When do Graph Neural Networks work and when not?

The principle for building Graph Foundation Model

Haitao Mao
Department of Computer Science and Engineering
Michigan State University
haitaoma@msu.edu

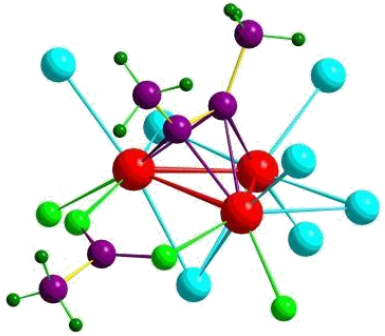
December, 2023

MICHIGAN STATE
UNIVERSITY

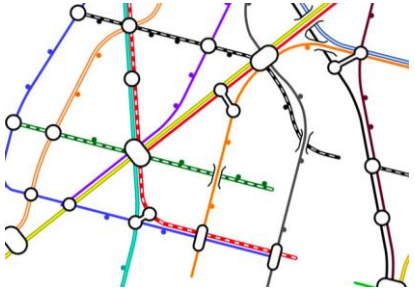
Graph can represent arbitrary data

Going Beyond regular data!

Social Network

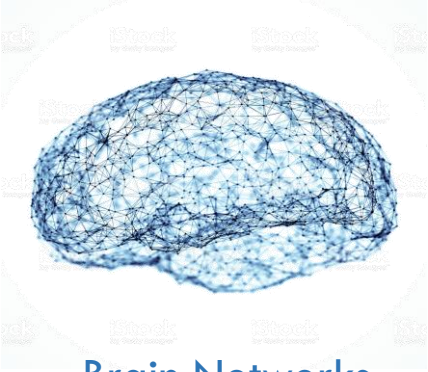


Molecules

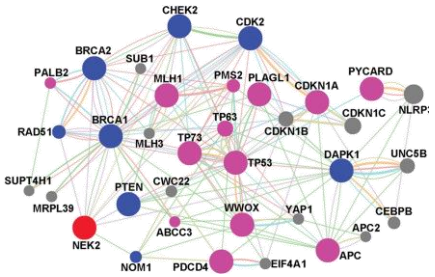


Transportation Networks

Web

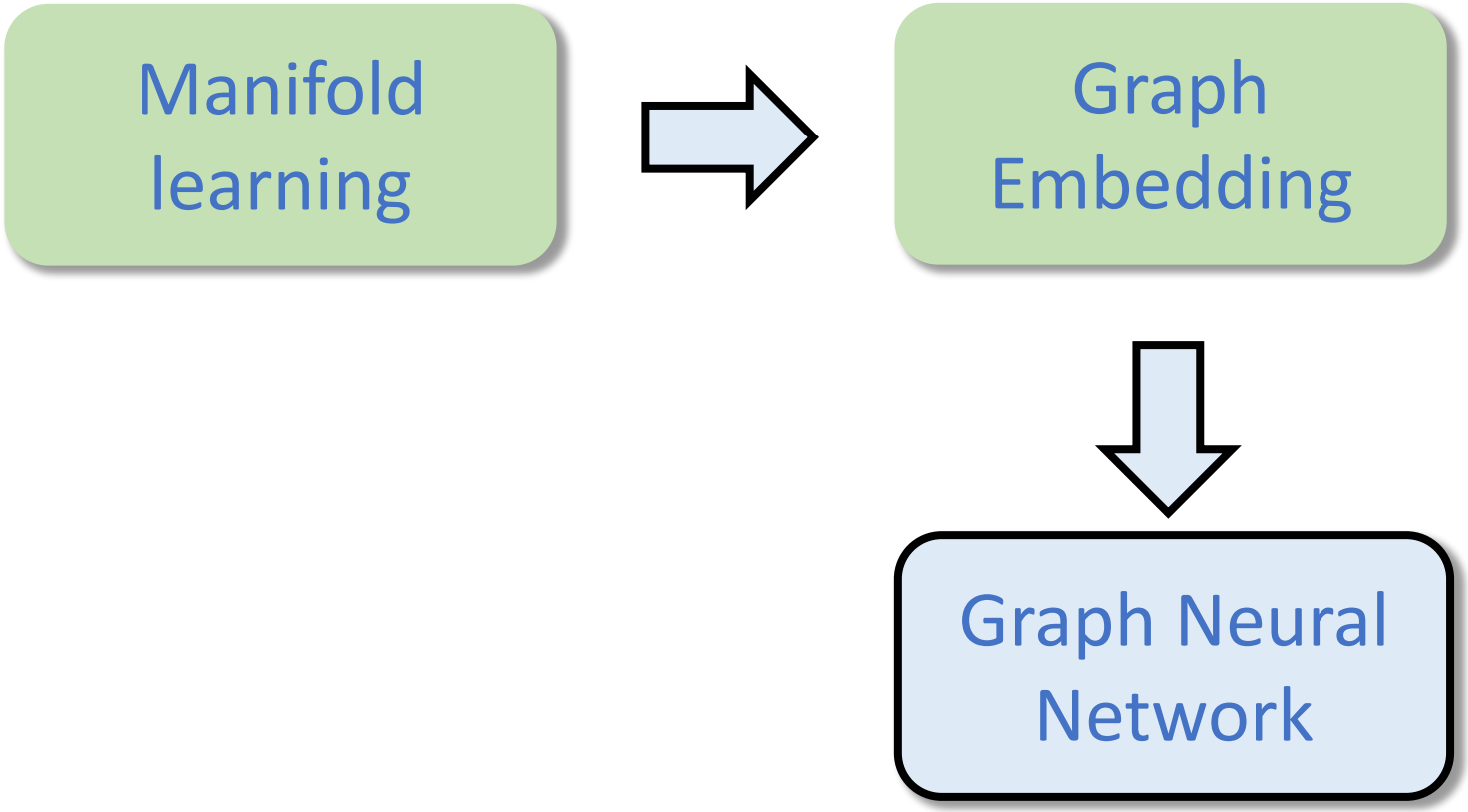


Brain Networks

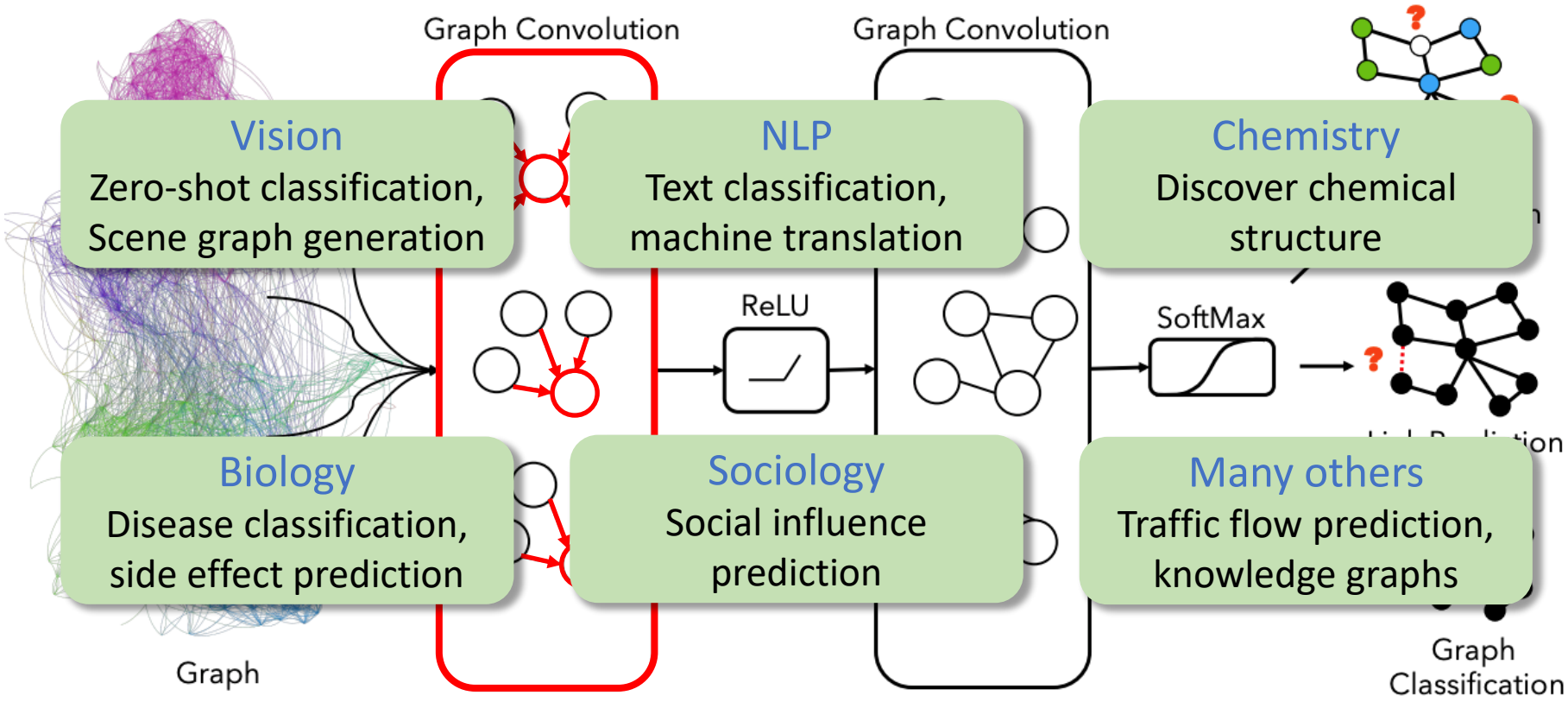


Gene Interactions

The stage for Graph Learning



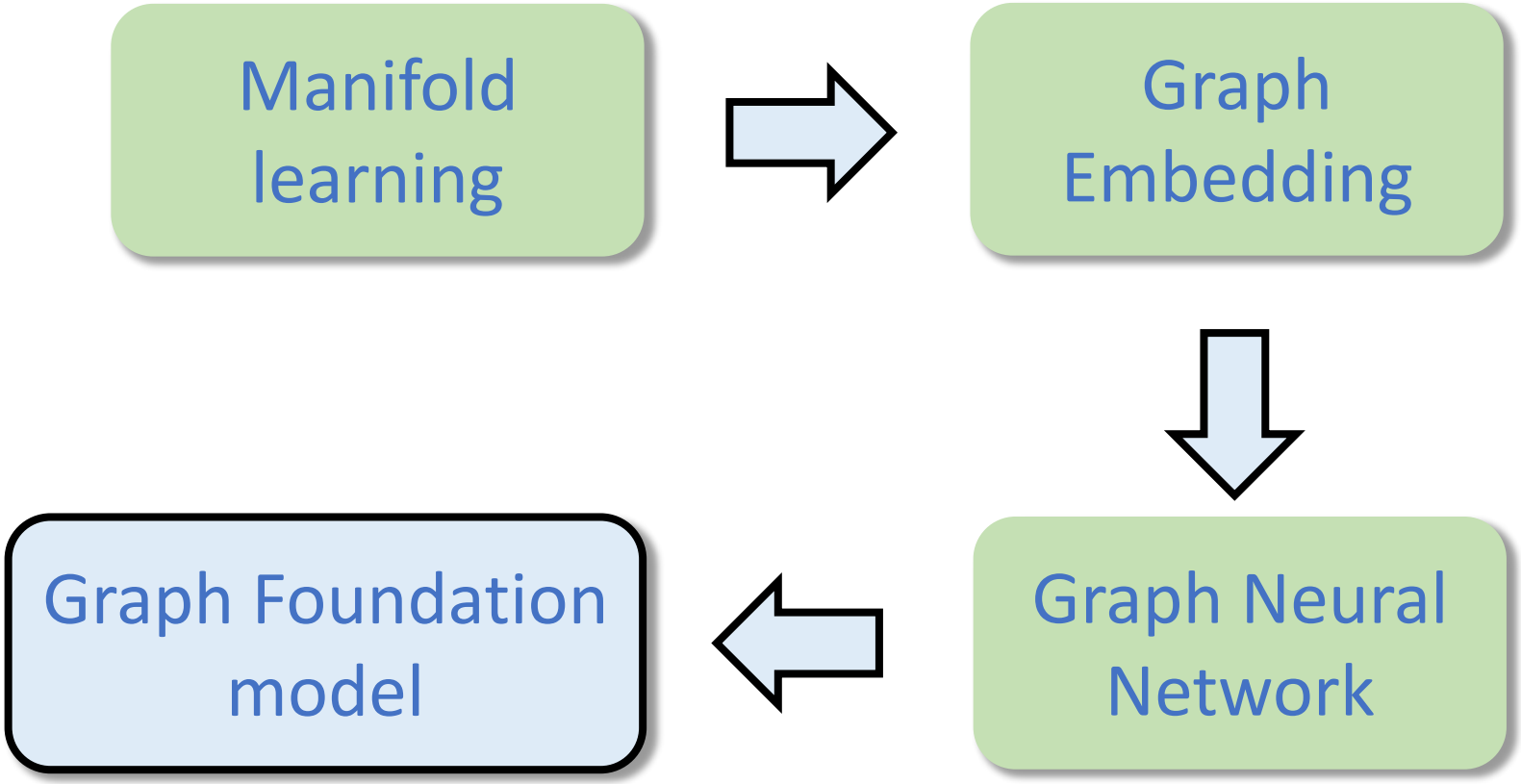
Graph Neural Networks (GNNs)



GNN achieve success in various applications

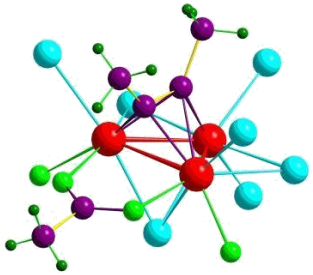
Image credit: Swapnil Gandhi

The stage for Graph Learning



The potential future Graph Foundation Model

Social Network



Molecules

Transportation

Graph Foundation Model (GFM)

Node classification

Graph classification

Link Prediction

Work for different datasets and different tasks

Obstacles for building Graph Foundation Model

Work for different datasets and different tasks

Obstacles for building Graph Foundation Model

Work for different datasets and different tasks

Table 1: Results on Cora, Citeseer, and Pubmed(%) under the existing evaluation setting. Highlighted are the results ranked **first**, **second**, and **third**.

	Models	Cora		Citeseer		Pubmed	
		MRR	AUC	MRR	AUC	MRR	AUC
Heuristic	CN	20.99	70.85	28.34	67.49	14.02	63.9
	AA	31.87	70.96	29.37	67.49	16.66	63.9
	RA	30.79	70.96	27.61	67.48	15.63	63.9
	Shortest Path	12.45	81.08	31.82	75.5	7.15	74.64
	Katz	27.4	81.17	38.16	75.37	21.44	74.86
Embedding	Node2Vec	**37.29 ± 8.82**	90.97 ± 0.64	44.33 ± 8.99	94.46 ± 0.59	34.61 ± 2.48	93.14 ± 0.18
	MF	14.29 ± 5.79	80.29 ± 2.26	24.80 ± 4.71	75.92 ± 3.25	19.29 ± 6.29	93.06 ± 0.43
	MLP	31.21 ± 7.90	95.32 ± 0.37	43.53 ± 7.26	94.45 ± 0.32	16.52 ± 4.14	98.34 ± 0.10
GNN	GCN	32.50 ± 6.87	95.01 ± 0.32	50.01 ± 6.04	95.89 ± 0.26	19.94 ± 4.24	98.69 ± 0.06
