed or not, the category

a) **Close-set shift:** label space unchanged
b) **Open-set shift:** new label classes emerge

Fig. 9. Illustration of graph transfer learning in graph data-centric close-set shift and open-set shift.



Category of Graph Transfer Learning

- **Category from DC-GML view**

Table 8. Summary of methods in graph transfer learning.

Graph Transfer Learning	Techniques	Categories
DANE [206]	Adversarial learning regularization	Close-set shift
UDA-GCN [181]	Adversarial learning with dual-GNN	Close-set shift
ACDNE [142]	Node affinity & topological proximity preservation	Close-set shift
OpenWGL [182]	Variational graph autoencoder	Open-set shift
PGL [101]	Class space decomposition	Open-set shift
SRGNN [221]	Central moment discrepancy (CMD) measurement	Close-set shift
SOGA [104]	Mutual information maximization	Close-set shift
DGDA [14]	Domain and semantic seperation	Close-set shift
SRNC [222]	Unified domain adaption GNN	Close-set/Open-set shifts

Part 4: Frontiers of Graph Data-centric MLOps

Outline for Graph Data-centric MLOps

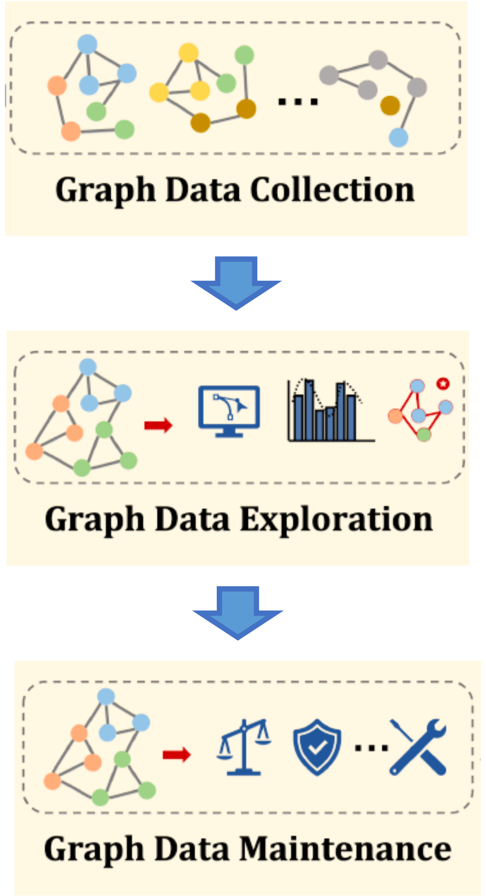
- ❖ **Overview of Graph Data-centric MLOps**
- ❖ **Techniques :**
  - Graph Data Crowdsourcing and Synthesis
  - Graph Data Understanding, Visualization, and Valuation
  - Graph Data Privacy and Security
  - Graph MLOps
- ❖ **Case Study in Graph MLOps:**

  【NeurIPS-2023】“GNNEvaluator: Evaluating GNN Performance On Unseen Graphs Without Labels”

Outline for Graph Data-centric MLOps

Phases	Goals	Methods & Tools
Graph Data Collection	Graph Data Crowdsourcing	Amazon Mechanical Turk [4], Tang et al. [152], Cao et al. [15]
	Graph Data Synthesis	SBMs [145], Koller et al. [78], Ying et al. [188], Unsupervised methods [106, 120], Semi-supervised methods [38, 111, 132, 155]
Graph Data Exploration	Graph Data Understanding & Visualization	NetworkX [31], igraph [60] Neo4j [107]
	Graph Data Valuation	GraphSVX [37]
Graph Data Maintenance	Graph Data Privacy	TrustworthyGNN [200], Zhang et al. [197], Liu et al. [92], Yu et al. [192], Mulle et al. [105], PGAS [198], Federatedscope-GNN [174], Tan et al. [151]
	Graph Data Security	Sandhu et al. [135], Abidi et al. [1], Li et al. [87]
Graph MLOps		Kubeflow [81], Amazon SageMaker [6], Amazon Neptune [179]

