Imbalanced Learning on Graphs:

Problems, Techniques, and Future Directions

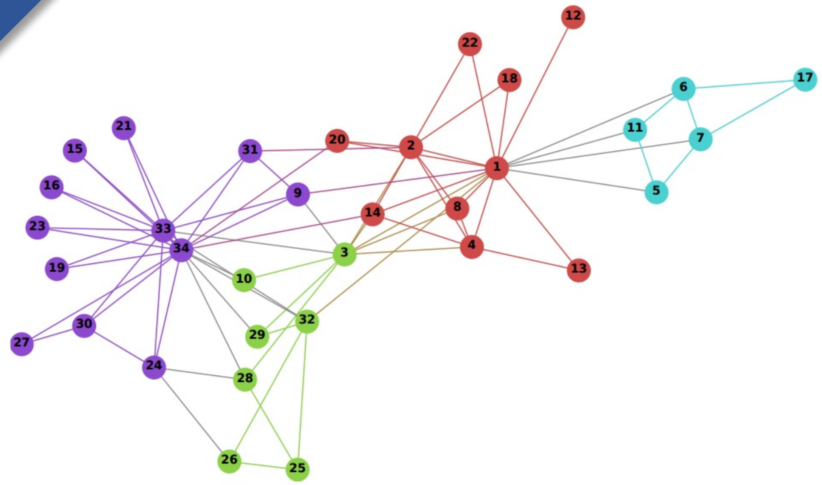
Zemin Liu
Senior Research Fellow
National University of Singapore
zeminliu@nus.edu.sg

Survey Link: https://arxiv.org/abs/2308.13821
GitHub Link: https://github.com/Xtra-Computing/Awesome-Literature-ILoGs

Outline

1. Introduction to Imbalanced Learning on Graphs (ILoGs)
2. Background
3. Overview of Taxonomies
4. Problems of ILoGs
5. Techniques of ILoGs
6. Future Directions
7. Conclusions

Graph Formalization

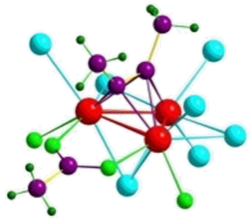


Graph. A **graph** can be represented as $G = \{\mathcal{V}, \mathcal{E}, \mathbf{X}_v, \mathbf{X}_e, \phi, \varphi, \mathcal{T}, \mathcal{R}\}$, where $\mathcal{V}$ is the set of nodes, $\mathcal{E}$ is the set of edges, $\mathbf{X}_v \in \mathbb{R}^{|\mathcal{V}| \times d_{\mathbf{x}_v}}$ and $\mathbf{X}_e \in \mathbb{R}^{|\mathcal{E}| \times d_{\mathbf{x}_e}}$ are the feature matrices of nodes and edges, respectively, and $\mathcal{T}$ and $\mathcal{R}$ are the sets of node types and edge types. For simplicity, we utilize $\mathbf{x}_v \in \mathbb{R}^{d_{\mathbf{x}_v}}$ and $\mathbf{x}_e \in \mathbb{R}^{d_{\mathbf{x}_e}}$ to denote the feature vectors of node v and edge e, respectively. The function $\phi : \mathcal{V} \rightarrow \mathcal{T}$ maps a node $v \in \mathcal{V}$ to its corresponding node type $\phi(v)$, while the function $\varphi : \mathcal{E} \rightarrow \mathcal{R}$ maps an edge $e_{\langle u,v \rangle} \in \mathcal{E}$ to its corresponding edge type $\varphi(e_{\langle u,v \rangle})$.

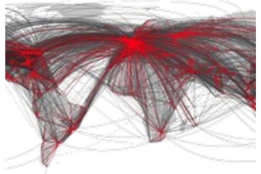
Graph Representation Learning

Graph Data

Social Networks

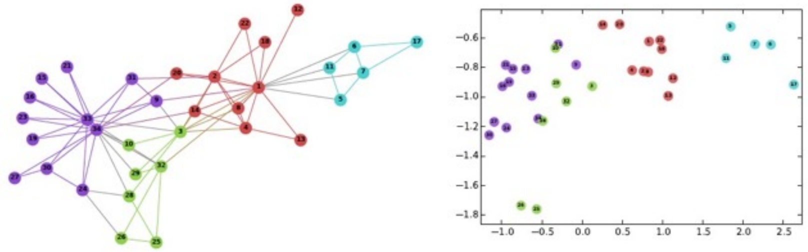


Molecular Networks

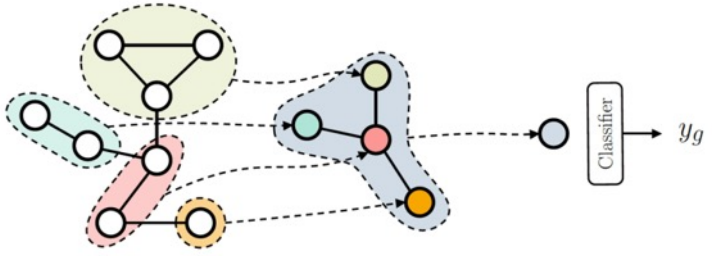


Transportation Networks

Graph Representation Learning



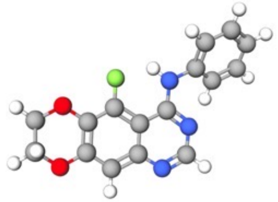
Node-Level Representation Learning



Graph-Level Representation Learning

Applications

Public Opinion Analysis



Drug Design

Transportation Prediction

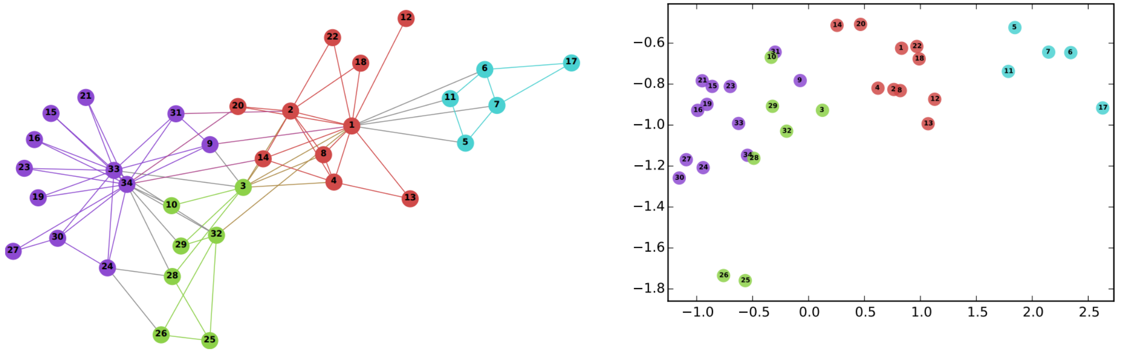
Graph Representation Learning Techniques

- Graph embedding approaches
  - DeepWalk [a], node2vec [b], …

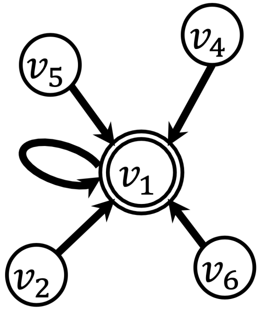
- Graph neural networks (GNNs) [c,d,e]

$$\mathbf{h}_v^l = \mathcal{M}(\mathbf{h}_v^{l-1}, \{\mathbf{h}_i^{l-1} : i \in \mathcal{N}_v\}; \theta^l)$$

Message passing function



(a) Input: Karate Graph (b) Output: Representation



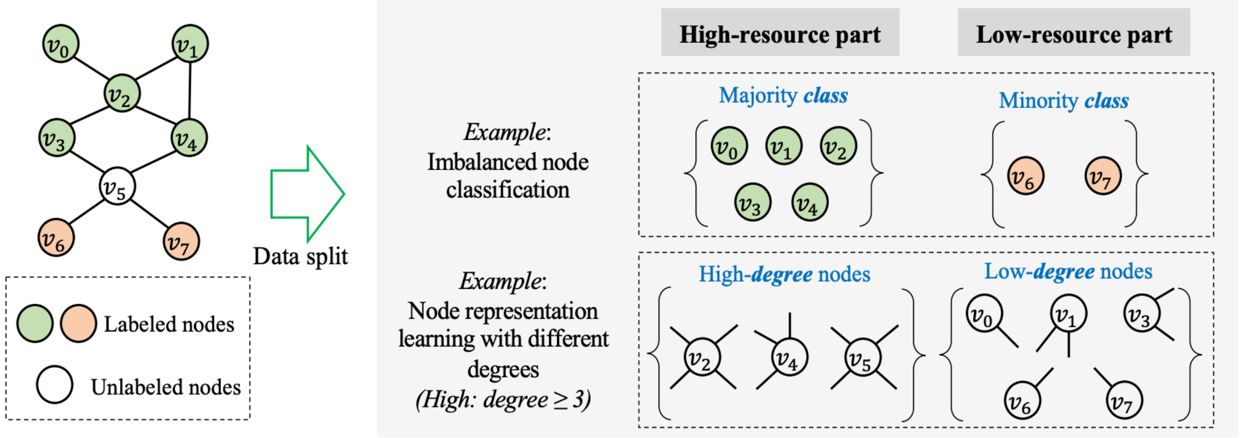
Neighborhood aggregation

[a] Perozzi B., et al. 2014. Deepwalk: Online learning of social representations. KDD.
[b] Grover A., et al. 2014. node2vec: Scalable feature learning for networks. KDD.
[c] Kipf, T. N., et al. 2017. Semi-supervised classification with graph convolutional networks. ICLR.
[d] Veličković, P., et al. 2018. Graph attention networks. ICLR.
[e] Hamilton W L., et al. 2017. Inductive representation learning on large graphs. NeurIPS.

Imbalance Phenomenon

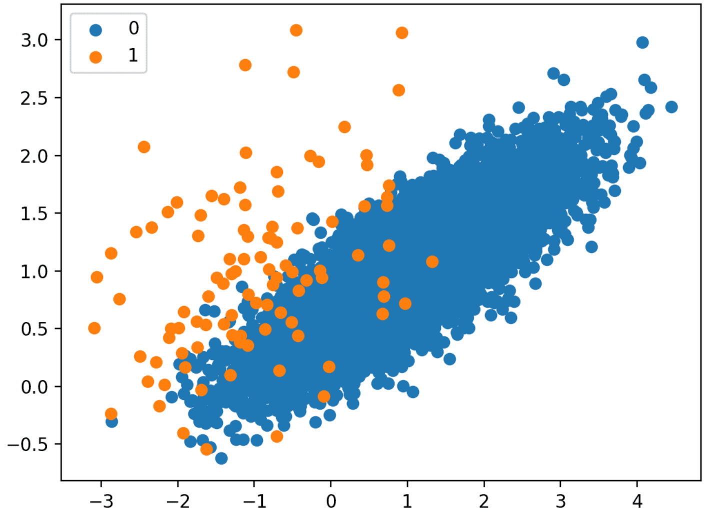
- **Information Abundance**
  - distributed differently across groups
    - *e.g.*, imbalanced classes: large classes *vs* small classes

- **High-resource** groups *vs* **Low-resource** groups
  - High-resource groups
    - Abundant data information
    - (Usually) High performance
  - Low-resource groups
    - Limited data information
    - (Usually) Low performance



(a) Input graph

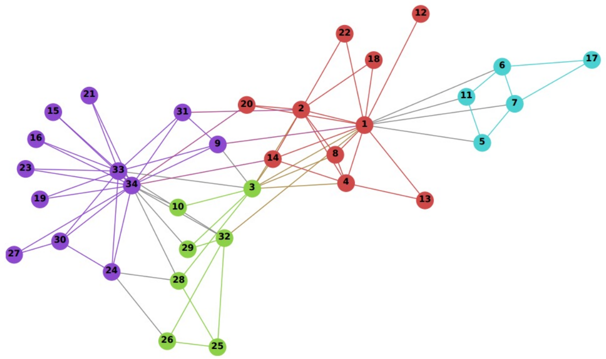
(b) Input with imbalanced graph resource distribution



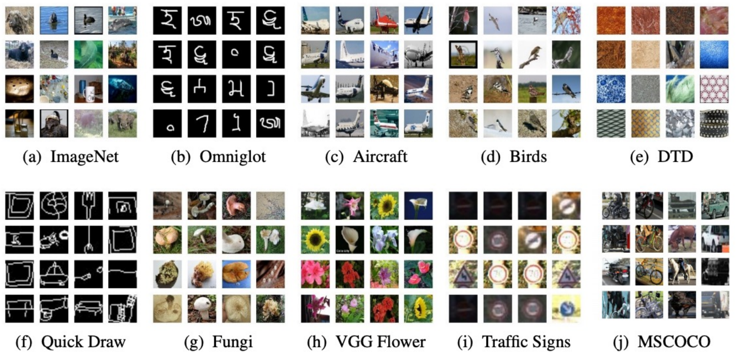
Imbalanced Learning on Graphs (ILoGs): Motivation

- **Graph data**
  - Different from vision and language data
  - Non-*i.i.d.*
  - Multifarious
    - Class, structure
- **Increasing volume of literature** on ILoGs
  - Problems
  - Techniques

  - Lacking a comprehensive framework to identify the commonalities and disparities