	GAT	31.86 ± 6.08	93.90 ± 0.32	48.69 ± 7.53	96.25 ± 0.20	18.63 ± 7.75	98.20 ± 0.07
	SAGE	**37.83 ± 7.75**	**95.63 ± 0.27**	47.84 ± 6.39	**97.39 ± 0.15**	22.74 ± 5.47	**98.87 ± 0.04**
	GAE	29.98 ± 3.21	95.08 ± 0.33	**63.33 ± 3.14**	**97.06 ± 0.22**	16.67 ± 0.19	97.47 ± 0.08
GNN+Pairwise Info	SEAL	26.69 ± 5.89	90.59 ± 0.75	39.36 ± 4.99	88.52 ± 1.40	**38.06 ± 5.18**	97.77 ± 0.40
	BUDDY	26.40 ± 4.40	95.06 ± 0.36	**59.48 ± 8.96**	96.72 ± 0.26	23.98 ± 5.11	98.2 ± 0.05
	Neo-GNN	22.65 ± 2.60	93.73 ± 0.36	53.97 ± 5.88	94.89 ± 0.60	31.45 ± 3.17	98.71 ± 0.05
	NCN	32.93 ± 3.80	**96.76 ± 0.18**	54.97 ± 6.03	97.04 ± 0.26	**35.65 ± 4.60**	**98.98 ± 0.04**
	NCNC	29.01 ± 3.83	**96.90 ± 0.28**	**64.03 ± 3.67**	**97.65 ± 0.30**	25.70 ± 4.48	**99.14 ± 0.03**
	NBFNet	**37.69 ± 3.97**	92.85 ± 0.17	38.17 ± 3.06	91.06 ± 0.15	**44.73 ± 2.12**	98.34 ± 0.02
	PEG	22.76 ± 1.84	94.46 ± 0.34	56.12 ± 6.62	96.15 ± 0.41	21.05 ± 2.85	96.97 ± 0.39

One GNN cannot perform across all dataset

Li et al., "Evaluating Graph Neural Networks for Link Prediction: Current Pitfalls and New Benchmarking", NeurIPS 2023 dataset track.

Obstacles for building Graph Foundation Model

Work for different datasets and different tasks

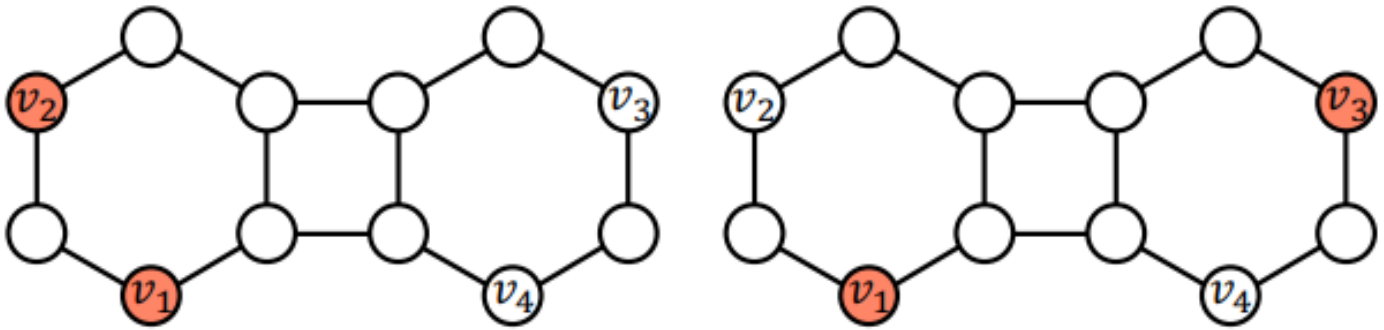
Heuristic evented in 1950s

	Katz	SAGE	BUDDY
POWER	29.85	6.99	19.88

GNNs may underperform simple heuristic in link prediction

Obstacles for building Graph Foundation Model

Work for different datasets and different tasks



Despite GCN performs well on node classification, it cannot distinguish node pairs (v_1, v_2) and (v_3, v_4), essential for Link prediction task

Why current GNNs cannot serve as the foundation?

Understanding for both data and model

We urgent to understand the inner work of both graph data and GNNs!!!

Model mechanism

- **When can GNN work and when not?**
- Why GNNs can not work across different tasks and datasets?

Data mechanism

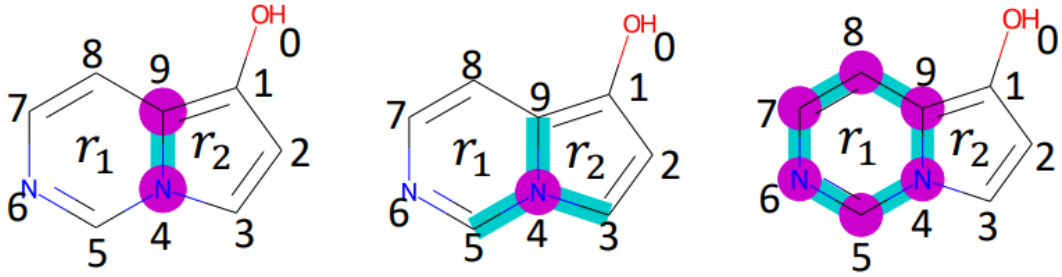
- What is the underlying data formation principle?
- What knowledge is shared? What are different between different datasets and tasks?

Understanding for both data and model

Model mechanism

Different GNNs can have different inductive biases.