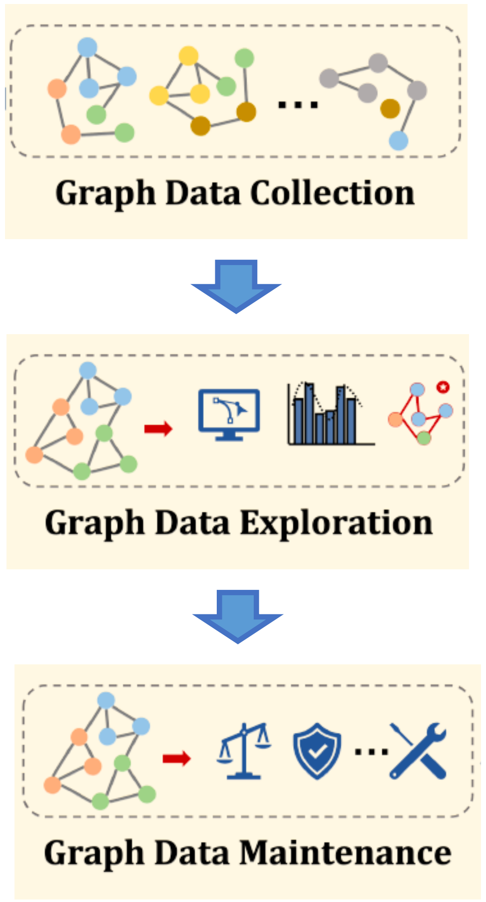
Graph data-centric view
Graph MLOps



Outline for Graph Data-centric MLOps

Phases	Goals	Methods & Tools
Graph Data Collection	Graph Data Crowdsourcing	Amazon Mechanical Turk [4], Tang et al. [152], Cao et al. [15]
	Graph Data Synthesis	SBMs [145], Koller et al. [78], Ying et al. [188], Unsupervised methods [106, 120], Semi-supervised methods [38, 111, 132, 155]
Graph Data Exploration	Graph Data Understanding & Visualization	NetworkX [31], igraph [60] Neo4j [107]
	Graph Data Valuation	GraphSVX [37]
Graph Data Maintenance	Graph Data Privacy	TrustworthyGNN [200], Zhang et al. [197], Liu et al. [92], Yu et al. [192], Mulle et al. [105], PGAS [198], Federatedscope-GNN [174], Tan et al. [151]
	Graph Data Security	Sandhu et al. [135], Abidi et al. [1], Li et al. [87]
Graph MLOps ☆		Kubeflow [81], Amazon SageMaker [6], Amazon Neptune [179]

Graph data-centric view
Graph MLOps

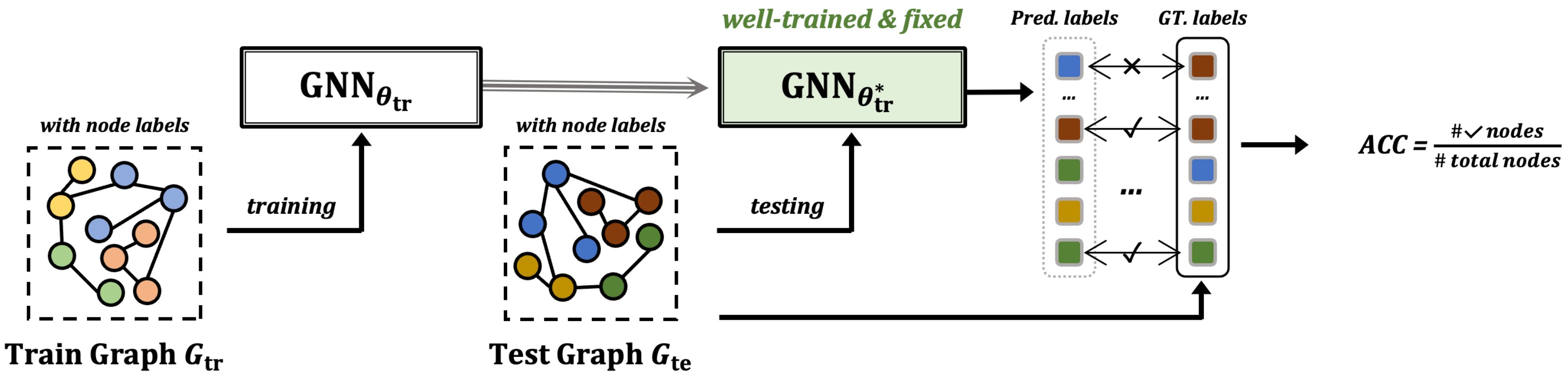


Key to practical deployment of GNNs — GNN Evaluation

Background of GNN Model Evaluation

--Case Study on Graph MLOps

Understanding and evaluating GNN models' performance is a vital step for GNN model deployment and serving.



(a) Conventional GNN Model Evaluation

For instance,

in financial transaction networks:

- **GNN model designers:** expect their developed GNNs to excel in identifying newly emerging suspicious transactions
- **Users:** ensure how they could trust well-trained GNNs to know suspicious transactions within their own data

In conventional model evaluation of GNNs, we have:

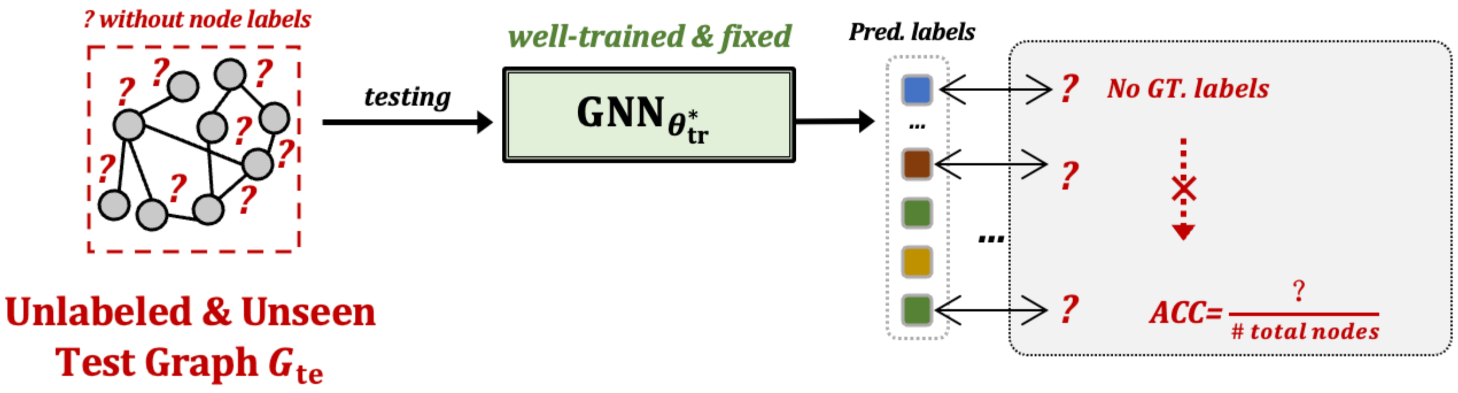
1) Seen test graph $\boldsymbol{G}_{\mathbf{te}}$ in the same distribution as the train graph $\boldsymbol{G}_{\mathbf{tr}}$
2) Known test graph labels for computing performance metric, e.g., Accuracy (ACC)