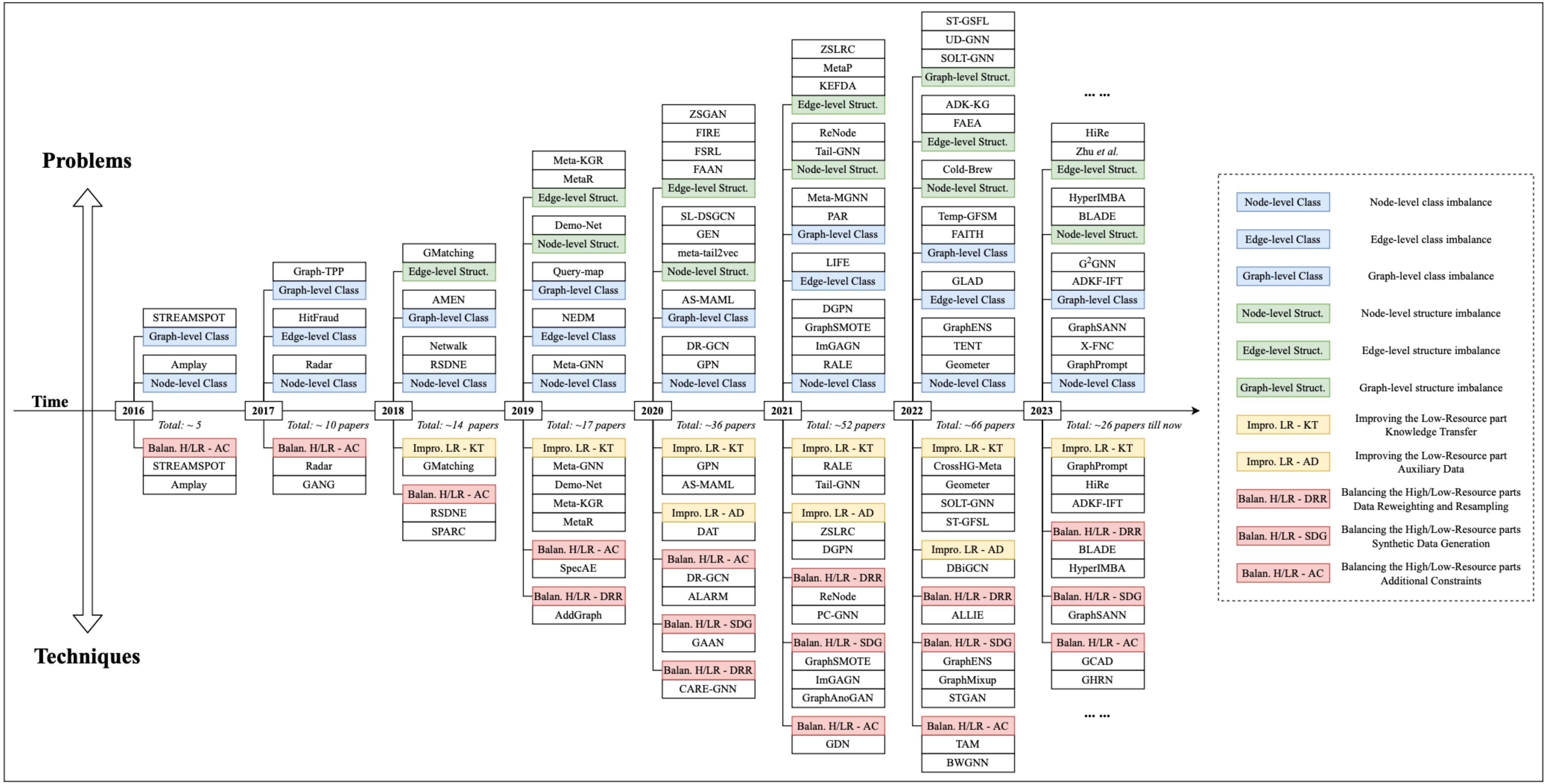


A graph dataset.



Some image datasets.

The Timeline of this Task



Imbalanced Learning on Graphs (ILoGs)

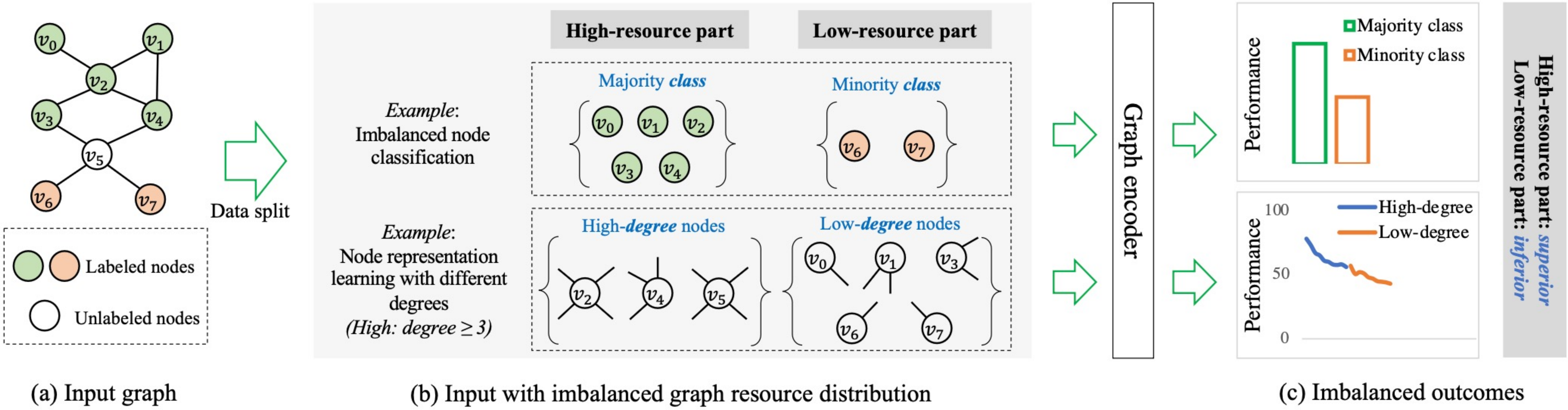


Fig. 2: Imbalanced learning on graphs: imbalanced graph resource distribution results in imbalanced outcomes.

Challenges and Solutions

- **Challenges**
  - Graph content -> a wide array of imbalance problems
    - *How to create an organized taxonomy to categorize these imbalance problems on graphs?*
  - Imbalance problems -> different techniques
    - *How to classify the literature from a technical perspective?*

- **Solutions**
  - Taxonomies
    - Problems and Techniques
  - Taxonomy of Problems
    - Class Imbalance and Structure Imbalance
    - Node-Level, Edge-Level, and Graph-Level
  - Taxonomy of Techniques
    - Imbalance types
      - *What imbalance types?*
    - The techniques to cope with each imbalance type
      - *How to cope with each type?*

Relationship with Existing Surveys

- **Imbalanced Learning Surveys**
  - Imbalanced classification
  - Few-shot learning
    - Anomaly detection
  - Long-tailed distribution
  - Characteristics:
    - Focus on imbalanced learning in a general context of in specific tasks, and lack comprehensive coverage of imbalanced learning on graphs

- **Graph-Related Imbalanced Learning Surveys**
  - Class-Imbalanced Learning
    - Anomaly detection
  - Few-shot classification
  - Fairness learning
  - Characteristics:
    - Focus on individual tasks and lack a comprehensive overview of imbalanced learning on graphs

- **Our Survey**
  - Provide a holistic view of imbalanced learning on graphs
    - Covering diverse tasks with a focus on both class imbalance and structure imbalance
  - Elucidate the shared traits and unique characteristics of the tasks
    - Offering fresh insights into their commonalities and differences within the sphere of imbalanced learning on graphs

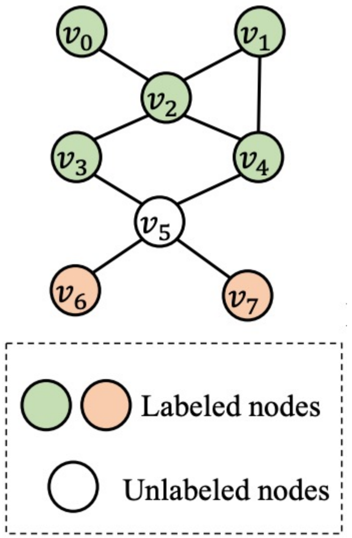
Contributions

- The first comprehensive survey of ILoGs
  - Serve as invaluable resource for both researchers and practitioners
- We propose two novel taxonomies
  - Problems and techniques
  - Facilitate a thorough understanding of existing literature
  - Provide a clear picture of the commonalities and distinctions
- Identify potential future research directions
  - Provide insights and guidance for those interested in advancing the SOTA in this fast-paced filed
- **Scenarios**
  - The scenarios that graph learning algorithms can involve
    - Imbalance is prevalent in the real-world scenarios

Outline

1. Introduction to Imbalanced Learning on Graphs (ILoGs)
2. Background
3. Overview of Taxonomies
4. Problems of ILoGs
5. Techniques of ILoGs
6. Future Directions
7. Conclusions

(Conventional) Imbalanced Learning



- Focus on handling **imbalanced classes**

Definition 1 (Conventional Imbalanced Learning). *In the context of conventional imbalanced learning, consider a labeled data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, which can be partitioned into K classes (groups) such that $\mathcal{D} = \bigcup_{1 \le j \le K} \mathcal{G}_j$ (given each group $\mathcal{G}_j = \{(x_i, y_i) : y_i = j\}$). There exists a notable imbalance in the number of labeled samples across these groups. Under this setting, the imbalanced distribution of samples across groups would give rise to biases in the performance of a learning algorithm. In particular, the low-resource groups, i.e., the classes with less labeled data, are usually marginalized by the learning model due to the domination of the high-resource groups, resulting in performance degradation for the former. The goal of imbalanced learning is to develop a balanced model that can improve the performance of low-resource groups, potentially reaching levels comparable to those of high-resource groups.*

Imbalance Ratio

Given K classes

$$\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{G}_i$$

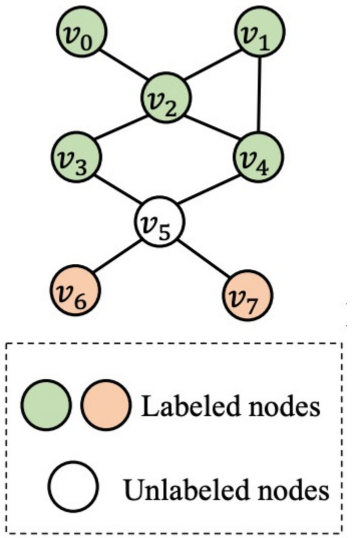
Order them in descending manner

$$\text{if } i < j, \text{ then } |\mathcal{G}_i| \ge |\mathcal{G}_j|$$

Imbalance ratio

$$|\mathcal{G}_1| / |\mathcal{G}_K|$$

Imbalanced Learning on Graphs



Definition 2 (Imbalanced Learning on Graphs). *In addition to the number of labeled instances, imbalance in graph data can stem from disparities in structural abundance across groups, leading to a more complex imbalance pattern. For a given graph dataset comprising a set of elements (i.e., nodes, edges, or (sub)graphs) represented as* $\mathcal{G} = \{x_i\}_{i=1}^{N}$, *these elements can be further grouped into* K *subsets, i.e.,* $\mathcal{G} = \bigcup_{1 \le i \le K} \mathcal{G}_i$, *according to specific criteria based on classes or structures, where* $1 < K \le N$. *It is important to note that these groups differ in terms of information abundance, which results in unequal performance among them when used as input for a learning model.*

Imbalance Ratio

Given K groups

$$\mathcal{G} = \{x_i\}_{i=1}^{N}, \text{ i.e., } \mathcal{G} = \bigcup_{1 \le i \le K} \mathcal{G}_i$$

Order them in descending manner

$$\text{if } i < j, \text{ then } s_{\mathcal{G}_i} \ge s_{\mathcal{G}_j}$$

Imbalance ratio:

$$s_{\mathcal{G}_1} / s_{\mathcal{G}_K}$$

What should be done if $s_{\mathcal{G}_i}$ is noncountable?

Existing Graph Imbalance Issues

Imbalance Types	Imbalance Tasks	Settings	Information Abundance s	Explanations
Node-Level Class Imbalance	Imbalanced node classification Node-level anomaly detection	A set of (or two) *node classes* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{C}_i$	$\lvert\mathcal{C}_i\rvert$ (# labeled nodes in each class $\mathcal{C}_i$)	Labeled nodes are unevenly distributed across classes.
	Few-shot node classification Zero-shot node classification	A set of base *node classes* $\mathcal{D}_b = \bigcup_{1 \le i \le K_1} \mathcal{C}_i$, and novel *node classes* $\mathcal{D}_n = \bigcup_{K_1 < i \le K_2} \mathcal{C}_i$	$\lvert\mathcal{C}_i\rvert$ (# labeled nodes in each class $\mathcal{C}_i$)	Base classes have abundant labeled nodes, while novel classes have few/no labeled nodes.
Edge-Level Class Imbalance	Few-shot link prediction	A set of base *graphs* $\mathcal{D}_b = \{G_i\}_{i=1}^{K_1}$ and novel *graphs* $\mathcal{D}_n = \{G_i\}_{i=K_1+1}^{K_2}$, where $G_i = \{\mathcal{V}_i, \mathcal{E}_i\}$	$\lvert\mathcal{E}_i\rvert$ (# edges in each graph G_i)	Base graphs have abundant edges, while novel graphs have limited edges.
	Edge-level anomaly detection	Two *edge classes* $\mathcal{D} = \mathcal{C}_1 \cup \mathcal{C}_2$	$\lvert\mathcal{C}_i\rvert$ (# labeled edges in each class $\mathcal{C}_i$)	Labeled edges are unevenly distributed across classes.
Graph-Level Class Imbalance	Imbalanced graph classification Graph-level anomaly detection	A set of (or two) *graph classes* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{C}_i$	$\lvert\mathcal{C}_i\rvert$ (# labeled graphs in each class $\mathcal{C}_i$)	Labeled graphs are unevenly distributed across classes.
	Few-shot graph classification	A set of base *graph classes* $\mathcal{D}_b = \bigcup_{1 \le i \le K_1} \mathcal{C}_i$, and novel *node classes* $\mathcal{D}_n = \bigcup_{K_1 < i \le K_2} \mathcal{C}_i$	$\lvert\mathcal{C}_i\rvert$ (# labeled graphs in each class $\mathcal{C}_i$)	Base classes have abundant labeled graphs, while novel classes have few labeled graphs.
Node-Level Structure Imbalance	Imbalanced node degrees	A set of *node groups* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{G}_i$, where $\mathcal{G}_i = \{v_j : d_j = i\}$ (d_j is the degree of node v_j)	d_j (the degree of each node v_j)	Head nodes have high degrees, while tail/cold-start nodes have few/no degrees.
	Node topology imbalance	A set of *node classes* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{C}_i$	The consistency between true class boundaries and influence boundaries of labeled nodes	Classes with more consistent boundaries tend to propagate label information more effectively.
	Long-tail entity embedding	A set of *entity groups* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{G}_i$, where $\mathcal{G}_i = \{e_j : d_j = i\}$ (d_j is the # triplets of entity e_j)	d_j (# triplets of each entity e_j)	Head entities have more triplets, while tail/cold-start entities have few/no triplets.
Edge-Level Structure Imbalance	Few-shot relation classification Zero-shot relation classification Few-shot reasoning on KGs	A set of base *relations* $\mathcal{D}_b = \bigcup_{1 \le i \le K_1} \mathcal{R}_i$, and novel *relations* $\mathcal{D}_n = \bigcup_{K_1 < i \le K_2} \mathcal{R}_i$	$\lvert\mathcal{R}_i\rvert$ (# labeled triplets of each relation $\mathcal{R}_i$)	Base relations have abundant labeled triplets, while novel relations have few/no labeled triplets.
Graph-Level Structure Imbalance	Imbalanced graph sizes	A set of *graph groups* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{G}_i$, where $\mathcal{G}_i = \{G_j : \lvert\mathcal{V}_j\rvert = i\}$ ($\lvert\mathcal{V}_j\rvert$ is the size of graph G_j)	$\lvert\mathcal{V}_j\rvert$ (the size of each graph G_j)	Head graph have large sizes, while tail graphs have small sizes.
	Imbalanced topology groups	A set of *topology motifs* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{M}_i$	$\lvert\mathcal{M}_i\rvert$ (# instances of each motif $\mathcal{M}_i$ in one class)	Motifs with more instances have stronger associations with the class than the less frequent motifs.