Homophily is a primary assumption, advanced GNNs can capture more diverse patterns, e.g., a ring

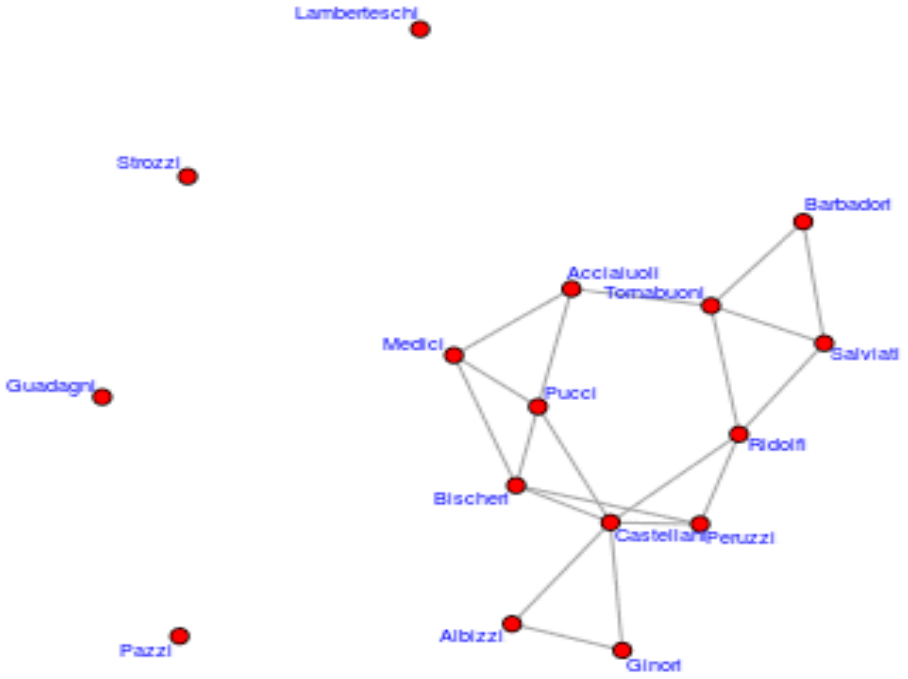


Understanding for both data and model

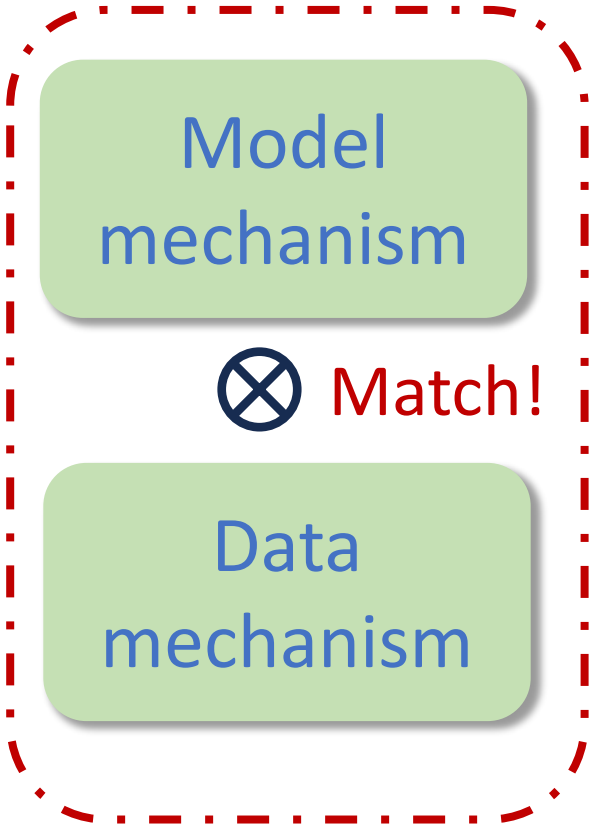
Data mechanism

Different graphs may have different data formulation principles

How is a link formed in a graph?



When can GNNs work and when not



Different GNNs may have different inductive biases

Different graphs may have different data formulation principles

GNNs work when the data mechanism matches the model mechanism!

This talk

When can GNNs work and when not?

Two tasks:

- Link Prediction
- Node Classification

Two mechanisms:

- Data mechanism
- Model mechanism

This talk

When can GNNs work and when not?

Multiple insights:

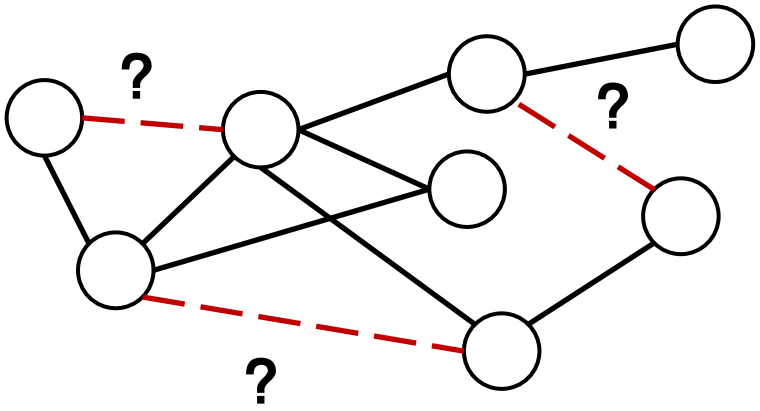
- What is transferable across different tasks and datasets?
- What is the essential difficulty for the model design?
- Basics and instructions for building the Graph Foundation Model

Revisiting Link Prediction: a Data Perspective

Haitao Mao, Juanhui Li, Harry Shomer, Bingheng Li
Wenqi Fan, Yao Ma, Tong Zhao, Neil Shah, Jiliang Tang

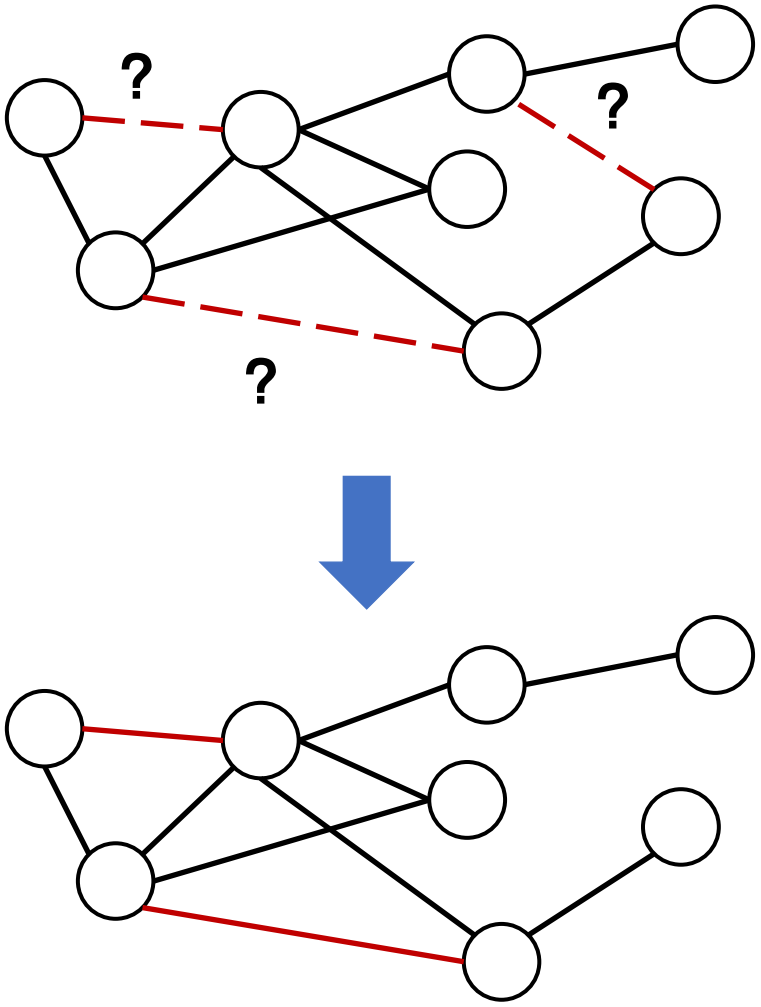
MICHIGAN STATE
UNIVERSITY

Link Prediction task



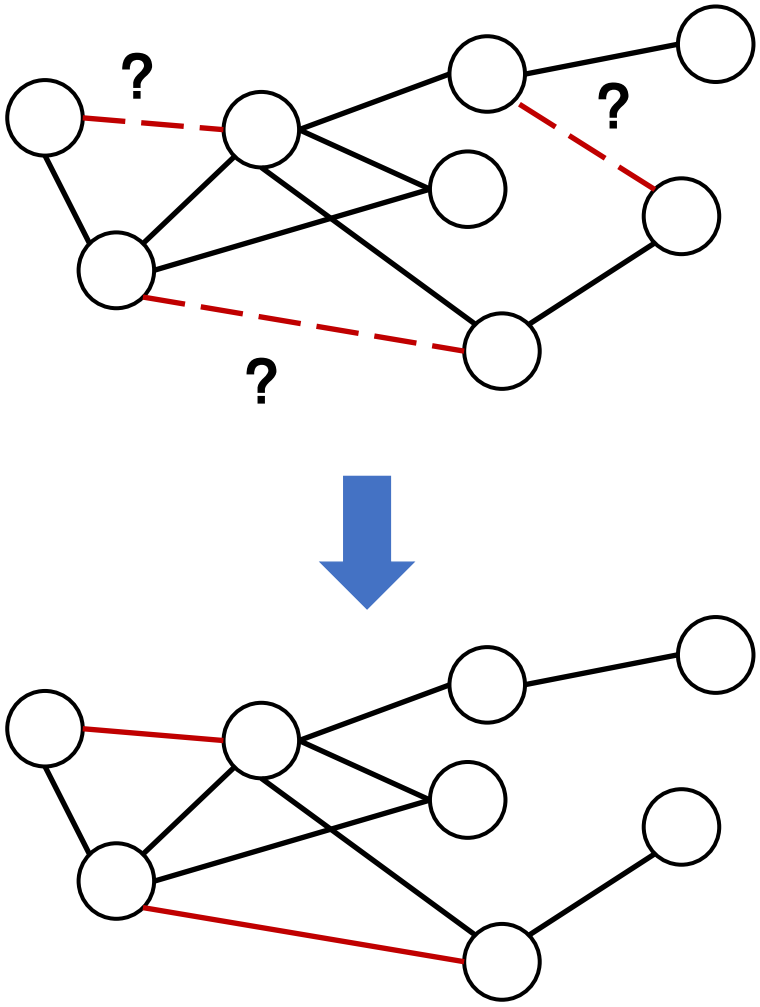
Predict the likelihood of whether two nodes connected

Link Prediction task



Predict the likelihood of whether two nodes connected

Link Prediction task



- Recommend Paper in Google scholar
- Recommend Friend in Twitter
- Predict Protein-Protein Interaction
- Predict Drug-Drug Joint effect