Zheng, X., Zhang, M., Chen, C., Molaei, S., Zhou, C., & Pan, S. (2023). GNNEvaluator: Evaluating GNN Performance On Unseen Graphs Without Labels. Advances in Neural Information Processing Systems (NeurIPS), 2023.

Background of GNN Model Evaluation

--Case Study on Graph MLOps

However, in real-world scenarios, the test graphs are typically **"unseen & lacking annotations"**



(b) Real-world GNN Model Evaluation

In real-world model evaluation of GNNs, we:

- X CAN NOT access the ground-truth labels of the test graph $\boldsymbol{G}_{\mathbf{te}}$
- X CAN NOT compute performance metric, e.g., Accuracy (ACC)
- X DO NOT know whether potential distribution shifts from the train graph $\boldsymbol{G}_{\mathbf{tr}}$

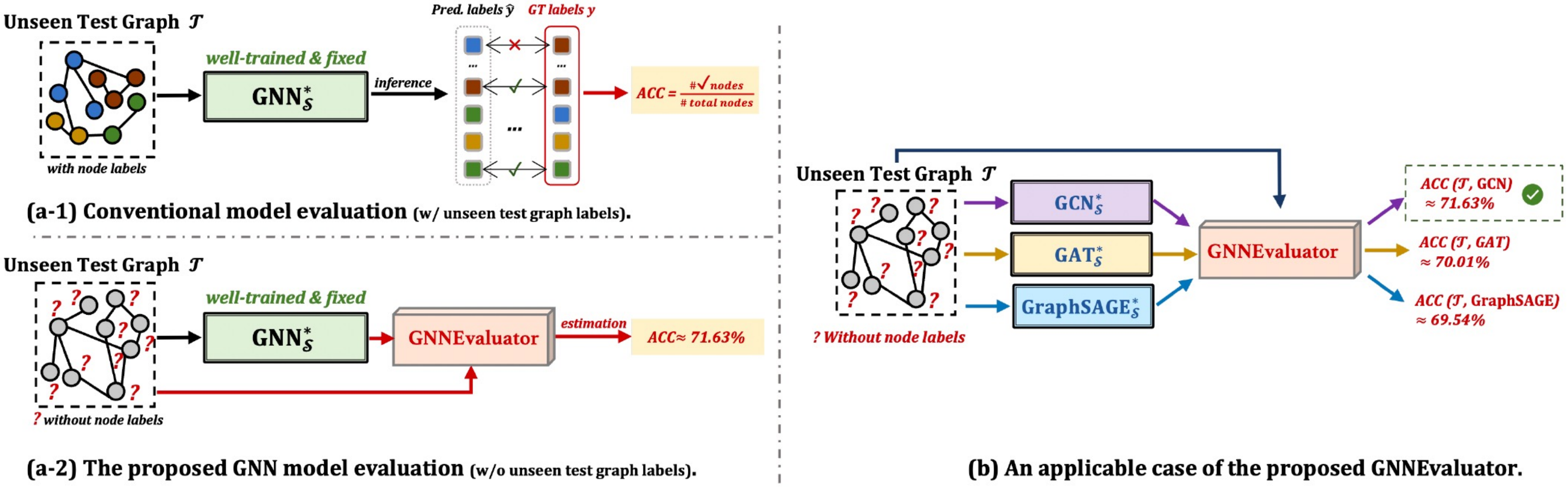
Zheng, X., Zhang, M., Chen, C., Molaei, S., Zhou, C., & Pan, S. (2023). GNNEvaluator: Evaluating GNN Performance On Unseen Graphs Without Labels. Advances in Neural Information Processing Systems (NeurIPS), 2023.

Background of GNN Model Evaluation

--Case Study on Graph MLOps

Given above scenarios, a natural question, i.e., **<u>"GNN model evaluation problem"</u>** arises:

In the absence of labels in an unseen test graph, can we estimate the performance of a well-trained GNN model?



Zheng, X., Zhang, M., Chen, C., Molaei, S., Zhou, C., & Pan, S. (2023). GNNEvaluator: Evaluating GNN Performance On Unseen Graphs Without Labels. Advances in Neural Information Processing Systems (NeurIPS), 2023.