TABLE 1: Category of existing graph imbalance issues.

Outline

1. Introduction to Imbalanced Learning on Graphs (ILoGs)
2. Background
3. Overview of Taxonomies
4. Problems of ILoGs
5. Techniques of ILoGs
6. Future Directions
7. Conclusions

Overview of Taxonomies

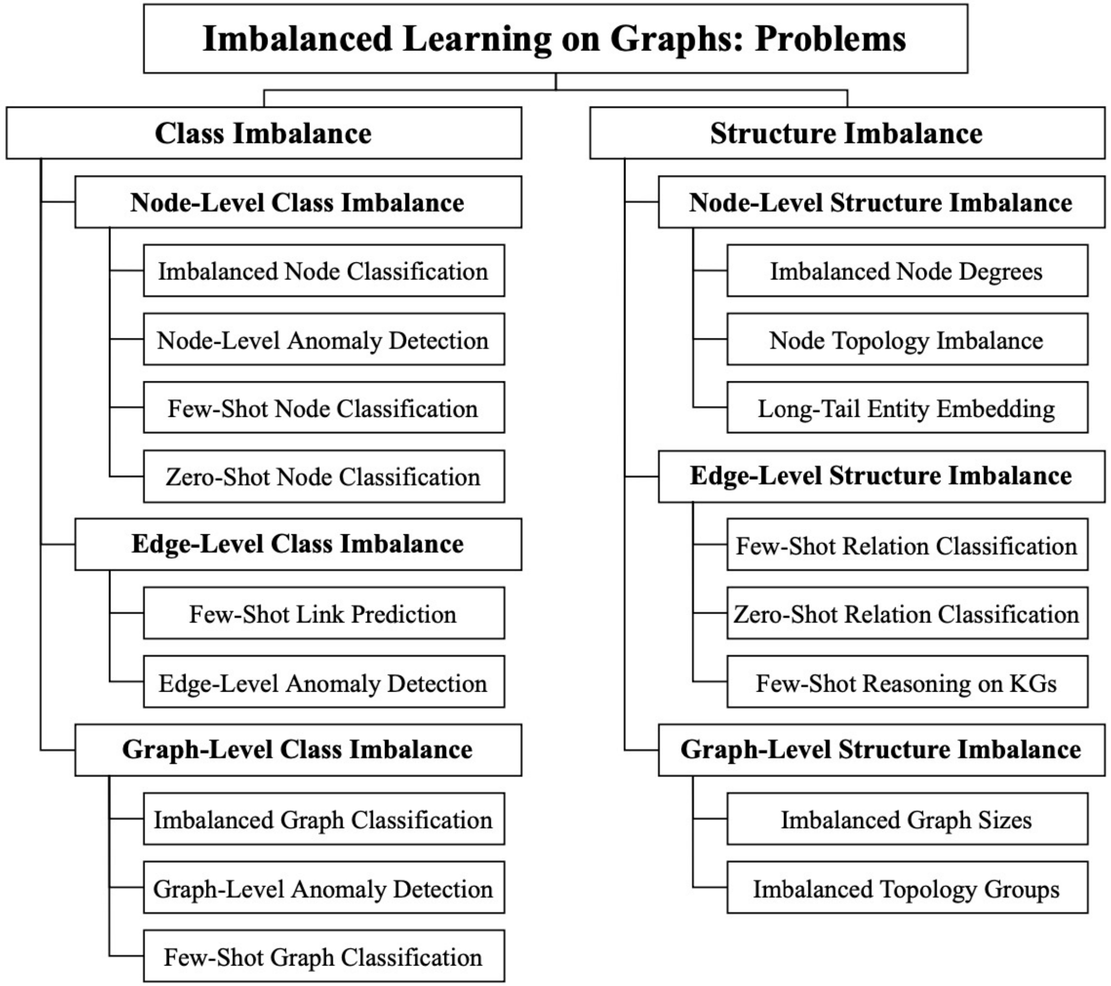


Fig. 3: Taxonomy of Problems.

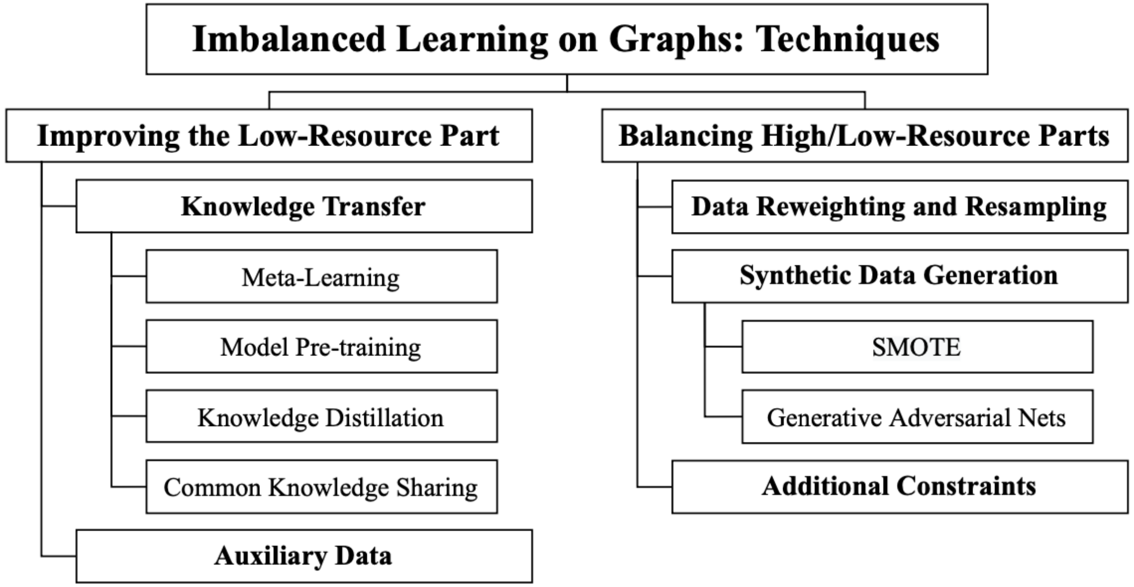


Fig. 4: Taxonomy of Techniques.

Outline

1. Introduction to Imbalanced Learning on Graphs (ILoGs)
2. Background
3. Overview of Taxonomies
4. Problems of ILoGs
5. Techniques of ILoGs
6. Future Directions
7. Conclusions

Problems of ILoGs

- **Categories**
  - Class Imbalance
  - Structure Imbalance

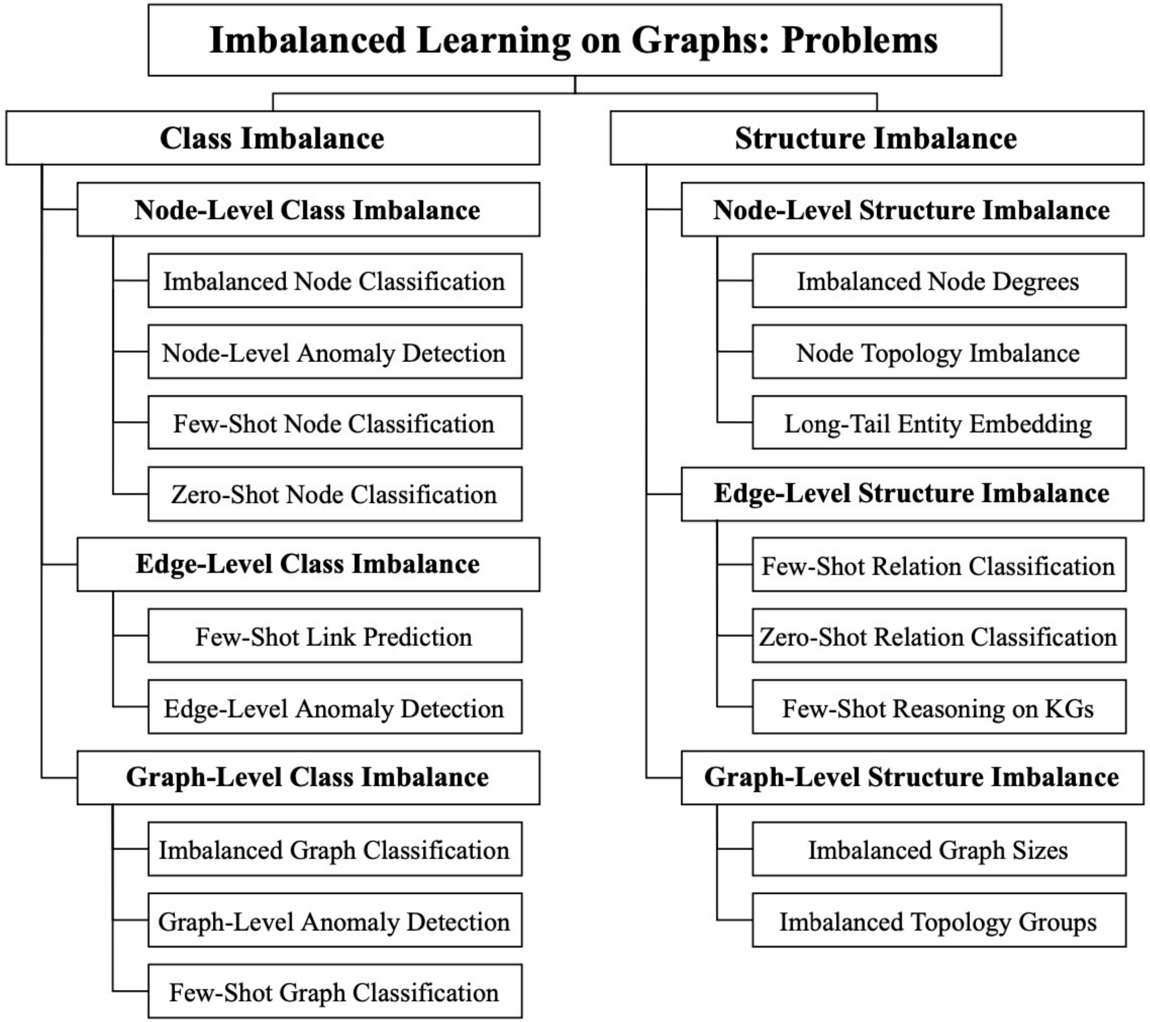


Fig. 3: Taxonomy of Problems.

Imbalanced Node Classification (1)

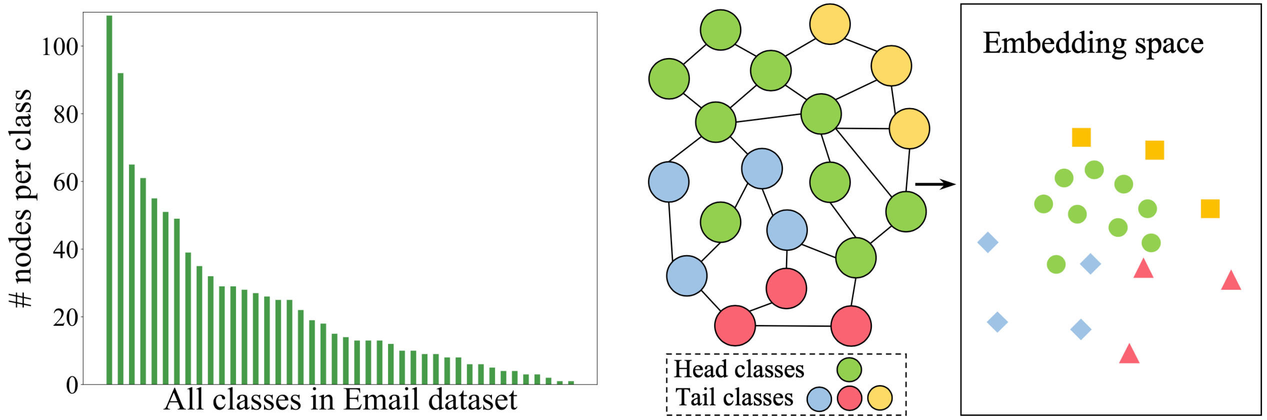
- Settings

  A set of (or two) *node classes* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{C}_i$

- Information Abundance

  $|\mathcal{C}_i|$ (# labeled nodes in each class $\mathcal{C}_i$)

- Explanations
  - Labeled nodes are unevenly distributed across classes.