⋮

Preliminary -- Link Prediction performance

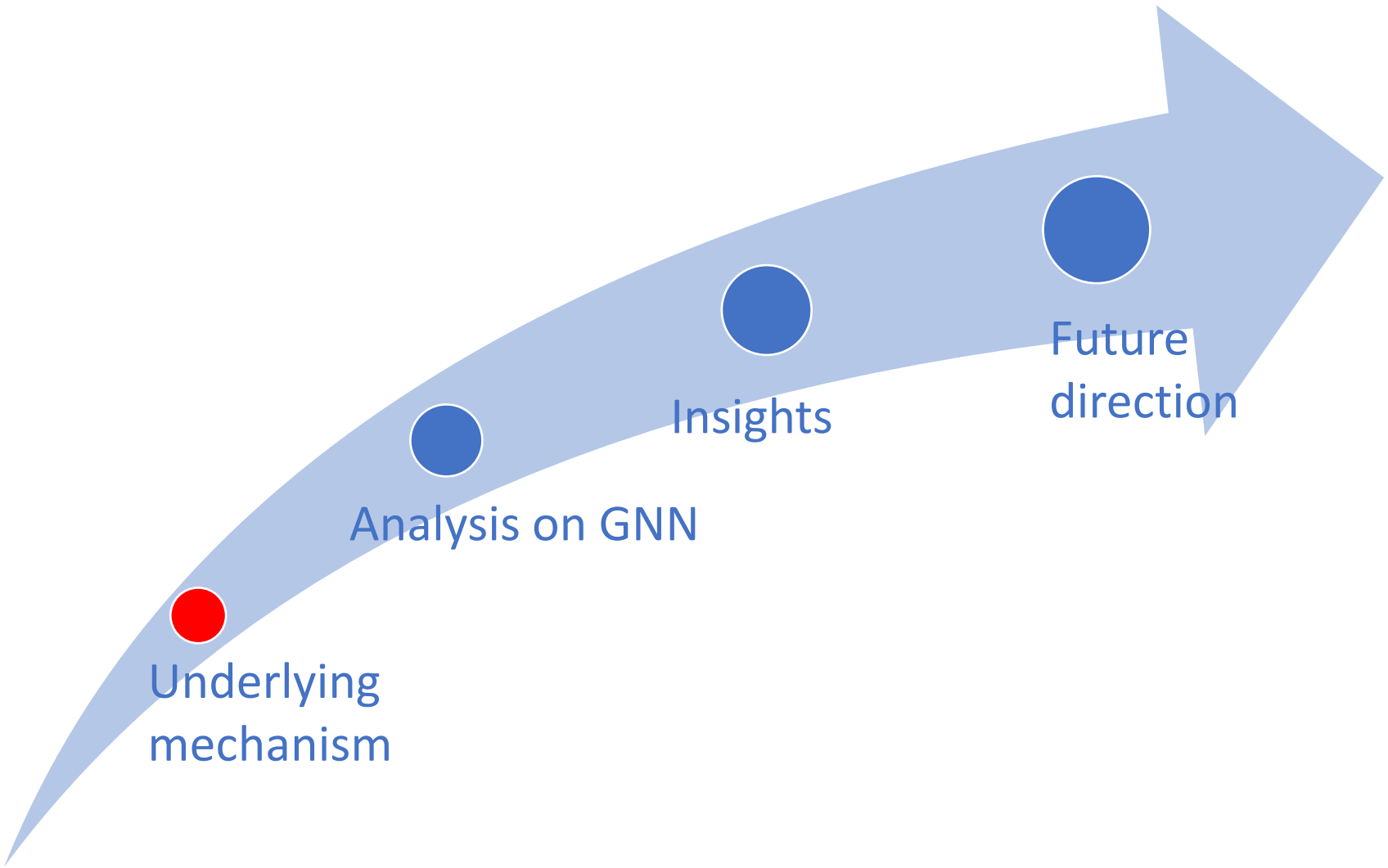
Table 1: Results on Cora, Citeseer, and Pubmed(%) under the existing evaluation setting. Highlighted are the results ranked **first**, **second**, and **third**.

	Models	Cora		Citeseer		Pubmed	
		MRR	AUC	MRR	AUC	MRR	AUC
Heuristic	CN	20.99	70.85	28.34	67.49	14.02	63.9
	AA	31.87	70.96	29.37	67.49	16.66	63.9
	RA	30.79	70.96	27.61	67.48	15.63	63.9
	Shortest Path	12.45	81.08	31.82	75.5	7.15	74.64
	Katz	27.4	81.17	38.16	75.37	21.44	74.86
Embedding	Node2Vec	**37.29 ± 8.82**	90.97 ± 0.64	44.33 ± 8.99	94.46 ± 0.59	34.61 ± 2.48	93.14 ± 0.18
	MF	14.29 ± 5.79	80.29 ± 2.26	24.80 ± 4.71	75.92 ± 3.25	19.29 ± 6.29	93.06 ± 0.43
	MLP	31.21 ± 7.90	95.32 ± 0.37	43.53 ± 7.26	94.45 ± 0.32	16.52 ± 4.14	98.34 ± 0.10
GNN	GCN	32.50 ± 6.87	95.01 ± 0.32	50.01 ± 6.04	95.89 ± 0.26	19.94 ± 4.24	98.69 ± 0.06
	GAT	31.86 ± 6.08	93.90 ± 0.32	48.69 ± 7.53	96.25 ± 0.20	18.63 ± 7.75	98.20 ± 0.07
	SAGE	**37.83 ± 7.75**	**95.63 ± 0.27**	47.84 ± 6.39	**97.39 ± 0.15**	22.74 ± 5.47	**98.87 ± 0.04**
	GAE	29.98 ± 3.21	95.08 ± 0.33	**63.33 ± 3.14**	**97.06 ± 0.22**	16.67 ± 0.19	97.47 ± 0.08
GNN+Pairwise Info	SEAL	26.69 ± 5.89	90.59 ± 0.75	39.36 ± 4.99	88.52 ± 1.40	**38.06 ± 5.18**	97.77 ± 0.40
	BUDDY	26.40 ± 4.40	95.06 ± 0.36	**59.48 ± 8.96**	96.72 ± 0.26	23.98 ± 5.11	98.2 ± 0.05
	Neo-GNN	22.65 ± 2.60	93.73 ± 0.36	53.97 ± 5.88	94.89 ± 0.60	31.45 ± 3.17	98.71 ± 0.05
	NCN	32.93 ± 3.80	**96.76 ± 0.18**	54.97 ± 6.03	97.04 ± 0.26	**35.65 ± 4.60**	**98.98 ± 0.04**
	NCNC	29.01 ± 3.83	**96.90 ± 0.28**	**64.03 ± 3.67**	**97.65 ± 0.30**	25.70 ± 4.48	**99.14 ± 0.03**
	NBFNet	**37.69 ± 3.97**	92.85 ± 0.17	38.17 ± 3.06	91.06 ± 0.15	**44.73 ± 2.12**	98.34 ± 0.02
	PEG	22.76 ± 1.84	94.46 ± 0.34	56.12 ± 6.62	96.15 ± 0.41	21.05 ± 2.85	96.97 ± 0.39

Can one GNN perform well? No consistent winning solution

Li et al., "Evaluating Graph Neural Networks for Link Prediction: Current Pitfalls and New Benchmarking", NeurIPS 2023 dataset track.

Outline


Underlying mechanism
Analysis on GNN
Insights
Future direction

Explore underlying mechanism

When do GNNs perform well on different node pairs and when not?

Model mechanism

Data mechanism

What is the mechanism underlying?

Explore underlying mechanism

Model mechanism

Data mechanism

Explore underlying mechanism

Model mechanism

Data mechanism

Estimate the proximity between node pairs

Explore underlying mechanism

Model mechanism

Data mechanism

How is a link formed in a graph

Explore underlying mechanism

Model mechanism

Data mechanism

Estimate the proximity between node pairs

How is a link formed in a graph

Similarity breeds connection. *This principle-the homophily principle-structures network ties of every type, including marriage, friendship, work, advice, support, information transfer, exchange, comembership, and other types of relationship*

McPherson et al., "Birds of a feather: Homophily in social networks.", Annual Review of Sociology. .

Explore underlying mechanism

Model mechanism

Data mechanism

How is a link formed in a graph

Important data factors

Feature Proximity (FP)

Local Structural Proximity (LSP)

Global Structural Proximity (GSP)

We describe those data factors with heuristic algorithms

Heuristic algorithms

Assign a score to each node pair based on specific heuristic function

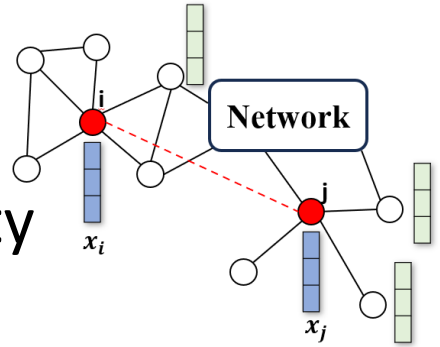
Name	Formula	Factor
common neighbors (CN)	$\lvert\Gamma(i) \cap \Gamma(j)\rvert$	LSP
Adamic-Adar (AA)	$\sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log \lvert\Gamma(k)\rvert}$	LSP
resource allocation (RA)	$\sum_{k \in \Gamma(j) \cap \Gamma(i)} \frac{1}{\lvert\Gamma(k)\rvert}$	LSP
Katz	$\sum_{l=1}^{\infty} \lambda^l \lvert\text{paths}^{\langle l \rangle}(i,j)\rvert$	GSP
Personal PageRank (PPR)	$[\pi_i]_j + [\pi_j]_i$	GSP
SimRank	$\gamma \frac{\sum_{a \in \Gamma(i)} \sum_{b \in \Gamma(j)} \text{score}(i,j)}{\lvert\Gamma(i)\rvert \cdot \lvert\Gamma(j)\rvert}$	GSP
Feature Homophily (FH)	$\text{dis}(x_i, x_j)$	FP

- Interpretable
- Simple
- Derived from expert knowledge

Important data factors

Feature Proximity (FP)

The node feature similarity

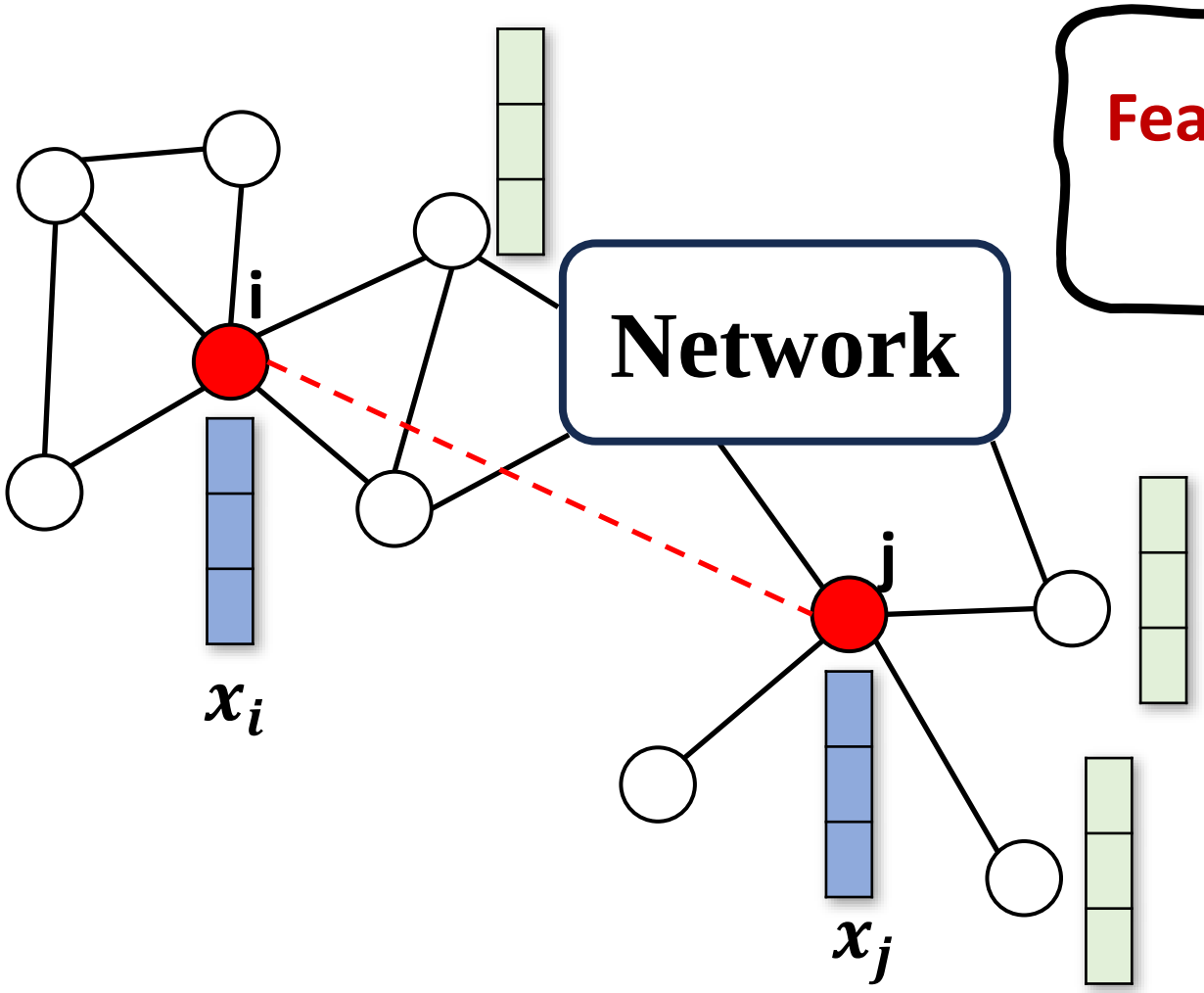


Local Structural Proximity (LSP)