Definition of GNN Model Evaluation

--Case Study on Graph MLOps

Definition of GNN Model Evaluation. Given the observed training graph $\mathcal{S}$, its well-trained model $\text{GNN}^*_{\mathcal{S}}$, and an unlabeled unseen graph $\mathcal{T}$ as inputs, the **goal** of GNN model evaluation aims to learn an accuracy estimation model $f_\phi(\cdot)$ parameterized by ϕ as:

$$\text{Acc}(\mathcal{T}) = f_\phi(\text{GNN}^*_{\mathcal{S}}, \mathcal{T}), \tag{2}$$

where $f_\phi : (\text{GNN}^*_{\mathcal{S}}, \mathcal{T}) \to a$ and $a \in \mathbb{R}$ is a scalar denoting the overall node classification accuracy $\text{Acc}(\mathcal{T})$ for all unlabeled nodes of $\mathcal{T}$. When the context is clear, we will use $f_\phi(\mathcal{T})$ for simplification.

To solve above problems,

We propose a two-stage GNN model evaluation framework with a "GNNEvaluator"

Note that our principal goal is to estimate well-trained GNN models' performance, rather than improve the generalization ability of new GNN models. In the whole evaluation process, the in-service GNN model is fixed

Zheng, X., Zhang, M., Chen, C., Molaei, S., Zhou, C., & Pan, S. (2023). GNNEvaluator: Evaluating GNN Performance On Unseen Graphs Without Labels. Advances in Neural Information Processing Systems (NeurIPS), 2023.

GNNEvaluator: Evaluating GNN Performance On Unseen Graphs Without Labels

--Case Study on Graph MLOps

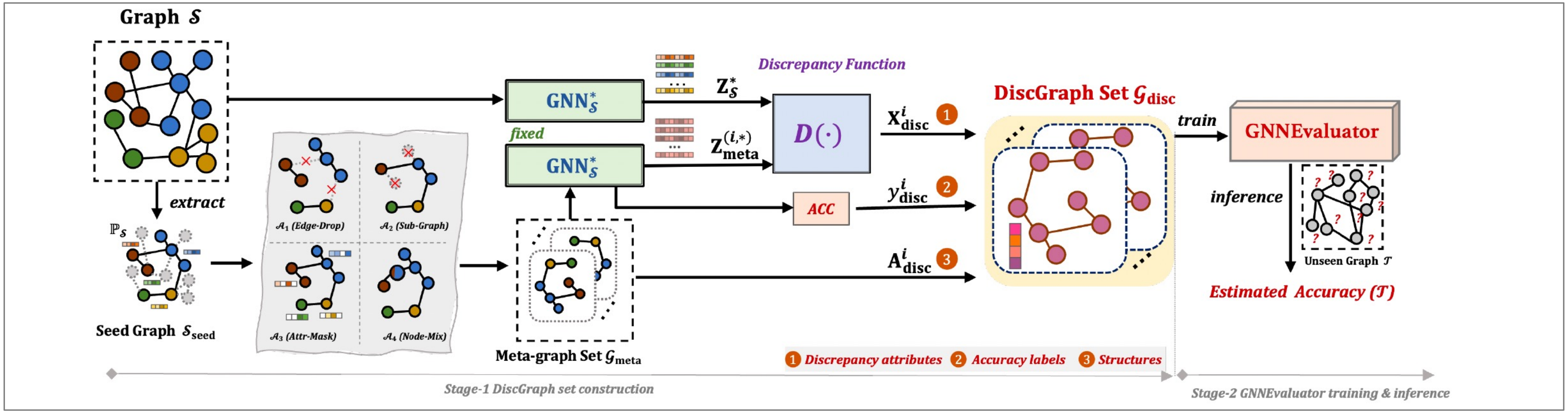


Figure.1 Overall two-stage framework of the proposed GNN model evaluation with GNNEvaluator

- **Stage-1: DiscGraph set construction**

  incorporating training-test graph discrepancies into DiscGraph node attributes $\mathbf{X}^i_{\text{disc}}$, structures $\mathbf{A}^i_{\text{disc}}$, and accuracy labels y^i_{disc}

- **Stage-2: GNNEvaluator training and inference**

  GNNEvaluator, train on DiscGraphs and output estimated ACC on the real-world test graph $\mathcal{T}$

Zheng, X., Zhang, M., Chen, C., Molaei, S., Zhou, C., & Pan, S. (2023). GNNEvaluator: Evaluating GNN Performance On Unseen Graphs Without Labels. Advances in Neural Information Processing Systems (NeurIPS), 2023.

Experiments on GNNEvaluator

--Case Study on Graph MLOps

The performance of our proposed GNNEvaluator in evaluating well-trained GNNs' node classification accuracy under all test evaluation cases and models

Table 1: Mean Absolute Error (MAE) performance of different GNN models across five random seeds. (GNNs are well-trained on the ACMv9 dataset and evaluated on the unseen and unlabeled Citationv2 and DBLPv8 datasets, *i.e.*, A→C and A→D, respectively. Best results are in bold.)