(a) Long-tailed class distribution (b) Head/tail classes and their embeddings

Figure 1: Illustration of imbalanced classes.

Imbalanced Node Classification (2)

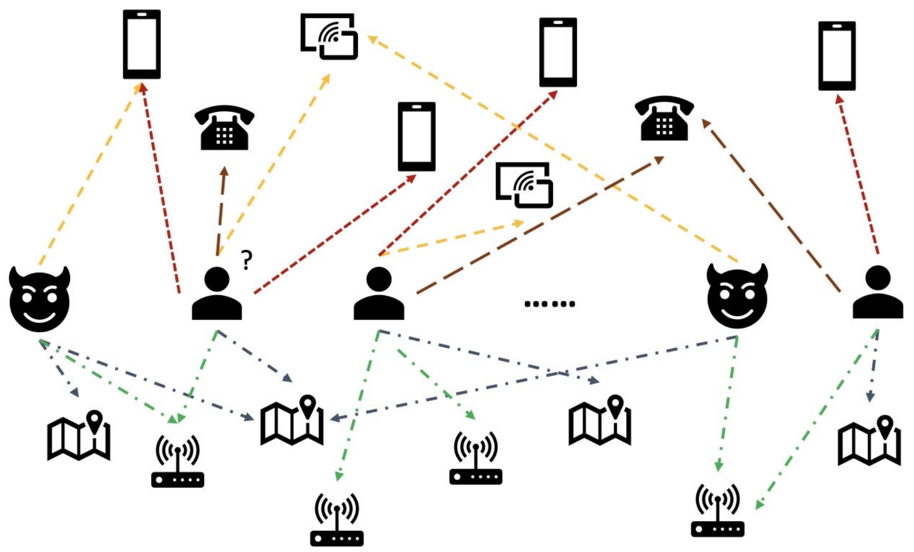
Techniques		Literature
Algo-level	Constraints Knowledge distillation	DRGCN [35], DPGNN [60], TAM [61] LTE4G [62]
Data-level	GAN SMOTE Mixup Resampling Reweighting	DRGCN [35], ImGAGN [37] GraphSMOTE [25] GraphMixup [64], GraphSANN [66], GraphENS [65] LTE4G [62], ALLIE [68] TAM [61]

TABLE 2: Summary of imbalanced node classification.

- Summary
  - **Challenge**
    - Achieving balanced information distribution across classes for uniform model training
  - Possible **Further Explorations**
    - Innovative techniques: e.g., diffusion models [a]

[a] Rombach R., et al. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR.

Node-Level Anomaly Detection



- Summary
  - A special case of Imbalanced Node Classification
  - Possible further exploration
    - Diffusion models [a]
    - Foundational models [b]
  - More references
    - a comprehensive survey [c]
    - Benchmarks: e.g., [d]
    - Leaderboards: e.g., [e]

Graph objects	Graph types	Base models	Literature
Node-level	Homogeneous	GA	[80]–[90]
		GE	[91]–[96]
		GNN	[76], [97]–[123]
	Heterogeneous	GA	[124]–[130]
		GE	[131], [132]
		GNN	[36], [74], [75], [122], [133]–[148]
	Dynamic	GA	[79], [149]–[156]
		GE	[157]–[161]
		GNN	[73], [162]–[166]
Edge-level	Homogeneous	GNN	[167], [168]
	Heterogeneous	GNN	[169]
	Dynamic	GA	[77], [170]
		GE	[78]
		GNN	[171], [172]
Graph-level	Homogeneous	GA	[173]–[178]
		GNN	[71], [72], [179], [180]
	Heterogeneous	GA	[181]
	Dynamic	GA	[182], [183]
		GNN	[184]

TABLE 3: Summary of anomaly detection on graphs.

[a] Rombach R., et al. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR.
[b] Bommasani R., et al. 2021. On the Opportunities and Risks of Foundation Models. arXiv.
[c] Ma X., et al. 2021. A Comprehensive Survey on Graph Anomaly Detection with Deep Learning. TKDE.
[d] Liu K., et al. 2022. BOND: Benchmarking Unsupervised Outlier Node Detection on Static Attributed Graphs. NeurIPS.
[e] https://dgraph.xinye.com/leaderboards/dgraphfin

Few-Shot Node Classification (1)

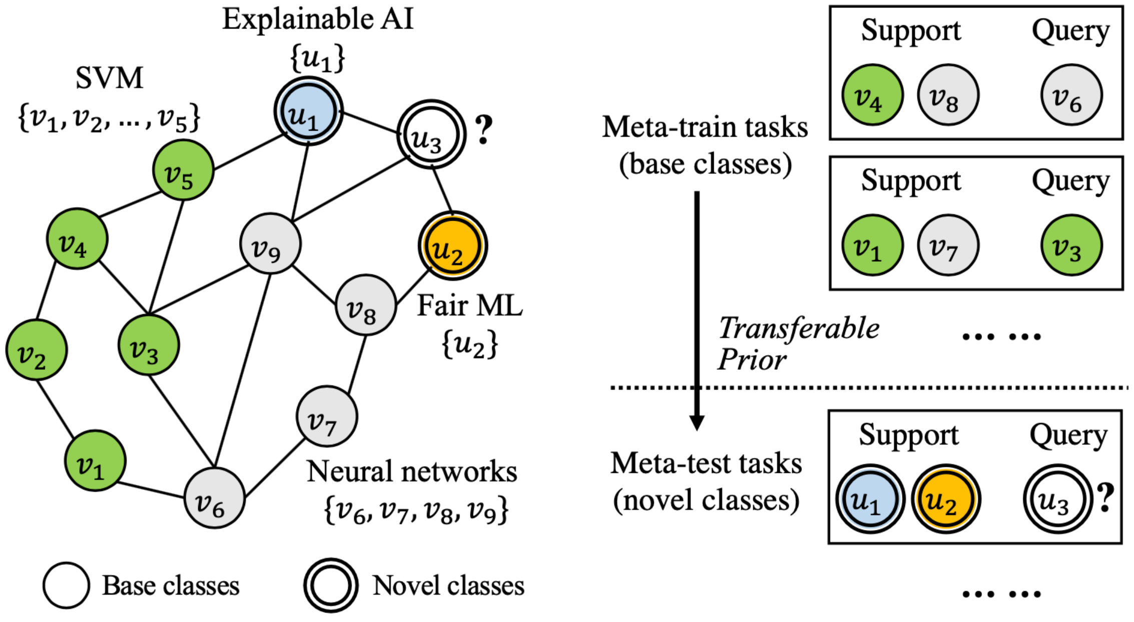
- Settings

  A set of base *node classes* $\mathcal{D}_b = \bigcup_{1 \leq i \leq K_1} \mathcal{C}_i$,
  and novel *node classes* $\mathcal{D}_n = \bigcup_{K_1 < i \leq K_2} \mathcal{C}_i$

- Information Abundance

  $|\mathcal{C}_i|$ (# labeled nodes in each class $\mathcal{C}_i$)

- Explanations
  - Base classes have abundant labeled nodes, while novel classes have few/no labeled nodes.



(a) Base and novel classes on graph

(b) Few-shot node classification

Few-Shot Node Classification (2)

Tasks	Meta-learning techniques			Other techniques		
	MAML	Prototypical network	Others	Label generation	Contrastive Learning	Prompting
Generic FSNC	[57], [187]–[189]	[26], [58], [190]–[193]	-	[194]	[195], [196]	[197]
Generalized FSNC	-	[198], [199]	[200]	-	-	-
Multi-label FSNC	-	[201]	-	-	-	-
FSNC with extremely weak supervision	-	-	-	[202]	-	-
FSNC on HINs	[30], [203]	-	-	-	-	-

TABLE 4: Summary of few-shot node classification.

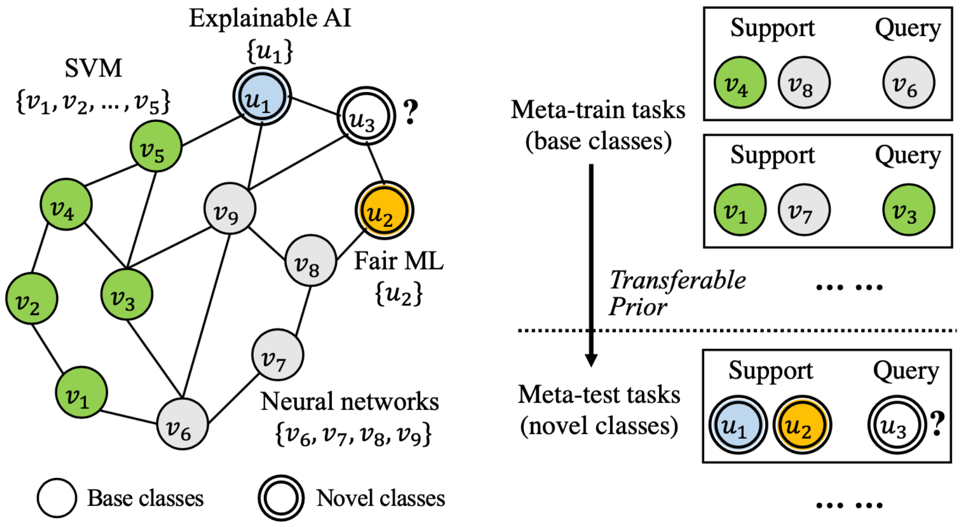
- Summary
  - **Challenge**
    - How to extract transferable knowledge from base classes to benefit novel classes
  - Possible **Further Exploration**
    - Specific settings remain largely underexplored
      - Generalized; multi-label; extremely weak supervision; FSNC on HINs
    - Innovative techniques
      - Prompt tuning [a]; generative models like diffusion [b]
  - **A Comprehensive Survey**
    - [c]

[a] Liu Z., et al. 2023. GraphPrompt: Unifying Pre-Training and Downstream Tasks for Graph Neural Networks. WWW.
[b] Rombach R., et al. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR.
[c] Zhang C., et al. 2022. Few-Shot Learning on Graphs. IJCAI.

Zero-Shot Node Classification

- **Characteristics**
  - Necessitate the absence of labeled data for novel classes during model training
  - A special case of Few-Shot Node Classification

- **Summary**
  - Still underexplored
    - Due to the absence of **descriptions** for elements like nodes, edges, or graphs
  - A formalization of ZSNC
    - [a]



(a) Base and novel classes on graph
(b) Few-shot node classification

[a] Wang Z., et al. 2021. Zero-shot Node Classification with Decomposed Graph Prototype Network. KDD.

Few-Shot Link Prediction

- Settings

  A set of base *graphs* $\mathcal{D}_b = \{G_i\}_{i=1}^{K_1}$ and novel *graphs* $\mathcal{D}_n = \{G_i\}_{i=K_1+1}^{K_2}$, where $G_i = \{\mathcal{V}_i, \mathcal{E}_i\}$

- Information Abundance

  $|\mathcal{E}_i|$ (# edges in each graph G_i)

- Explanations
  - Base graphs have abundant edges, while novel graphs have limited edges.

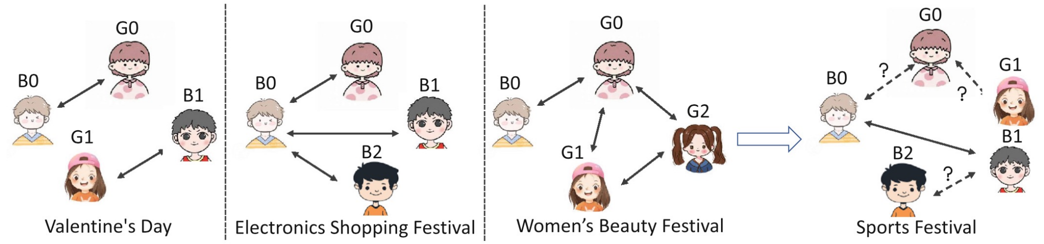


Fig. 1. An example of overlapping EBSNs generated by product share records on an e-commerce platform, where B and G denote boys and girls, respectively.

- Summary
  - Still underexploited
  - Other settings
    - Few-shot link prediction across different sections of a single graph

Edge-Level Anomaly Detection

- Summary
  - A special case of Imbalanced Edge Classification
  - Challenge
    - The highly imbalanced distribution of normal and abnormal edges
  - Possible further exploration
    - Investigation on HINs
  - More references
    - a comprehensive survey [a]

Graph objects	Graph types	Base models	Literature
Node-level	Homogeneous	GA GE GNN	[80]–[90] [91]–[96] [76], [97]–[123]
	Heterogeneous	GA GE GNN	[124]–[130] [131], [132] [36], [74], [75], [122], [133]–[148]
	Dynamic	GA GE GNN	[79], [149]–[156] [157]–[161] [73], [162]–[166]
Edge-level	Homogeneous	GNN	[167], [168]
	Heterogeneous	GNN	[169]
	Dynamic	GA GE GNN	[77], [170] [78] [171], [172]
Graph-level	Homogeneous	GA GNN	[173]–[178] [71], [72], [179], [180]
	Heterogeneous	GA	[181]
	Dynamic	GA GNN	[182], [183] [184]

TABLE 3: Summary of anomaly detection on graphs.

[a] Ma X., et al. 2021. A Comprehensive Survey on Graph Anomaly Detection with Deep Learning. TKDE.