Global Structural Proximity (GSP)

Heuristic algorithms

Feature homophily: $dis(x_i, x_j)$

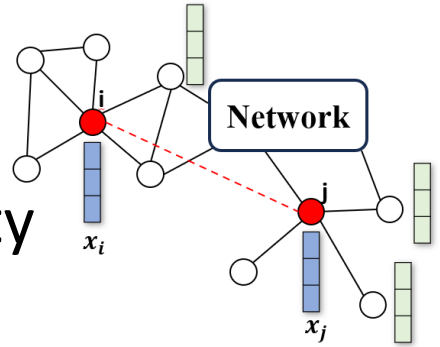


Feature similarity matters

Important data factors

Feature Proximity (FP)

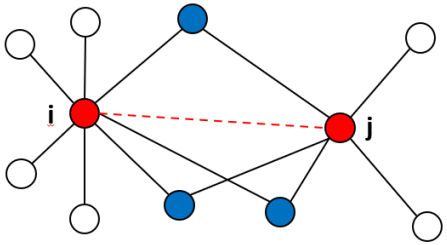
The node feature similarity



Local Structural Proximity (LSP)

The overlapping between immediate neighbors

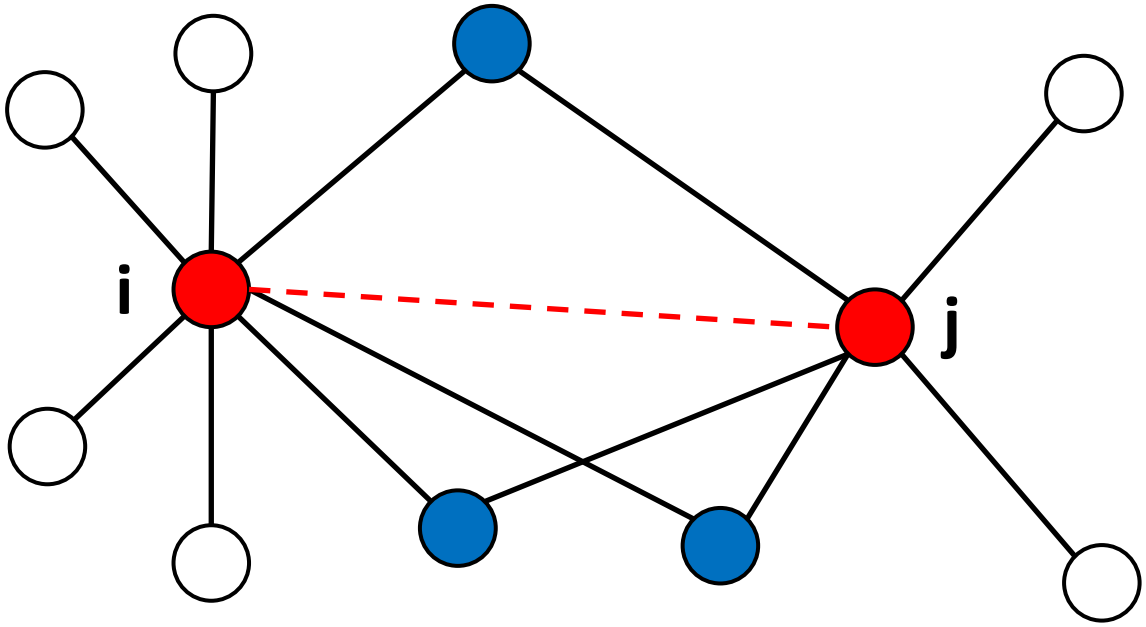
CN, RA, AA, … …



Global Structural Proximity (GSP)

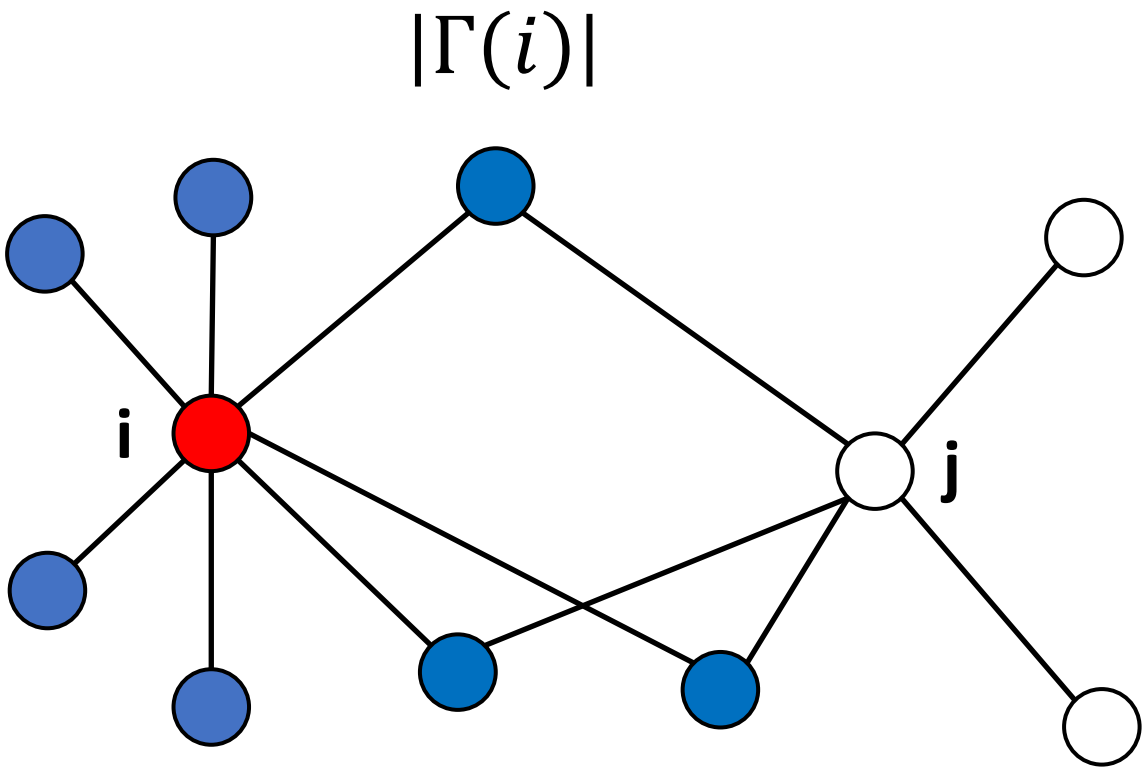
Heuristic algorithms

Common Neighbors (CN): $|\Gamma(i) \cap \Gamma(j)|$



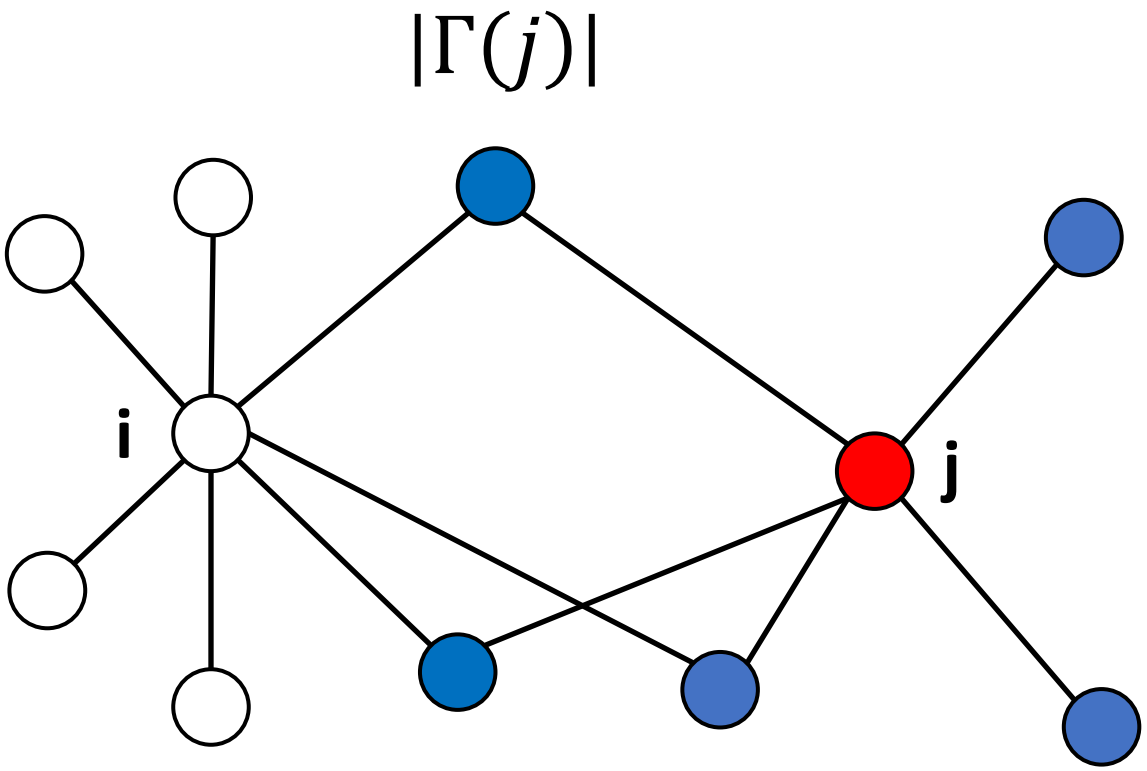
Notations: $\Gamma(i)$ is the neighbor set of node i in the graph