Methods	ACMv9→Citationv2						ACMv9→DBLPv8					
	GCN	SAGE	GAT	GIN	MLP	*Avg.*	GCN	SAGE	GAT	GIN	MLP	*Avg.*
ATC-MC [8]	4.49	8.40	4.37	18.40	34.33	14.00	21.96	24.20	30.30	24.06	26.62	25.43
ATC-MC-c [8]	**2.41**	5.74	4.67	22.00	51.41	17.25	31.15	30.55	30.18	29.71	45.81	33.48
ATC-NE [8]	3.97	8.02	4.28	17.35	38.87	14.50	22.93	24.78	30.50	23.74	31.13	26.62
ATC-NE-c [8]	4.44	6.09	**3.30**	23.95	44.62	16.48	34.42	28.31	27.02	30.28	39.28	31.86
Thres. ($\tau = 0.7$) [6]	32.64	35.81	33.63	50.76	35.28	37.63	9.59	12.14	14.30	32.67	39.72	21.68
Thres. ($\tau = 0.8$) [6]	26.30	29.60	26.18	49.25	35.87	33.44	**2.63**	7.44	14.47	32.20	40.31	19.41
Thres. ($\tau = 0.9$) [6]	17.56	21.34	16.38	46.53	36.08	27.58	8.20	7.42	16.07	31.47	40.56	20.74
AutoEval-G [6]	18.94	26.19	26.12	50.86	32.40	30.90	2.77	**2.54**	7.25	48.68	29.95	18.24
`GNNEvaluator` **(Ours)**	4.85	**4.11**	12.23	**10.14**	**22.20**	**10.71**	11.80	14.88	**6.36**	**13.78**	**17.49**	**12.86**

Table 2: Mean Absolute Error (MAE) performance of different GNN models across five random seeds. (GNNs are well-trained on the Citationv2 dataset and evaluated on the unseen and unlabeled ACMv9 and DBLPv8 datasets, i.e., C→A and C→D, respectively.Best results are in bold.)

Methods	Citationv2→ACMv9						Citationv2→DBLPv8					
	GCN	SAGE	GAT	GIN	MLP	*Avg.*	GCN	SAGE	GAT	GIN	MLP	*Avg.*
ATC-MC [8]	9.50	13.40	8.28	35.51	43.40	22.02	22.57	**1.37**	21.87	29.24	35.20	22.05
ATC-MC-c [8]	6.93	11.75	**6.70**	38.93	57.43	24.35	33.67	4.92	28.23	30.89	52.59	30.06
ATC-NE [8]	8.86	13.04	7.87	34.88	47.49	22.42	23.97	1.86	23.74	28.96	39.72	23.65
ATC-NE-C [8]	7.73	13.94	7.63	41.17	62.96	26.69	37.16	4.66	29.43	31.66	58.95	32.37
Thres. ($\tau = 0.7$) [6]	37.33	36.61	31.68	58.91	34.33	39.77	10.70	23.05	12.74	34.60	38.29	23.88
Thres. ($\tau = 0.8$) [6]	29.62	28.95	22.77	57.48	34.53	34.67	5.65	15.01	7.61	34.36	38.43	20.21
Thres. ($\tau = 0.9$) [6]	19.59	19.06	11.37	55.72	34.56	28.06	10.65	8.28	8.07	34.00	38.44	19.89
AutoEval-G [6]	23.01	31.24	26.74	59.66	35.02	28.28	**2.57**	16.52	6.96	19.20	32.24	24.59
`GNNEvaluator` **(Ours)**	**5.45**	**8.53**	9.61	**29.77**	**28.52**	**16.38**	11.64	7.02	**5.58**	**6.46**	**22.87**	**10.71**

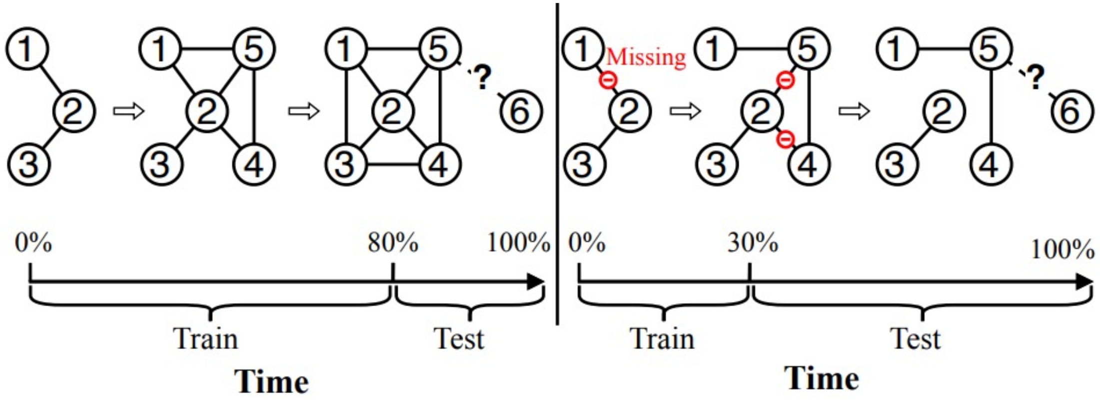
- Experiments on 3 real-world graph datasets in 6 cases potential domain shift, each evaluating 5 models:
- *Consistent outstanding performance over all GNN models and cases!*

Zheng, X., Zhang, M., Chen, C., Molaei, S., Zhou, C., & Pan, S. (2023). GNNEvaluator: Evaluating GNN Performance On Unseen Graphs Without Labels. Advances in Neural Information Processing Systems (NeurIPS), 2023.