Imbalanced Graph Classification

- Settings

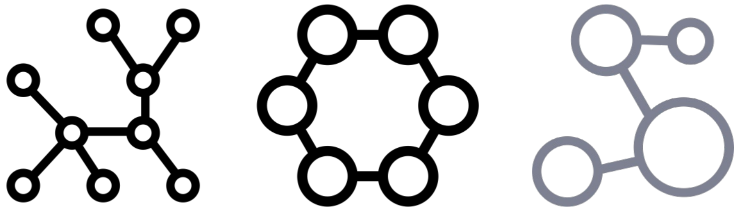
  A set of (or two) *graph classes* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{C}_i$

- Information Abundance

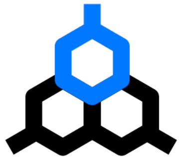
  $|\mathcal{C}_i|$ (# labeled graphs in each class $\mathcal{C}_i$)

- Explanations
  - Labeled graphs are unevenly distributed across classes.

- Summary
  - Scenarios
    - e.g., imbalanced chemical compound classification
  - Still underexploited



Class A



Class B

Graph-Level Anomaly Detection

- Summary
  - A special case of Imbalanced Graph Classification
  - Challenge
    - The highly imbalanced distribution of normal and abnormal graphs
  - Approaches
    - Determine node- or edge-level anomaly scores and aggregate to gauge graph-level anomalies
    - Graph-level embedding -> anomaly scores
  - Possible further exploration
    - Investigation on HINs
  - More references
    - a comprehensive survey [a]

Graph objects	Graph types	Base models	Literature
Node-level	Homogeneous	GA	[80]–[90]
		GE	[91]–[96]
		GNN	[76], [97]–[123]
	Heterogeneous	GA	[124]–[130]
		GE	[131], [132]
		GNN	[36], [74], [75], [122], [133]–[148]
	Dynamic	GA	[79], [149]–[156]
		GE	[157]–[161]
		GNN	[73], [162]–[166]
Edge-level	Homogeneous	GNN	[167], [168]
	Heterogeneous	GNN	[169]
	Dynamic	GA	[77], [170]
		GE	[78]
		GNN	[171], [172]
Graph-level	Homogeneous	GA	[173]–[178]
		GNN	[71], [72], [179], [180]
	Heterogeneous	GA	[181]
	Dynamic	GA	[182], [183]
		GNN	[184]

TABLE 3: Summary of anomaly detection on graphs.

[a] Ma X., et al. 2021. A Comprehensive Survey on Graph Anomaly Detection with Deep Learning. TKDE.

Few-Shot Graph Classification (1)

- Settings

  A set of base *graph classes* $\mathcal{D}_b = \bigcup_{1 \le i \le K_1} \mathcal{C}_i$,
  and novel *graph classes* $\mathcal{D}_n = \bigcup_{K_1 < i \le K_2} \mathcal{C}_i$

- Information Abundance

  $|\mathcal{C}_i|$ (# labeled graphs in each class $\mathcal{C}_i$)

- Explanations
  - Base classes have abundant labeled graphs, while novel classes have few labeled graphs.

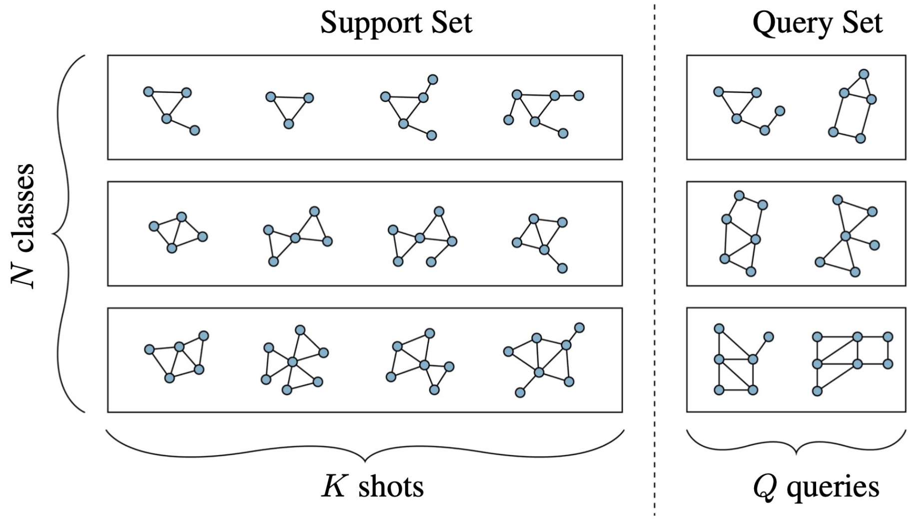


Figure 1: An N-way K-shot episode. In this example, there are $N = 3$ classes. Each class has $K = 4$ supports yielding a support set with size $N * K = 12$. The class information provided by the supports is exploited to classify the queries. We test the classification accuracy on all N classes. In this figure there are $Q = 2$ queries for each class, thus the query set has size $N * Q = 6$.

Few-Shot Graph Classification (2)

Tasks	Meta-learning techniques		Other techniques
	MAML	Prototypical network	
Generic FSGC	[219]	[220], [221]	Adaptive step controller [219], super-class graph [222], task correlations [220]
Cross-domain FSGC	-	-	Data augmentation [223]
Few-shot temporal graph classification	[224]	[224]	-
Few-shot molecular property prediction	[225]–[228]	-	Meta-task reweighting [226], implicit function theorem [229]

TABLE 5: Summary of few-shot graph classification.

- Summary
  - **Challenge**
    - Effectively transferring knowledge from base graph classes to novel graph classes to enhance the performance of the latter
  - Possible **further exploration**
    - E.g., cross-domain scenario FSGC; FSGC on temporal graphs
    - FSGC on HINs

Imbalanced Node Degrees (1)

- Tasks
  - Tail node embedding
  - Cold-start node embedding

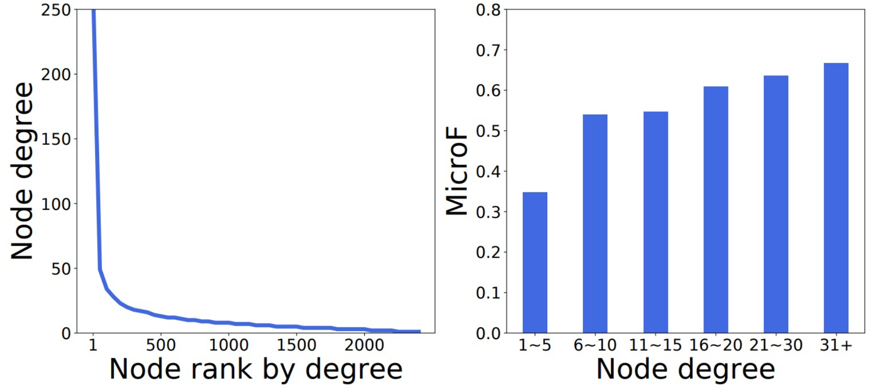
- Settings

  A set of *node groups* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{G}_i$, where
  $\mathcal{G}_i = \{v_j : d_j = i\}$ (d_j is the degree of node v_j)

- Information Abundance

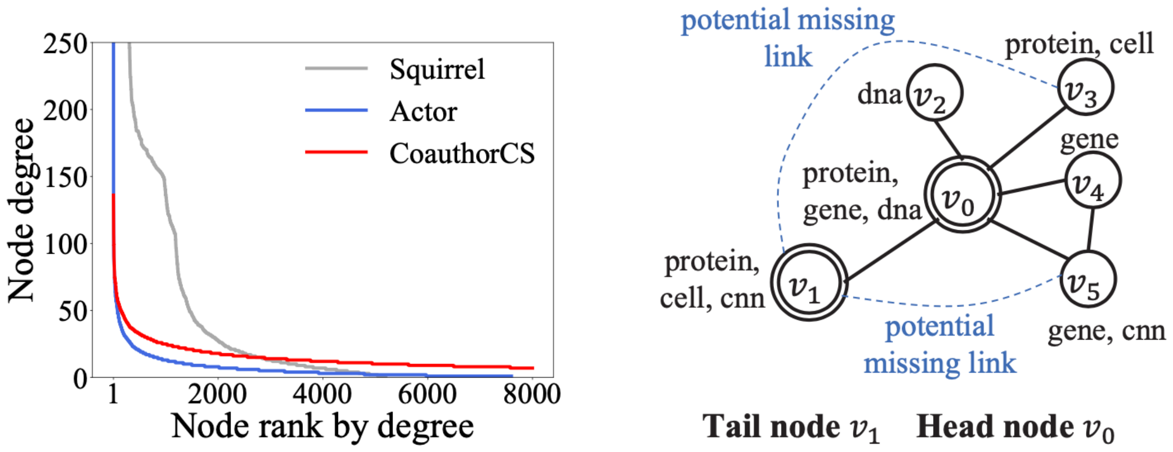
  d_j (the degree of each node v_j)

- Explanations
  - Head nodes have high degrees, while tail/cold-start nodes have few/no degrees.



(a) Degree distribution (b) Classification performance

Figure 1: Distribution of node degree and its relationship to the quality of embedding vector on the Wiki network.



(a) Long-tailed node distribution (b) Toy citation network

Figure 1: Illustration of tail nodes.

Imbalanced Node Degrees (2)

Tasks	Degree-aware modulation	Meta-learning	Knowledge distillation	Other techniques
Tail node embedding	[233], [234]	[18], [235], [236]	-	Neighborhood translation [27] Hybrid-order proximities [237] Reweighting [238]–[240]
Cold-start node embedding	-	-	[56]	-
Node topology imbalance	-	-	-	Reweighting [241], [242] Graph geometric embedding [243]
Long-tail entity embedding on KGs	[244]	[245]–[248]	-	Open knowledge enrichment [249] Synthetic data generation [250]

TABLE 6: Summary of node-level structure imbalance.

- Summary
  - **Challenge**
    - Efficient knowledge transfer from head nodes to tail or cold-start nodes
  - Possible **further exploration**
    - Cold-start node embedding is still underexploited
    - The possible usage of knowledge distillation for tail node embedding

Node Topology Imbalance

- Settings

  A set of *node classes* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{C}_i$

- Information Abundance
  - The consistency between true class boundaries and influence boundaries of labeled nodes

- Explanations
  - Classes with more consistent boundaries tend to propagate label information more effectively.

- Summary
  - Still an underexploited problem

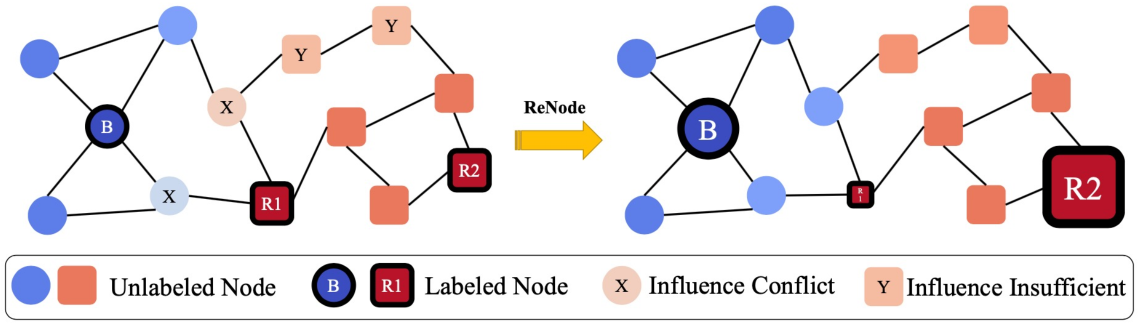


Figure 1: Schematic diagram of the topology-imbalance issue in node representation learning. The color and the hue denote the type and the intensity of each node's received influence from the labeled nodes, respectively. The left shows that nodes close to the boundary have the risk of information conflict and nodes far away from labeled nodes have the risk of information insufficient. The right shows that our method can decrease the training weights of labeled nodes (R1) close to the class boundary and increase the weights of labeled nodes (B and R2) close to the class centers, thus relieving the topology-imbalance issue.

Imbalanced Graph Sizes

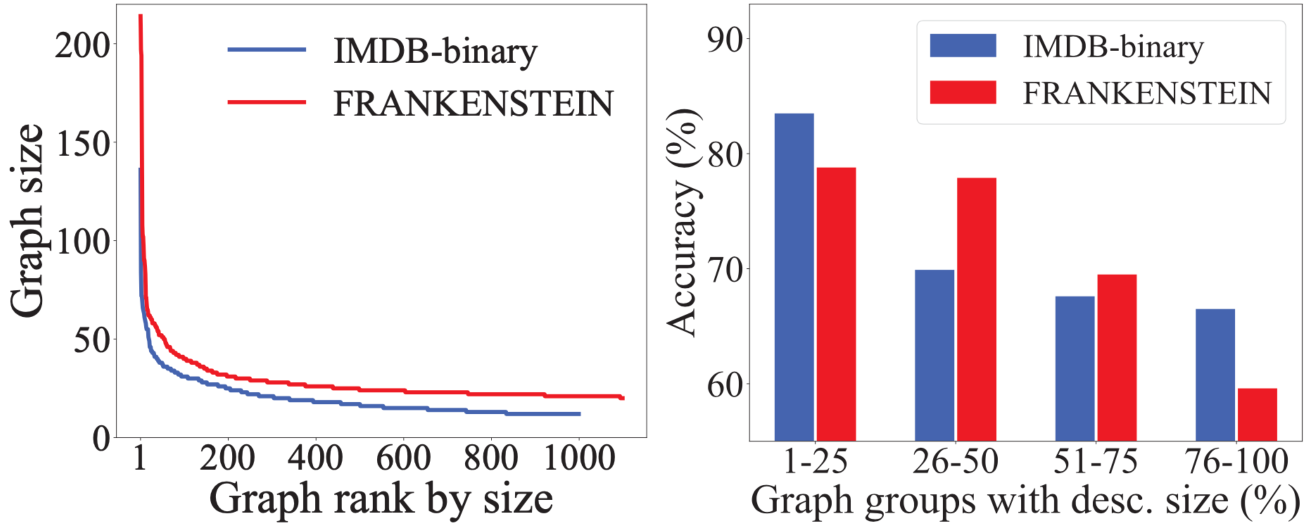
- Settings

A set of *graph groups* $\mathcal{D} = \bigcup_{1\le i\le K} \mathcal{G}_i$, where $\mathcal{G}_i = \{G_j : |\mathcal{V}_j| = i\}$ ($|\mathcal{V}_j|$ is the size of graph G_j)

- Information Abundance

$|\mathcal{V}_j|$ (the size of each graph G_j)

- Explanations
    - Head graph have large sizes, while tail graphs have small sizes.
- Summary
    - Still underexploited



(a) Long-tailed graph distribution (b) Graph classification performance

Figure 1: Illustration of long-tailed distribution.