Heuristic algorithms



Notations: $\Gamma(i)$ is the neighbor set of node i in the graph

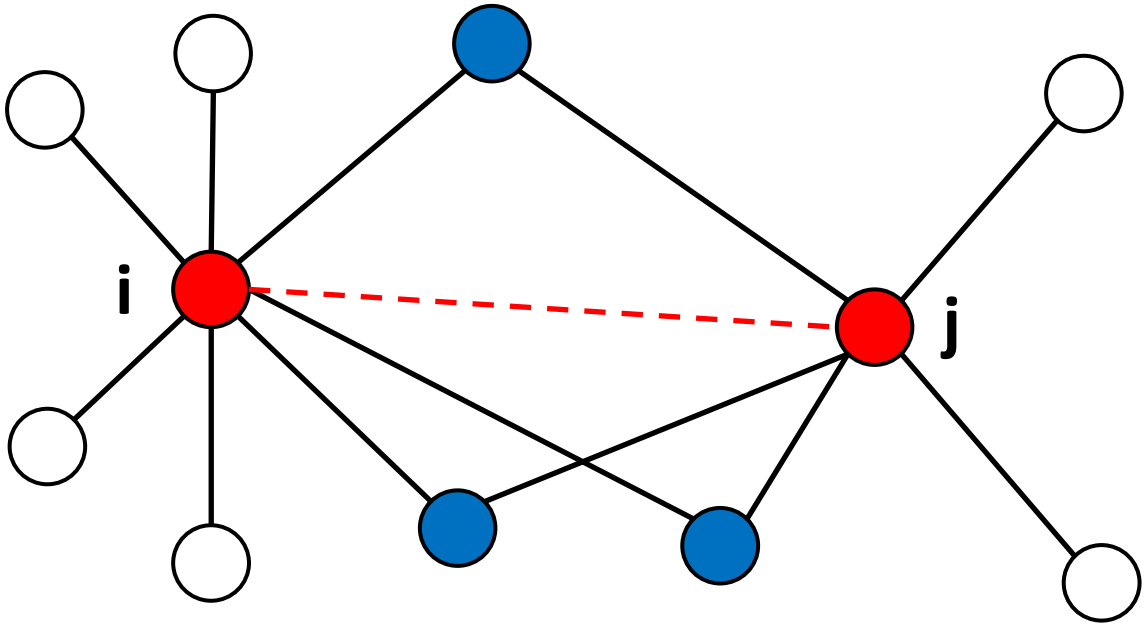
Heuristic algorithms



Notations: $\Gamma(j)$ is the neighbor set of node j in the graph

Heuristic algorithms

Common Neighbors (CN): $|\Gamma(i) \cap \Gamma(j)|$



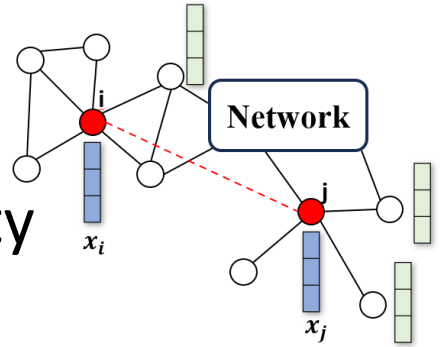
Only first-order neighbors

Notations: $\Gamma(i)$ is the neighbor set of node i in the graph

Important data factors

Feature Proximity (FP)

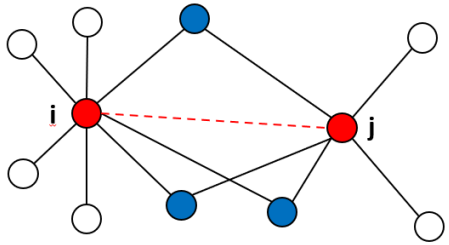
The node feature similarity



Local Structural Proximity (LSP)

The overlapping between immediate neighbors

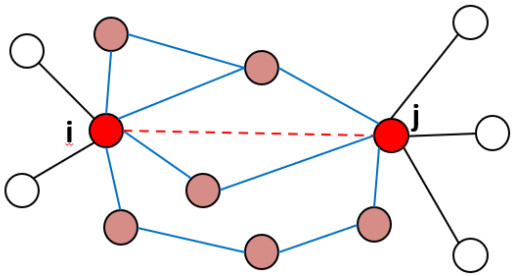
CN, RA, AA, … …



Global Structural Proximity (GSP)

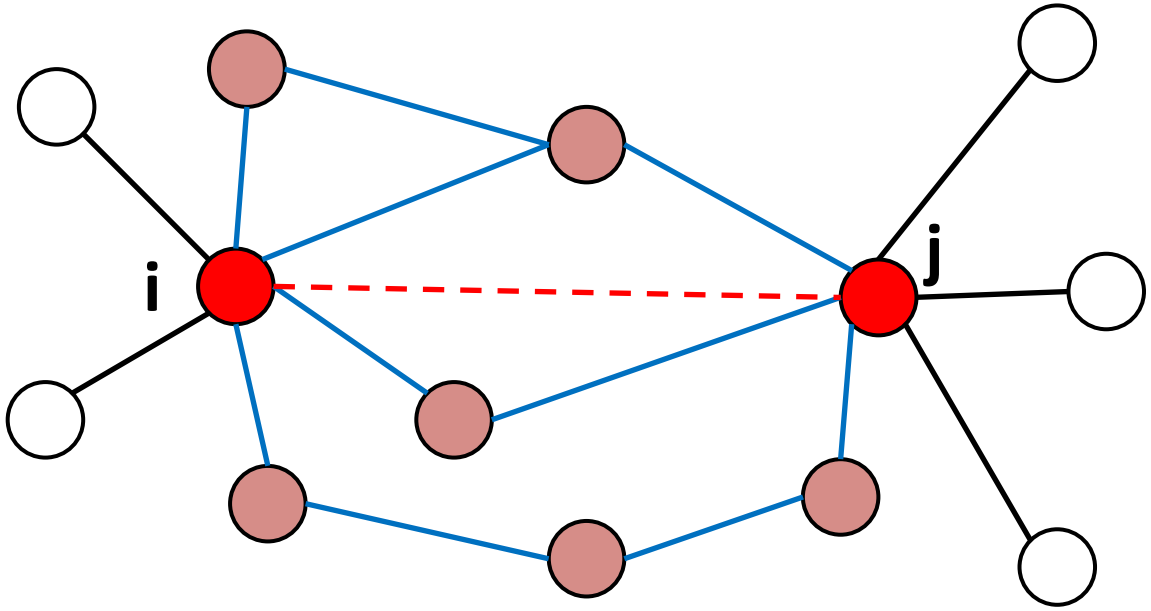
The number of paths with different length

Katz, SimRank, PPR, … …



Heuristic algorithms

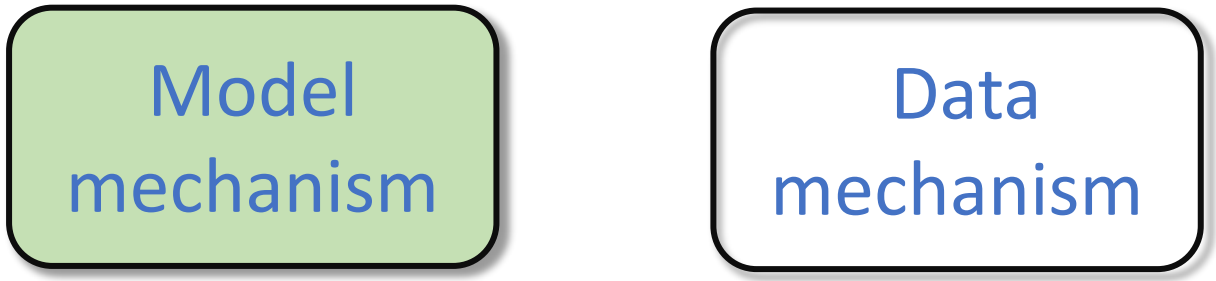
Katz Index: $\sum_{l=1}^{\infty} \boldsymbol{\beta}^{\boldsymbol{l}} |walks(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{l}|$



Higher-order neighbors on the entire graph

Notations: $\beta < 1$ is damping factor

Explore underlying mechanism



- Important data factors

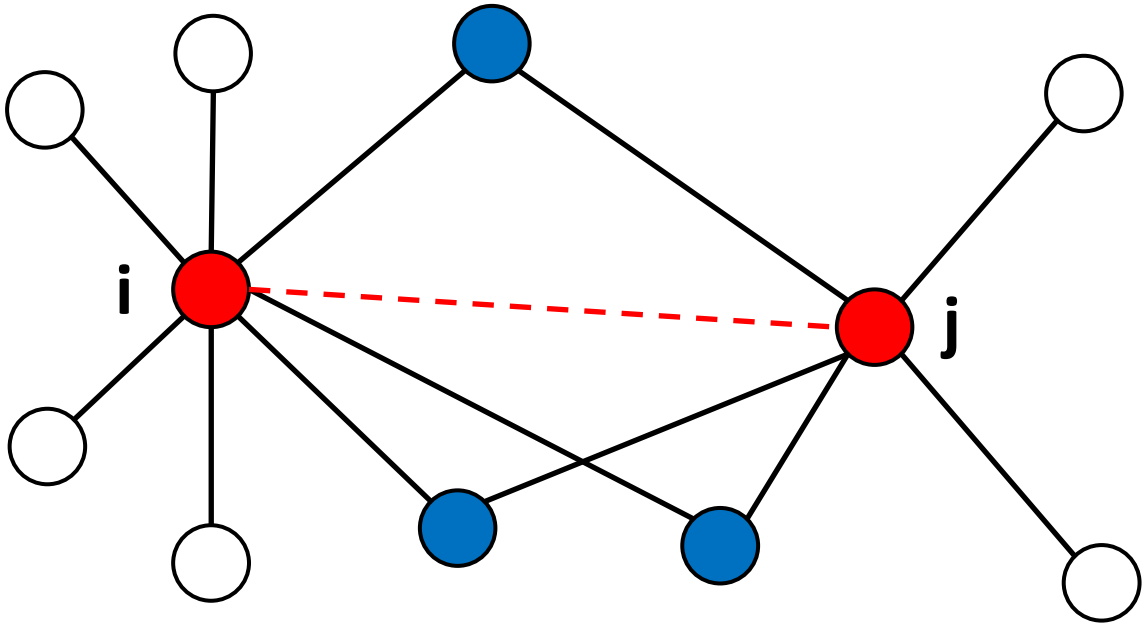
Inspired

- Graph Neural Networks for Link Prediction (GNN4LP)

Different GNNs for Link Prediction incorporate different data factors as the inductive bias.