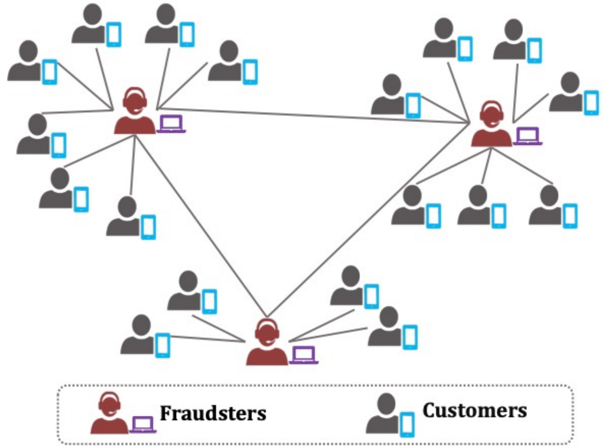
Part 5: Future Directions & Conclusion

Promising Future Directions

❖ Exploration of complex and dynamic graph data



Dynamic Graph



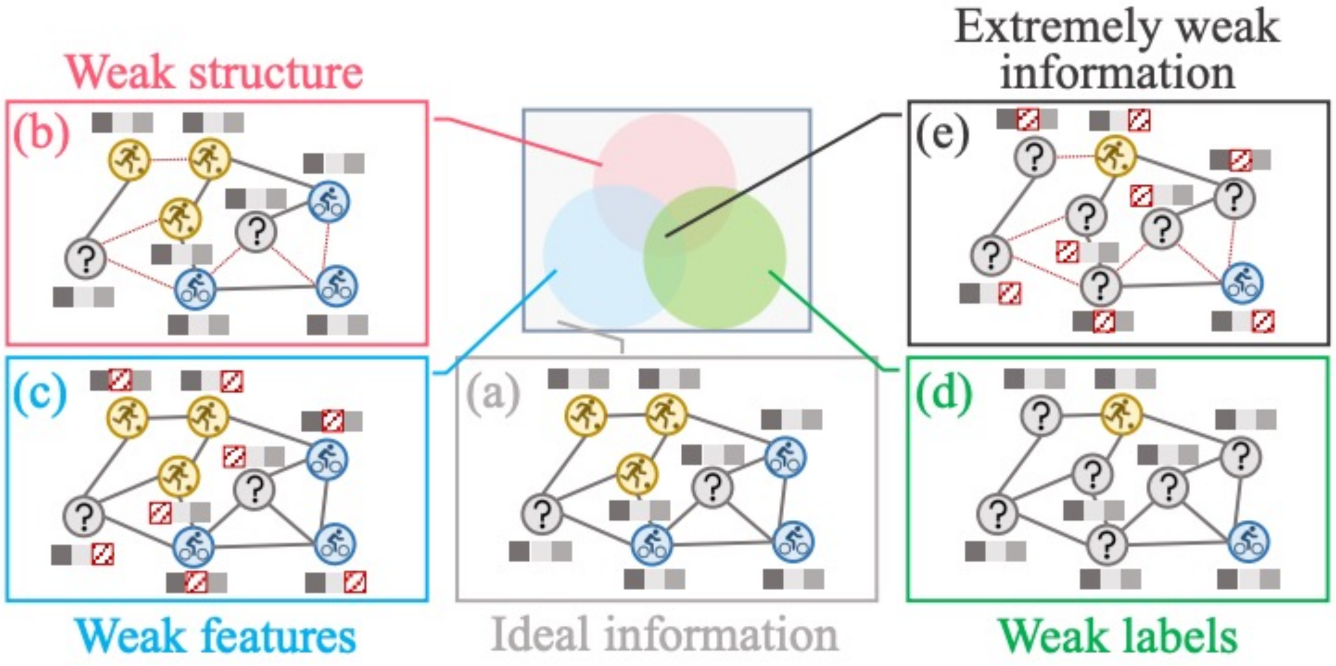
Heterophilic Graph

[1] Luo, L., Haffari, G., & Pan, S. (2023, February). Graph sequential neural ode process for link prediction on dynamic and sparse graphs. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (pp. 778-786).

[2] Zheng, X., Liu, Y., Pan, S., Zhang, M., Jin, D., & Yu, P. S. (2022). Graph neural networks for graphs with heterophily: A survey. arXiv preprint arXiv:2202.07082.