Imbalanced Topology Groups

- Settings

A set of *topology motifs* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{M}_i$

- Information Abundance

$|\mathcal{M}_i|$ (# instances of each motif $\mathcal{M}_i$ in one class)

- Explanations
  - Motifs with more instances have stronger associations with the class than the less frequent motifs.
- Summary
  - Still underexploited

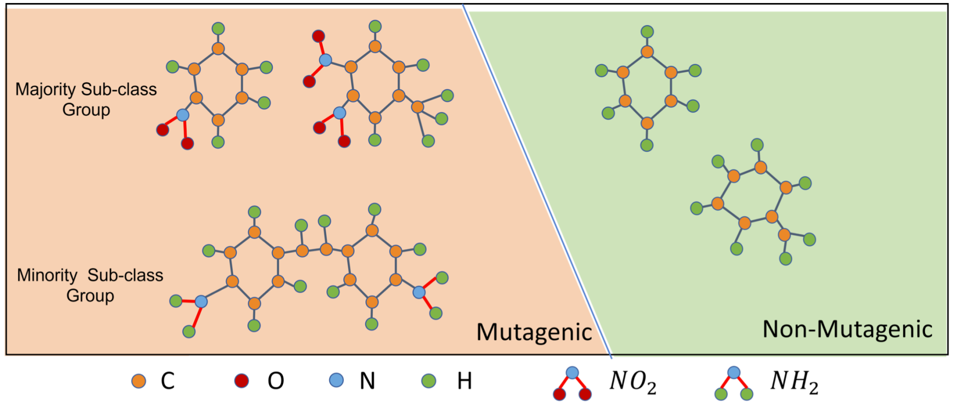


Figure 1: Example of sub-class topology level graph imbalance on dataset Mutag [12]. Molecular graphs of the Mutagenic class have two topology groups, one with motif NO_2 and another one with motif NH_2 [13]. The NO_2 group is much larger in quantity compared to the NH_2 group.

Existing Graph Imbalance Issues

Imbalance Types	Imbalance Tasks	Settings	Information Abundance s	Explanations
Node-Level Class Imbalance	Imbalanced node classification Node-level anomaly detection	A set of (or two) *node classes* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{C}_i$	$\lvert\mathcal{C}_i\rvert$ (# labeled nodes in each class $\mathcal{C}_i$)	Labeled nodes are unevenly distributed across classes.
	Few-shot node classification Zero-shot node classification	A set of base *node classes* $\mathcal{D}_b = \bigcup_{1 \le i \le K_1} \mathcal{C}_i$, and novel *node classes* $\mathcal{D}_n = \bigcup_{K_1 < i \le K_2} \mathcal{C}_i$	$\lvert\mathcal{C}_i\rvert$ (# labeled nodes in each class $\mathcal{C}_i$)	Base classes have abundant labeled nodes, while novel classes have few/no labeled nodes.
Edge-Level Class Imbalance	Few-shot link prediction	A set of base *graphs* $\mathcal{D}_b = \{G_i\}_{i=1}^{K_1}$ and novel *graphs* $\mathcal{D}_n = \{G_i\}_{i=K_1+1}^{K_2}$, where $G_i = \{\mathcal{V}_i, \mathcal{E}_i\}$	$\lvert\mathcal{E}_i\rvert$ (# edges in each graph G_i)	Base graphs have abundant edges, while novel graphs have limited edges.
	Edge-level anomaly detection	Two *edge classes* $\mathcal{D} = \mathcal{C}_1 \cup \mathcal{C}_2$	$\lvert\mathcal{C}_i\rvert$ (# labeled edges in each class $\mathcal{C}_i$)	Labeled edges are unevenly distributed across classes.
Graph-Level Class Imbalance	Imbalanced graph classification Graph-level anomaly detection	A set of (or two) *graph classes* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{C}_i$	$\lvert\mathcal{C}_i\rvert$ (# labeled graphs in each class $\mathcal{C}_i$)	Labeled graphs are unevenly distributed across classes.
	Few-shot graph classification	A set of base *graph classes* $\mathcal{D}_b = \bigcup_{1 \le i \le K_1} \mathcal{C}_i$, and novel *node classes* $\mathcal{D}_n = \bigcup_{K_1 < i \le K_2} \mathcal{C}_i$	$\lvert\mathcal{C}_i\rvert$ (# labeled graphs in each class $\mathcal{C}_i$)	Base classes have abundant labeled graphs, while novel classes have few labeled graphs.
Node-Level Structure Imbalance	Imbalanced node degrees	A set of *node groups* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{G}_i$, where $\mathcal{G}_i = \{v_j : d_j = i\}$ (d_j is the degree of node v_j)	d_j (the degree of each node v_j)	Head nodes have high degrees, while tail/cold-start nodes have few/no degrees.
	Node topology imbalance	A set of *node classes* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{C}_i$	The consistency between true class boundaries and influence boundaries of labeled nodes	Classes with more consistent boundaries tend to propagate label information more effectively.
	Long-tail entity embedding	A set of *entity groups* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{G}_i$, where $\mathcal{G}_i = \{e_j : d_j = i\}$ (d_j is the # triplets of entity e_j)	d_j (# triplets of each entity e_j)	Head entities have more triplets, while tail/cold-start entities have few/no triplets.
Edge-Level Structure Imbalance	Few-shot relation classification Zero-shot relation classification Few-shot reasoning on KGs	A set of base *relations* $\mathcal{D}_b = \bigcup_{1 \le i \le K_1} \mathcal{R}_i$, and novel *relations* $\mathcal{D}_n = \bigcup_{K_1 < i \le K_2} \mathcal{R}_i$	$\lvert\mathcal{R}_i\rvert$ (# labeled triplets of each relation $\mathcal{R}_i$)	Base relations have abundant labeled triplets, while novel relations have few/no labeled triplets.
Graph-Level Structure Imbalance	Imbalanced graph sizes	A set of *graph groups* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{G}_i$, where $\mathcal{G}_i = \{G_j : \lvert\mathcal{V}_j\rvert = i\}$ ($\lvert\mathcal{V}_j\rvert$ is the size of graph G_j)	$\lvert\mathcal{V}_j\rvert$ (the size of each graph G_j)	Head graph have large sizes, while tail graphs have small sizes.
	Imbalanced topology groups	A set of *topology motifs* $\mathcal{D} = \bigcup_{1 \le i \le K} \mathcal{M}_i$	$\lvert\mathcal{M}_i\rvert$ (# instances of each motif $\mathcal{M}_i$ in one class)	Motifs with more instances have stronger associations with the class than the less frequent motifs.

TABLE 1: Category of existing graph imbalance issues.

Outline

1. Introduction to Imbalanced Learning on Graphs (ILoGs)
2. Background
3. Overview of Taxonomies
4. Problems of ILoGs
5. Techniques of ILoGs
6. Future Directions
7. Conclusions

Techniques of ILoGs

- Categories
  - Improving the Low-Resource Part
    - *Examples*
      - few-shot node classification
      - tail/cold-start node representation learning
  - Balancing High/Low-Resource Parts
    - *Examples*
      - imbalanced node/edge/graph classification

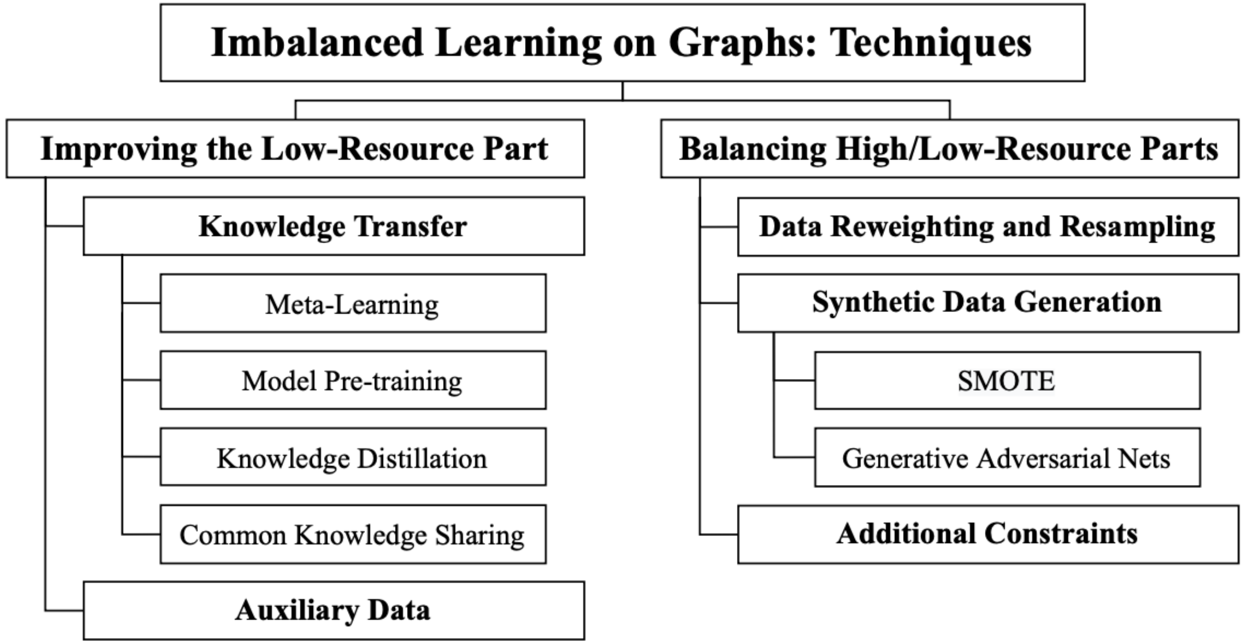
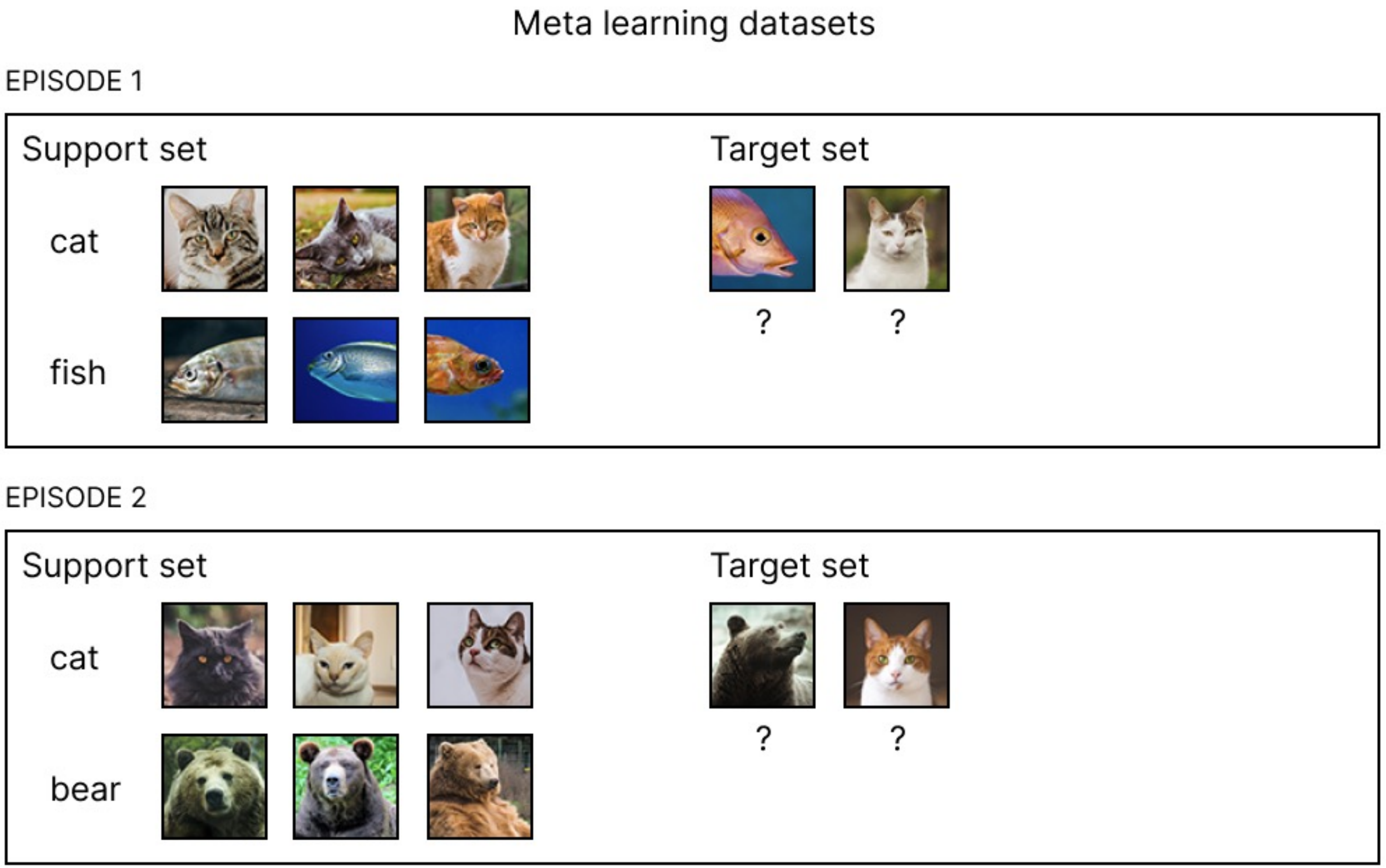


Fig. 4: Taxonomy of Techniques.