Local Structural Proximity

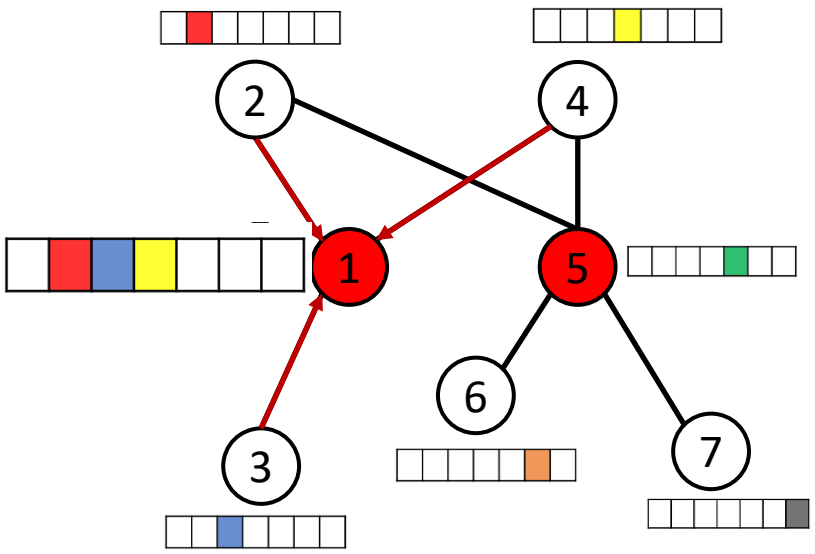
Common Neighbors (CN): $|\Gamma(i) \cap \Gamma(j)|$



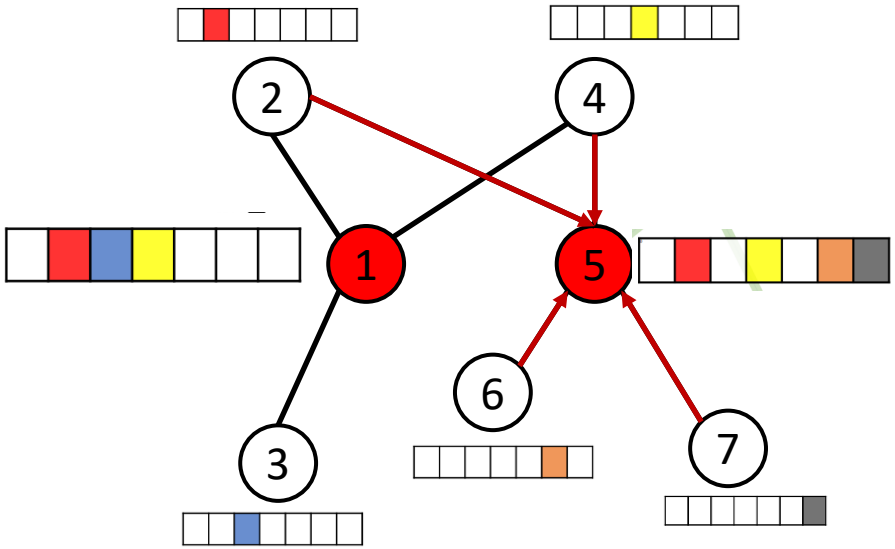
Only first-order neighbors

Notations: $\Gamma(i)$ is the neighbor set of node i in the graph

Graph Neural Network -- Neighborhood Overlap GNN

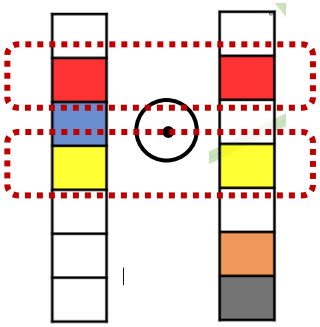


Graph Neural Network -- Neighborhood Overlap GNN

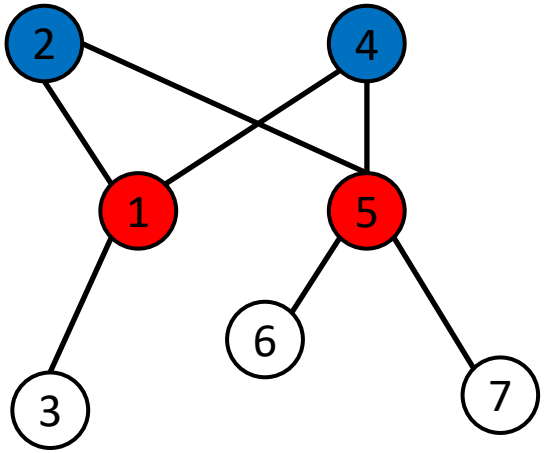


Graph Neural Network -- Neighborhood Overlap GNN

Two common neighbors

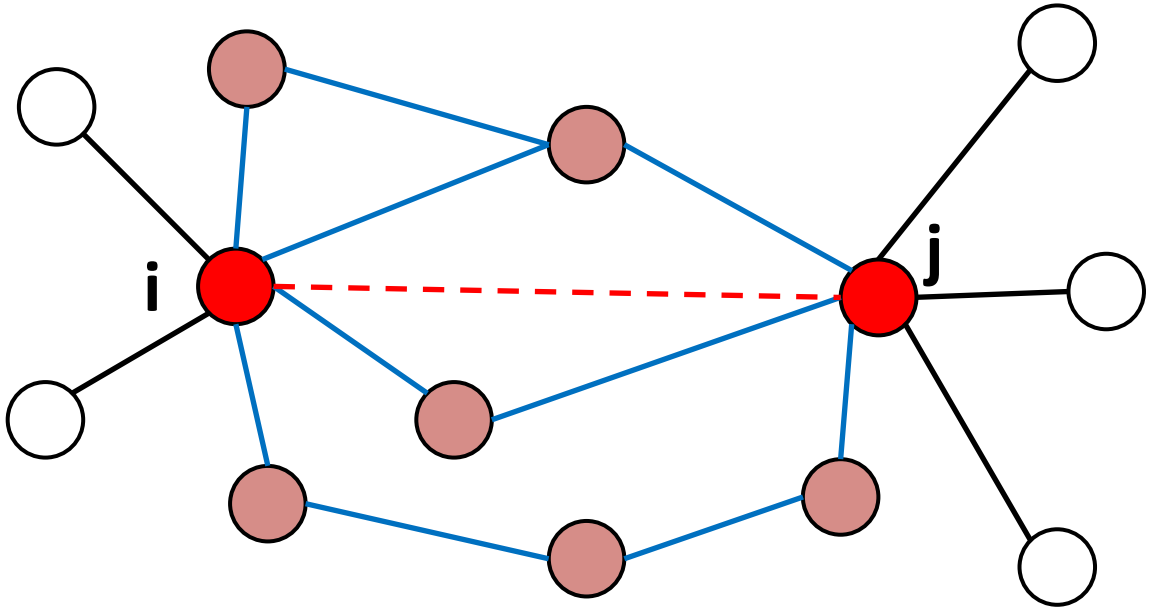


Identify the important Neighborhood similarity information



Global Structural Proximity

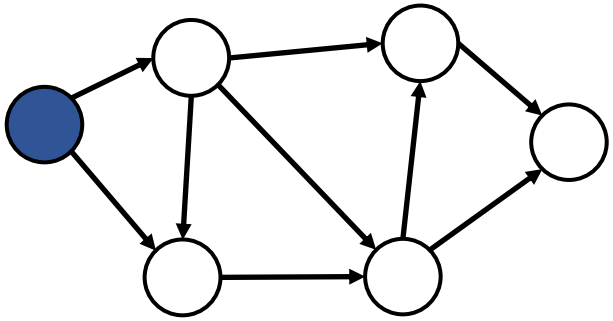
Katz Index: $\sum_{l=1}^{\infty} \boldsymbol{\beta}^{\boldsymbol{l}} |walks(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{l}|$



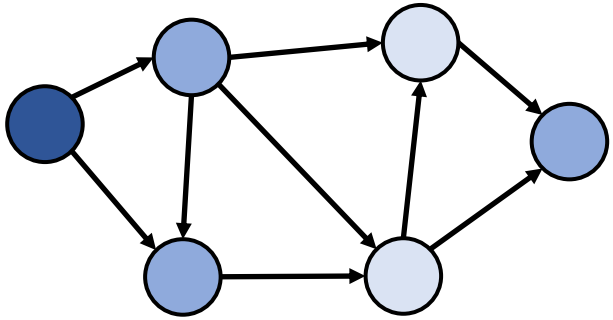
Higher-order neighbors on the entire graph