Promising Future Directions

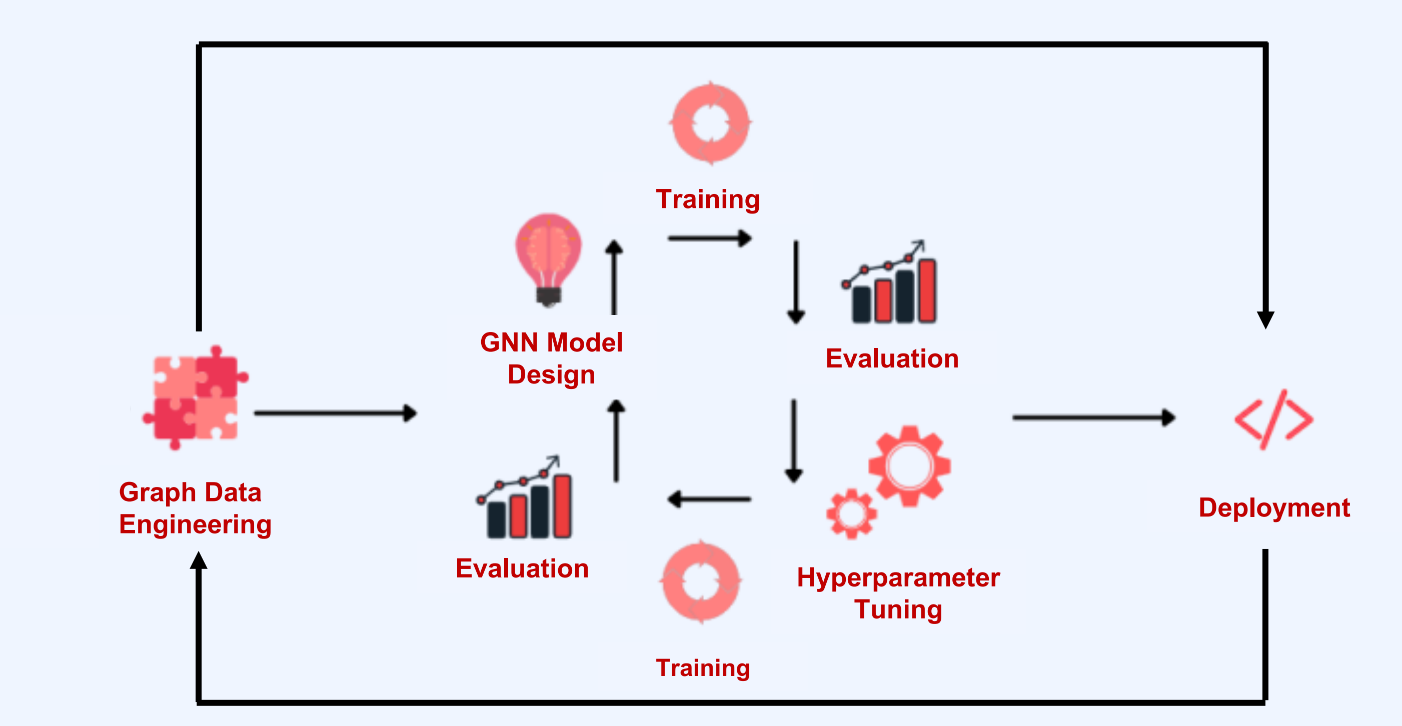
❖ **General and automatic graph data improvement.**



Liu, Y., Ding, K., Wang, J., Lee, V., Liu, H., & Pan, S. (2023). Learning Strong Graph Neural Networks with Weak Information. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23).

Promising Future Directions

- ❖ **Standardized graph data benchmarks**
- ❖ **Collaborative development of graph data and model**
- ❖ **Comprehensive graph data lifecycle management pipelines**