Improving the Low-Resource Part: Knowledge Transfer – Meta-Learning

- Meta-Learning: Learning to learn



Improving the Low-Resource Part: Knowledge Transfer – Meta-Learning

- MAML [a]
  - Model-Agnostic Meta-Learning

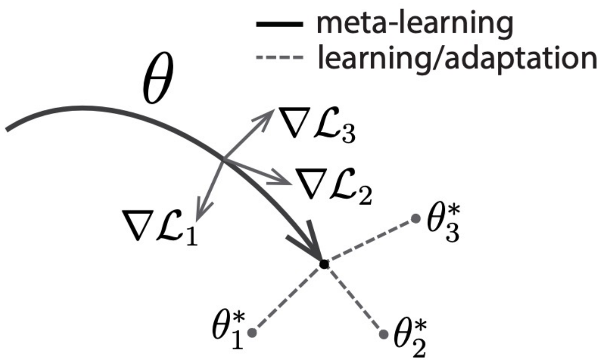
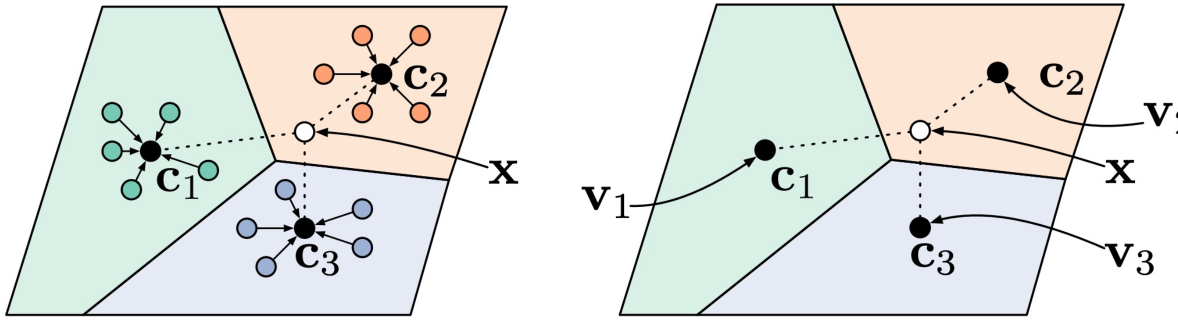


Figure 1. Diagram of our model-agnostic meta-learning algorithm (MAML), which optimizes for a representation θ that can quickly adapt to new tasks.

- Prototypical network [b]



(a) Few-shot (b) Zero-shot

Figure 1: Prototypical networks in the few-shot and zero-shot scenarios. **Left**: Few-shot prototypes $\mathbf{c}_k$ are computed as the mean of embedded support examples for each class. **Right**: Zero-shot prototypes $\mathbf{c}_k$ are produced by embedding class meta-data $\mathbf{v}_k$. In either case, embedded query points are classified via a softmax over distances to class prototypes: $p_\phi(y = k|\mathbf{x}) \propto \exp(-d(f_\phi(\mathbf{x}), \mathbf{c}_k))$.

[a] Finn C., et al. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. ICML 2017.
[b] Snell J., et al. Prototypical Networks for Few-shot Learning. NeurIPS 2017.

Improving the Low-Resource Part: Knowledge Transfer – Model Pre-training

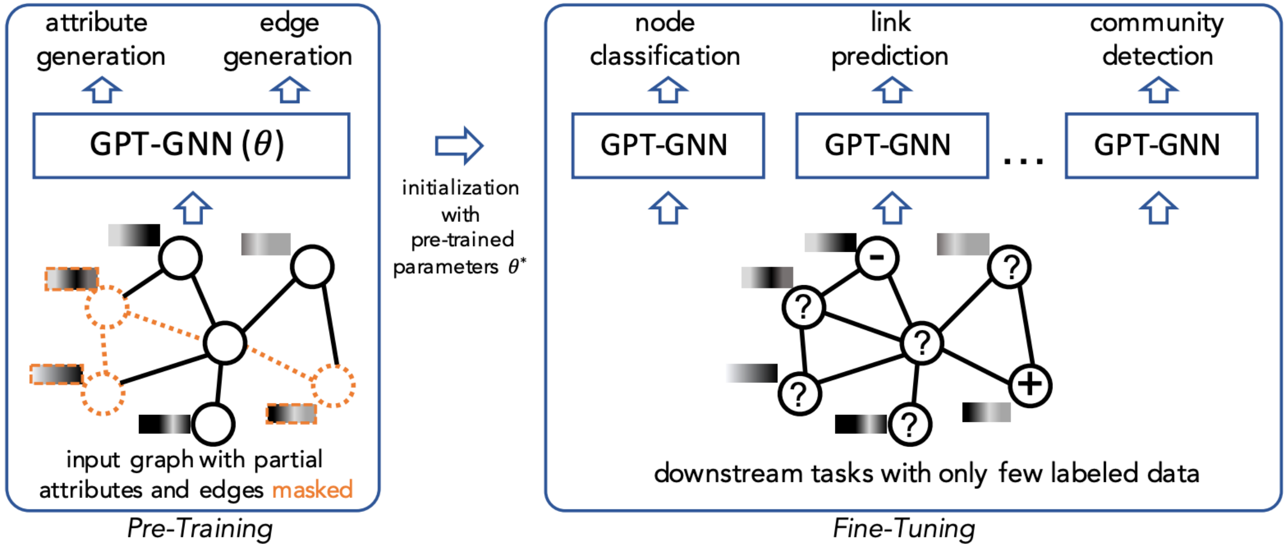
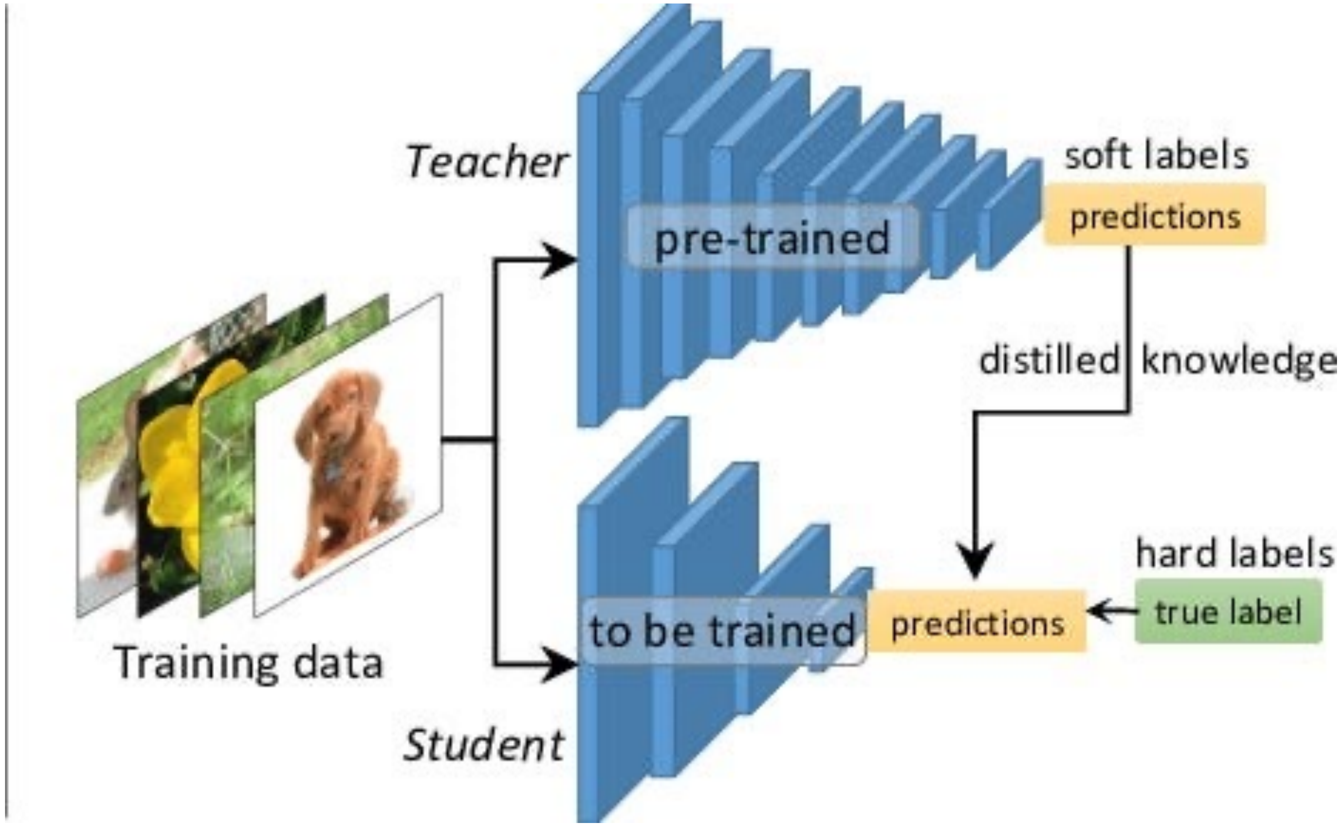


Figure 1: The pre-training and fine-tuning flow of GPT-GNN: First, a GNN is pre-trained with the self-supervised learning task—attribute and structure generations. Second, the pre-trained model and its parameters are then used to initialize models for downstream tasks on the input graph or graphs of the same domain.

[a] Hu Z., et al. GPT-GNN: Generative Pre-Training of Graph Neural Networks. KDD 2020.

Improving the Low-Resource Part: Knowledge Transfer – Knowledge Distillation

- Knowledge: Teacher Model → Student Model



[a] https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764.

Improving the Low-Resource Part: Knowledge Transfer – Common Knowledge Sharing

- Knowledge: High-Resource part → Low Resource Part

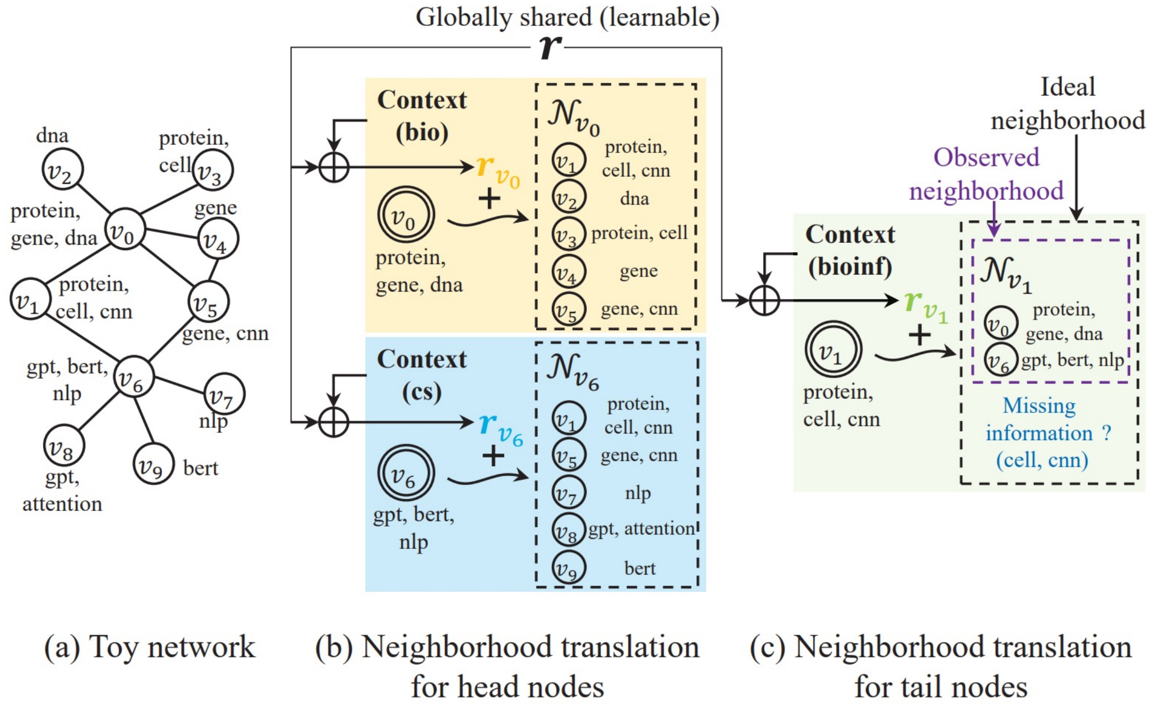


Figure 2: Illustration of neighborhood translation.

Key points:
- Identity (or consistency) between high- and low-resource parts
- High-resource → sufficient → learn knowledge
- Low-resource → insufficient → incorporate knowledge → fulfill this identity

[a] Liu Z., et al. Tail-GNN: Tail-Node Graph Neural Networks. KDD 2021.