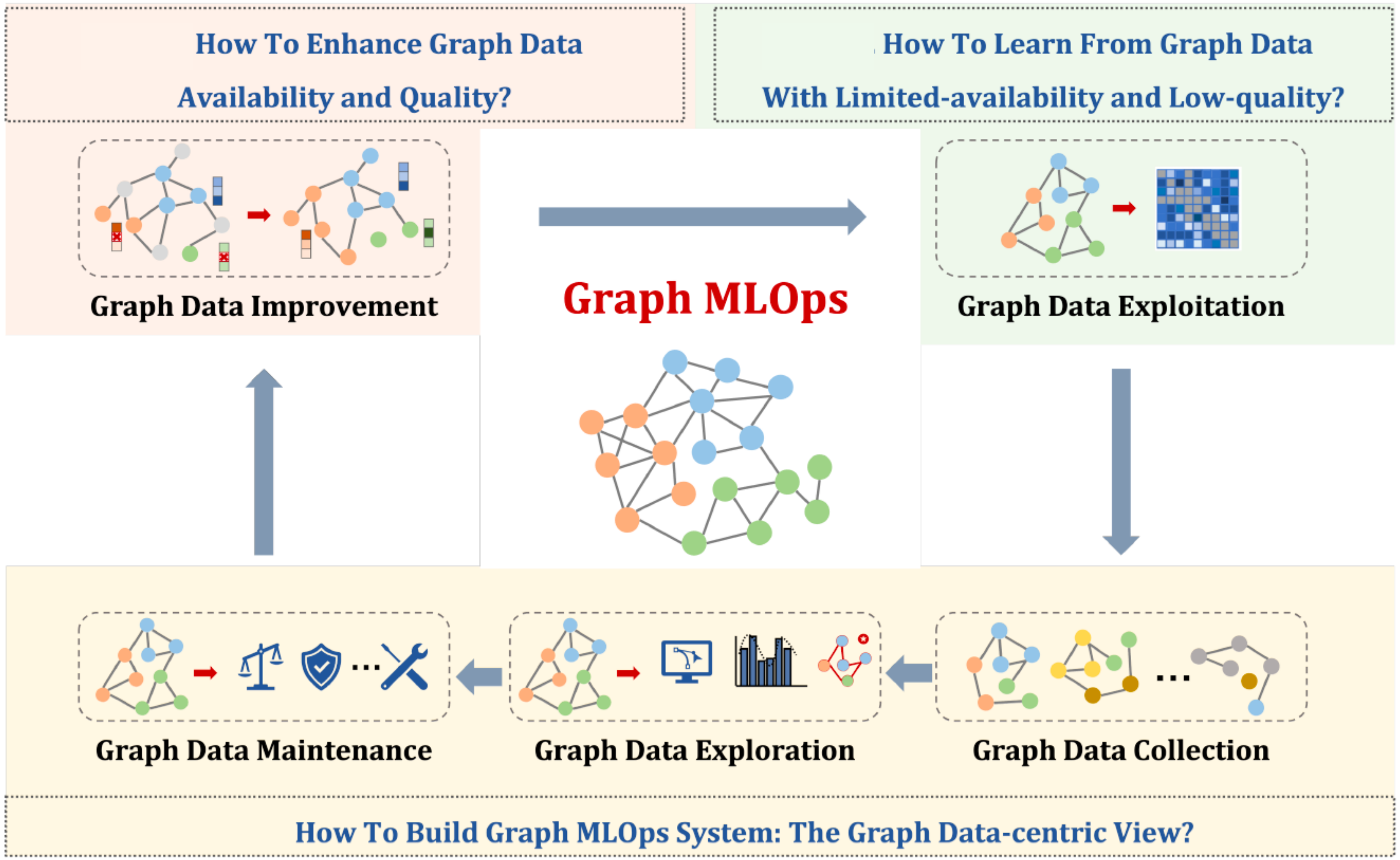


Promising Future Directions

- **Exploration of complex and dynamic graph data**
- **General and automatic graph data improvement**
- **Standardized graph data benchmarks**
- **Collaborative development of graph data and model**
- **Comprehensive graph data lifecycle management pipelines**

Conclusion

Promising Data-centric Graph Machine Learning (DC-GML)



Conclusion

Three Core Research Questions

RQ1: How to enhance graph data availability and quality?

↓

Enhanced Graph Data

RQ2: How to learn from graph data with limited-availability and low-quality?

↓

Enhanced Graph Data → *Enhanced Graph ML Models*

RQ3: How to build graph MLOps systems from the graph data-centric view?

↓

Systematic & Comprehensive
Data-centric Graph Machine Learning (DC-GML)

Conclusion

Comprehensive Taxonomy

Graph Data Improvement

- **Graph Structure Enhancement**
  - *Graph Structure Learning || Graph Sparsification || Graph Diffusion*
- **Graph Feature Enhancement**
  - *Graph Feature Completion || Graph Feature Denoising*
- **Graph Label Enhancement**
  - *Graph Pseudo-labelling || Graph Label Denoising || Graph Class-imbalanced Sampling*
- **Graph Size Enhancement**
  - *Graph Size Reduction || Graph Data Augmentation*

Graph Data Exploitation

- **Graph Self-supervised Learning**
- **Graph Semi-supervised Learning**
- **Graph Active Learning**
- **Graph Transfer Learning**

Graph Data Collection

- **Graph Data Crowdsourcing & Synthesis**

Graph Data Exploration

- **Graph Data Understanding, Visualization & Valuation**

Graph Data Maintenance

- **Graph Data Privacy & Security**

Fig. 2. The framework and taxonomy of data-centric graph machine learning (DC-GML).

Conclusion

Extensive & Open Potentials of DC-GML

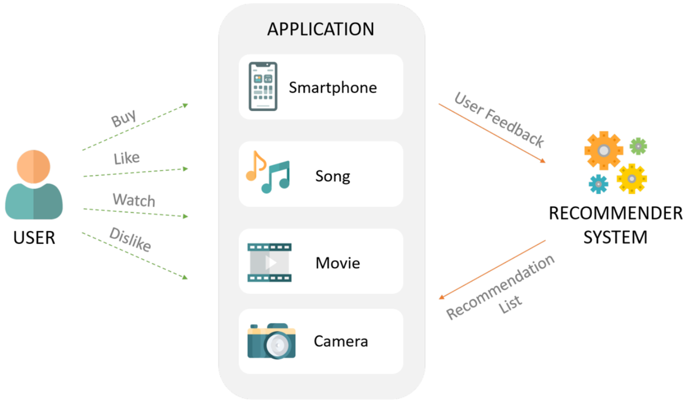
A. Standardized graph machine learning workflow

B. Enhanced graph data understanding

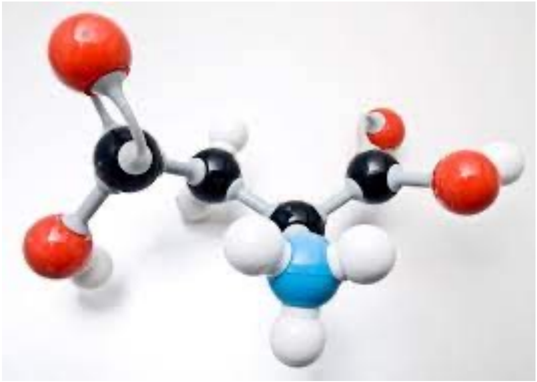
C. Better graph learning model performance

D. Wider graph data application range

Recommender Systems



Chemistry



… continual and broader applications in DC-GML…

Thanks!

Towards Data-centric Graph Machine Learning

Xin Zheng[1], Shirui Pan[2]

[1] Monash University

[2] Griffith University

Data-centric Graph ML Review & Outlook

DC-GML GitHub Collection