Improving the Low-Resource Part: Auxiliary Data

- Auxiliary data: supplemental information
  - Text data

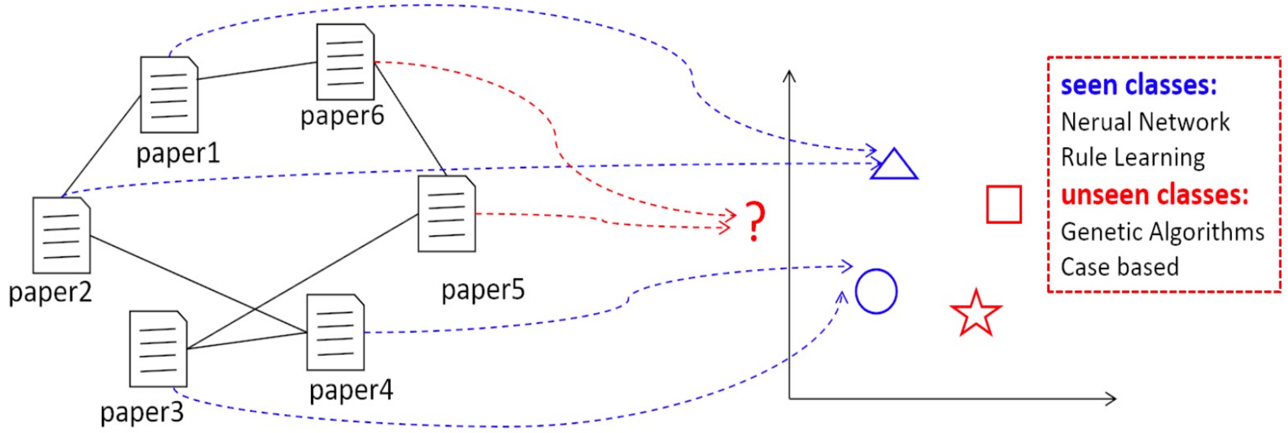


Figure 1: An example of zero-shot node classification.

[a] Yue Q., et al. Dual Bidirectional Graph Convolutional Networks for Zero-shot Node Classification. KDD 2022.

Balancing the High- and Low-Resource Parts: Data Reweighting and Resampling

- Data reweighting and resampling
  - Balancing the contribution of different parts

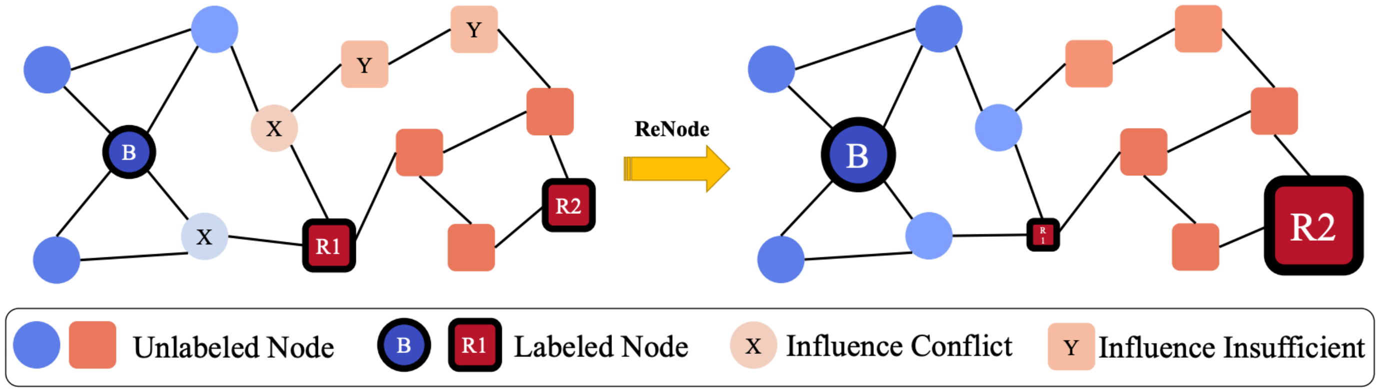
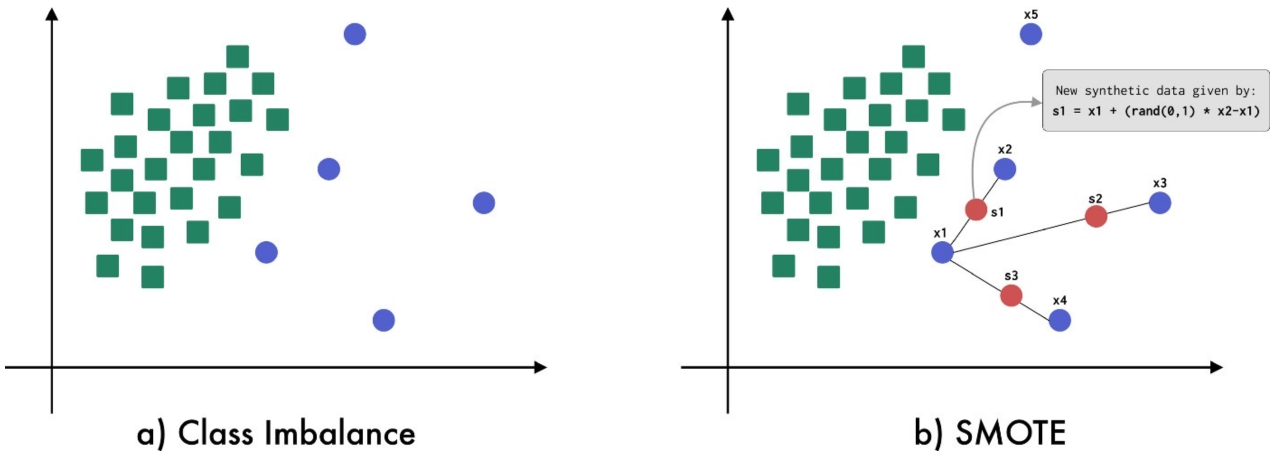


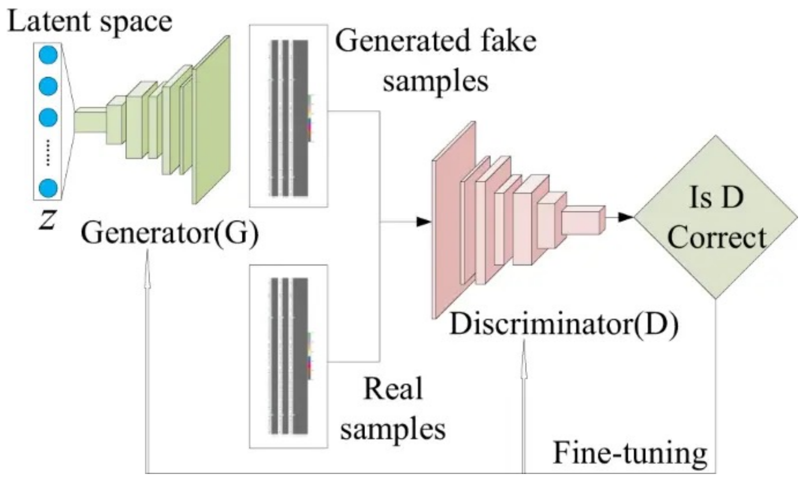
Figure 1: Schematic diagram of the topology-imbalance issue in node representation learning. The color and the hue denote the type and the intensity of each node's received influence from the labeled nodes, respectively. The left shows that nodes close to the boundary have the risk of information conflict and nodes far away from labeled nodes have the risk of information insufficient. The right shows that our method can decrease the training weights of labeled nodes (R1) close to the class boundary and increase the weights of labeled nodes (B and R2) close to the class centers, thus relieving the topology-imbalance issue.

[a] Chen D., et al. Topology-Imbalance Learning for Semi-Supervised Node Classification. NeurIPS 2021.

Balancing the High- and Low-Resource Parts: Synthetic Data Generation

- SMOTE [a] (synthetic minority over-sampling technique)
  - Mixup [b]
- GAN [c] (generative adversarial nets)





[a] Chawla N. V., et al. SMOTE: Synthetic Minority Over-sampling Technique. JAIR 2002.
[b] Zhang H., et al. mixup: Beyond empirical risk minimization. ICLR 2018.
[c] Goodfellow I., et al. Generative Adversarial Networks. NeurIPS 2014.

Literature Categorization

	Techniques			Literature
Improving the low-resource part	Knowledge transfer	Meta-learning	Optimization-based	[18], [31], [57], [75], [107], [187]–[189], [192], [202], [203], [208], [209], [219], [224]–[228], [236], [245], [254], [255], [257], [258], [272], [278]–[280], [282], [287]
			Metric-based	[26], [33], [58], [190], [191], [193], [198], [199], [201], [220], [221], [223], [253], [256], [259]–[269], [273], [276]
		Model pre-training		GNN parameters transfer [194], contrastive learning [195], [196], prompting [197]
		Knowledge distillation		GNNs to MLPs [56], KG models to MLPs [277], Random knowledge distillation [72]
		Common knowledge sharing	Data sharing	Super-classes [222]
			Model sharing	[27], [34], [233], [234]
	Auxiliary data			Alignment data [244], auxiliary descriptions [215], [216], [276]
Balancing high/low-resource parts	Reweighting and resampling			Reweighting [60], [74], [119], [124], [146], [178], [239], [241], [243], [288], resampling [36], [68], [82], [130], [133], [135], [142]–[144], [144], [147], [164], [171], [240], [242]
	Synthetic data generation	SMOTE		SMOTE [25], Mixup [64]–[66]
		GAN		[37], [101], [109], [163], [179]
		Other methods		Predictive data generation [73], [134], label generation [76], [166], [169]
	Additional constraints			Condition relax constraints [138], [212]–[214], imbalance constraints [61], [111], class separation constraints [35], [71], [77]–[100], [102]–[106], [108], [110], [112]–[118], [120]–[123], [125]–[129], [131], [132], [136], [137], [139]–[143], [145]–[162], [165], [167], [168], [170], [172]–[178], [180]–[184], [289]

TABLE 8: Literature categorization of imbalanced learning on graphs w.r.t. the taxonomy of techniques.

Appropriate Techniques Selection

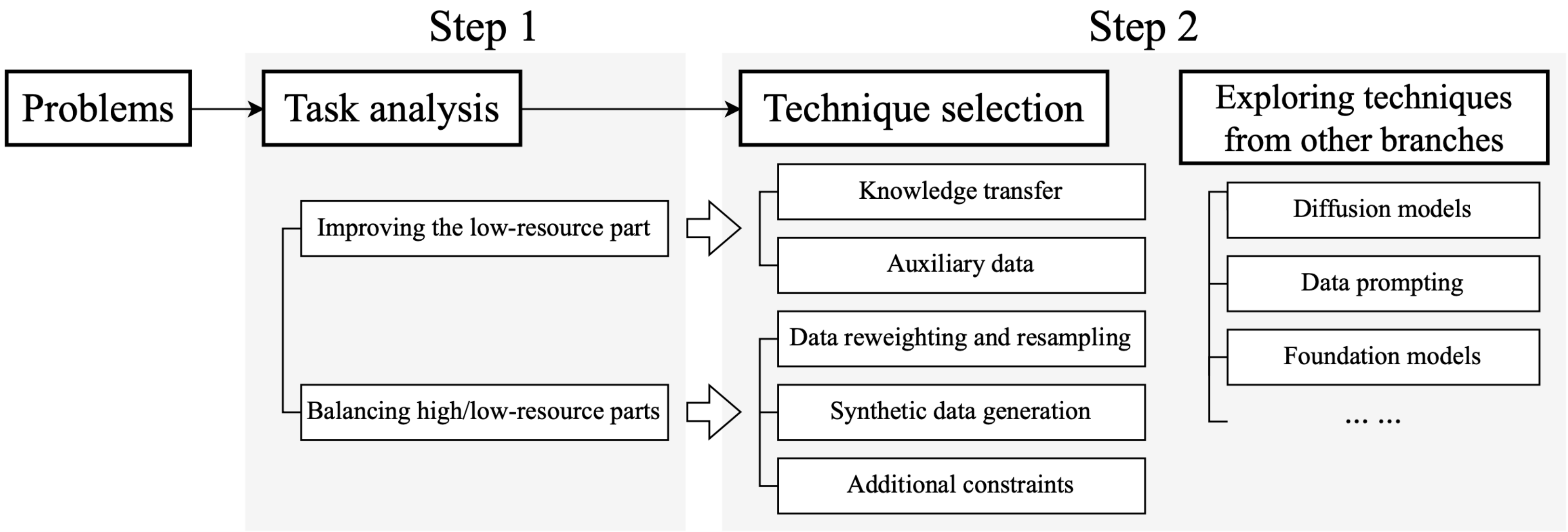


Fig. 5: Procedure of techniques selection.

Outline

1. Introduction to Imbalanced Learning on Graphs (ILoGs)
2. Background
3. Overview of Taxonomies
4. Problems of ILoGs
5. Techniques of ILoGs
6. Future Directions
7. Conclusions

Future Directions

- Future Directions of **Problems**
  - Class Imbalance
    - Existing attention: node-level imbalance
    - Edge-Level imbalance
    - Generic imbalanced graph classification
    - Zero-shot graph classification
      - May require text information
  - Structure Imbalance
    - Existing attention: node-level structure imbalance
    - Node-level: Generalized node degree [a]
    - Graph-level: Imbalanced graph-sizes
- Future Directions of **Techniques**
  - Cross-branch technique exploration
  - Novel technique exploration
    - Diffusion [b], foundation models [c]

[a] Liu Z., et al. 2023. On Generalized Degree Fairness in Graph Neural Networks. AAAI.
[b] Rombach R., et al. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR.
[c] Bommasani R., et al. 2021. On the Opportunities and Risks of Foundation Models. arXiv.

Outline

1. Introduction to Imbalanced Learning on Graphs (ILoGs)
2. Background
3. Overview of Taxonomies
4. Problems of ILoGs
5. Techniques of ILoGs
6. Future Directions
7. Conclusions

Conclusions

- **Task** of this talk
  - A comprehensive review of the literature on ILoGs

- Two **comprehensive taxonomies** of ILoGs
  - Problems
    - Class imbalance
      - Node, edge, graph
    - Structure imbalance
      - Node, edge, graph
  - Techniques
    - The type of imbalance
    - The corresponding strategies to rectify these imbalance

- **Future directions**
  - Problems
  - Techniques

Survey Paper Link

GitHub Repository Link

Thanks!

Zemin Liu
Senior Research Fellow
National University of Singapore
zeminliu@nus.edu.sg