Notations: $\beta < 1$ is damping factor

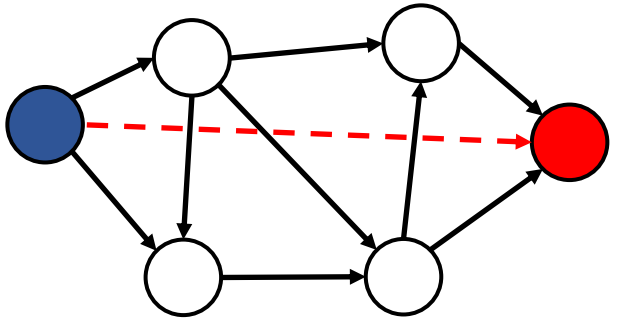
Graph Neural Network -- NBFNet

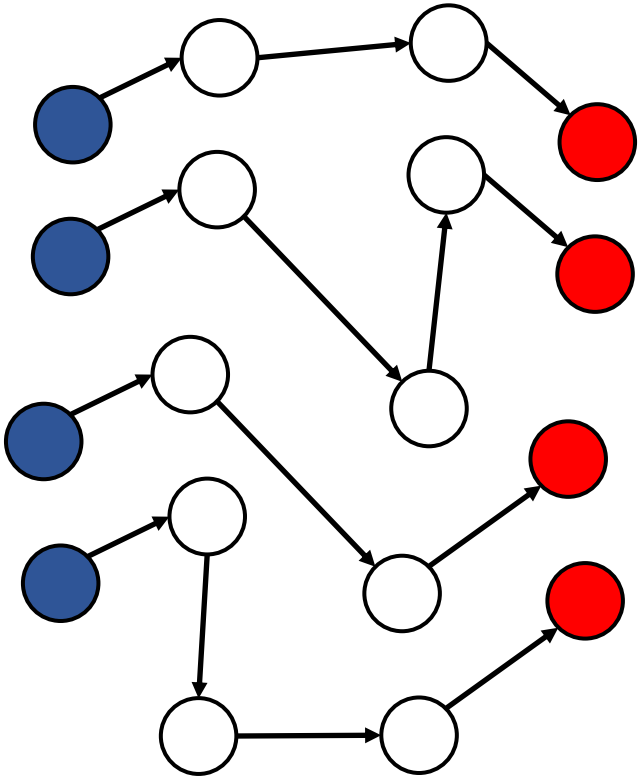


GNN



MLP





0.21

0.05

0.28

0.11

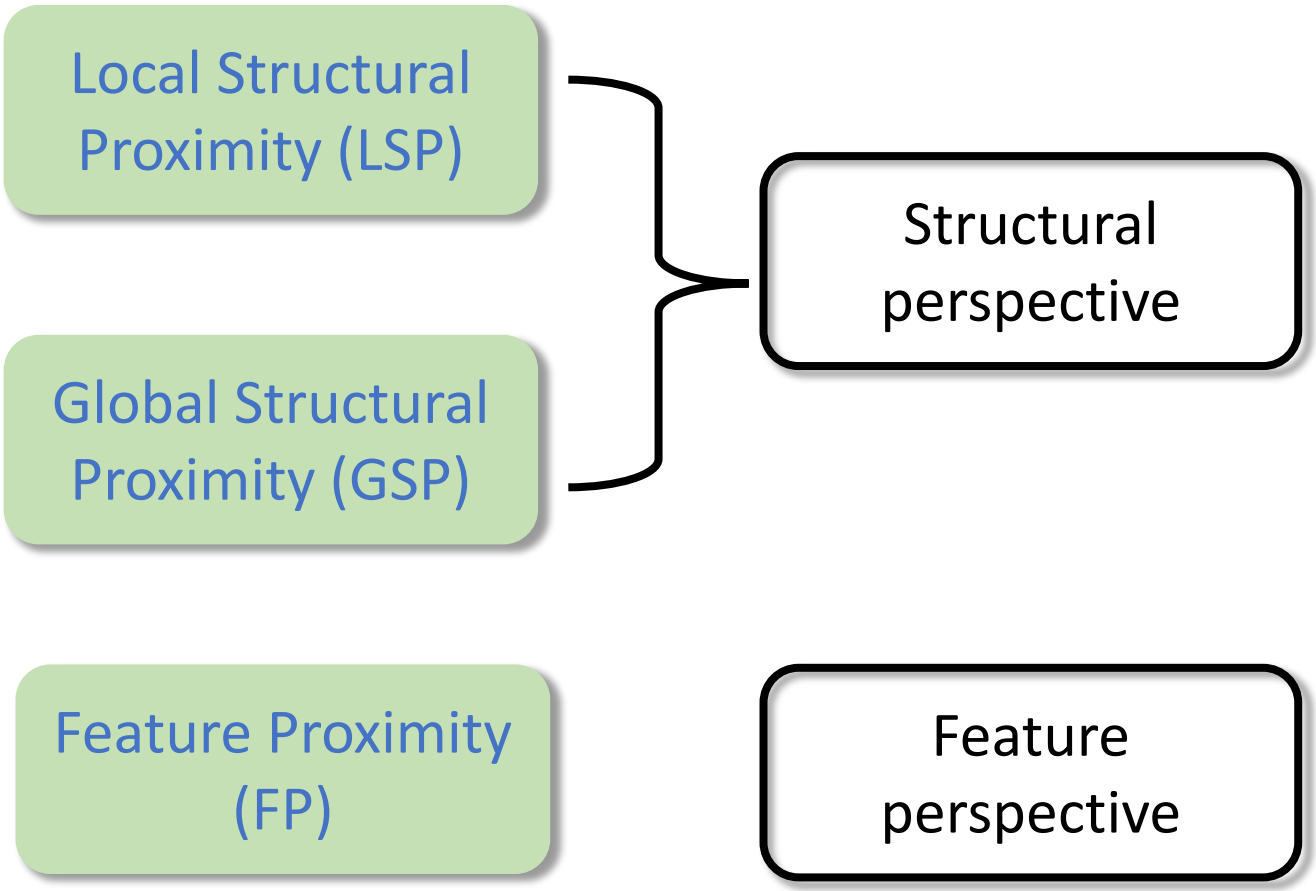
Identify important path information

A recap

Local Structural Proximity (LSP)	The overlapping between immediate neighbors
Global Structural Proximity (GSP)	The number of paths with different length
Feature Proximity (FP)	The node feature similarity

Rather than concepts, we need a theoretical model for a rigorous discussion.

Desired properties for the theoretical model



Desired properties for the theoretical model

Structural perspective

Feature perspective

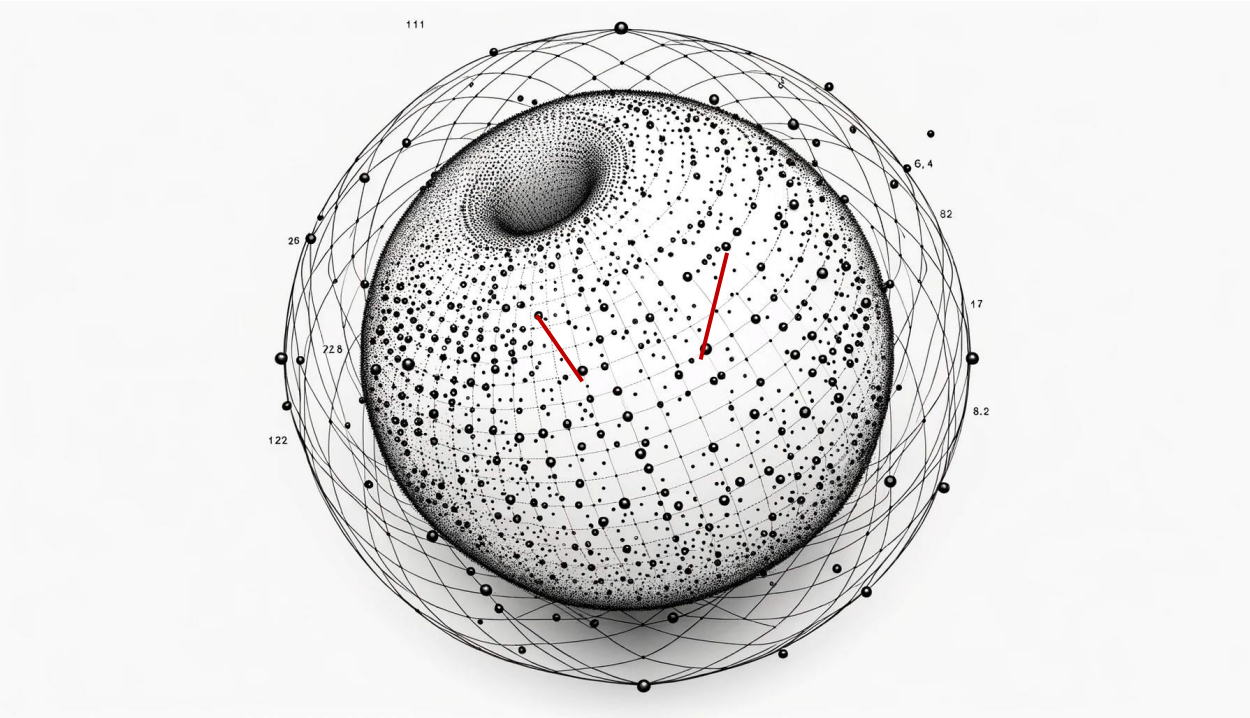
Instead of focusing on one node, similarity between a pair of node should be the focus.

Desired properties for the theoretical model

Structural perspective

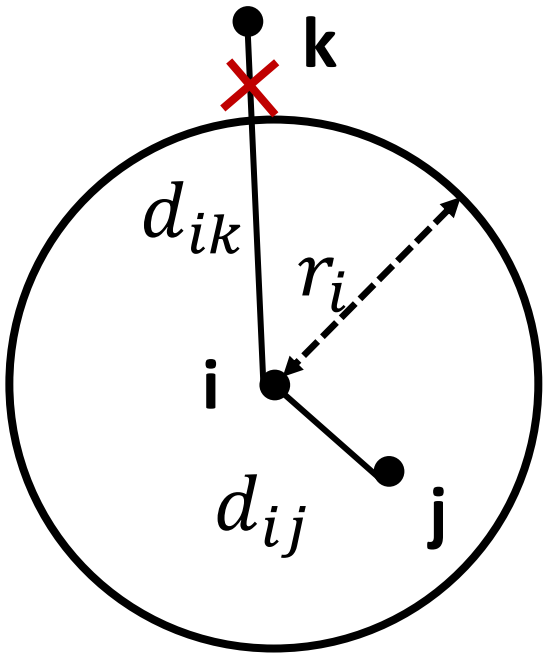
Feature perspective

Latent space Model



Each node is associated with a location in a D-dimensional latent space

Structural perspective



Structural perspective

- r_i is the accept threshold for node i

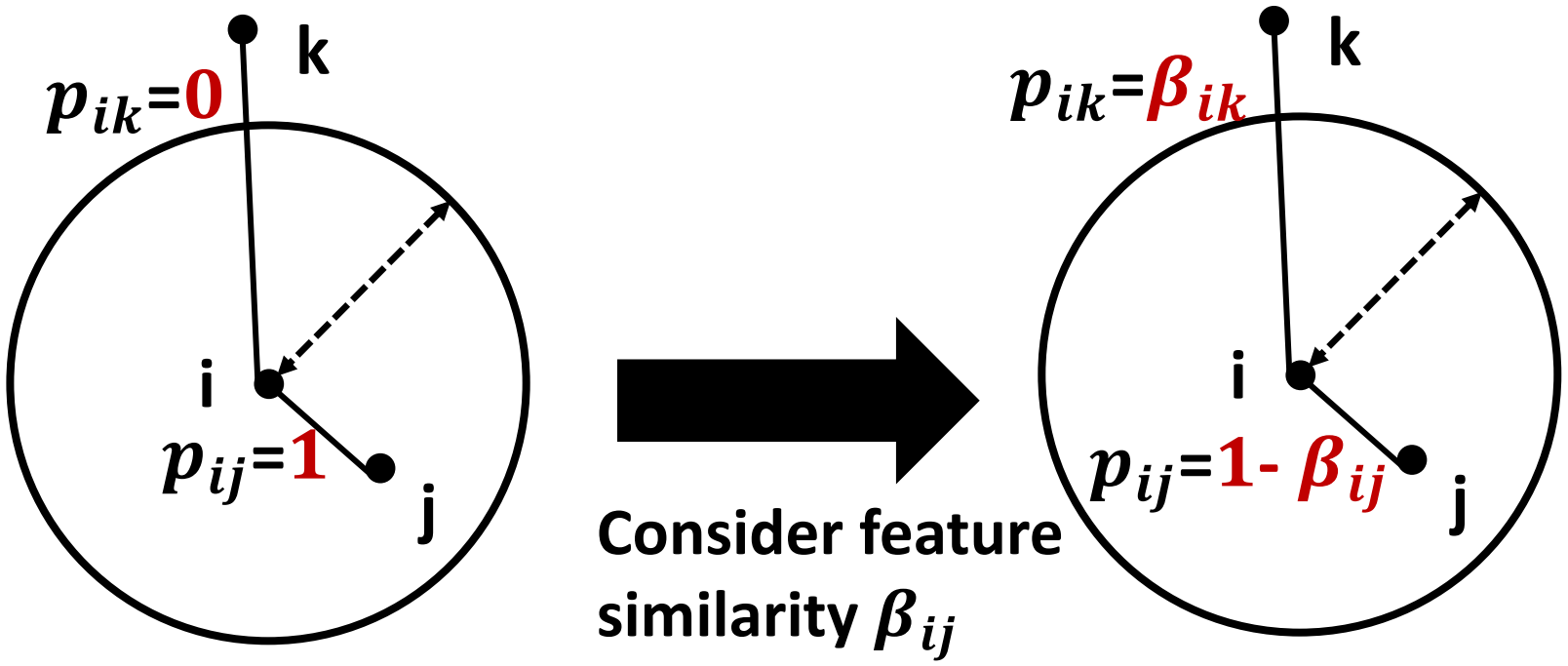
d is the structural similarity distance

- When $d_{ij} \leq r_i$, there is an edge connected between i and j
- When $d_{ik} > r_i$, there is no edge connected between i and k

d_{ij} can be viewed as the social closeness

Feature perspective

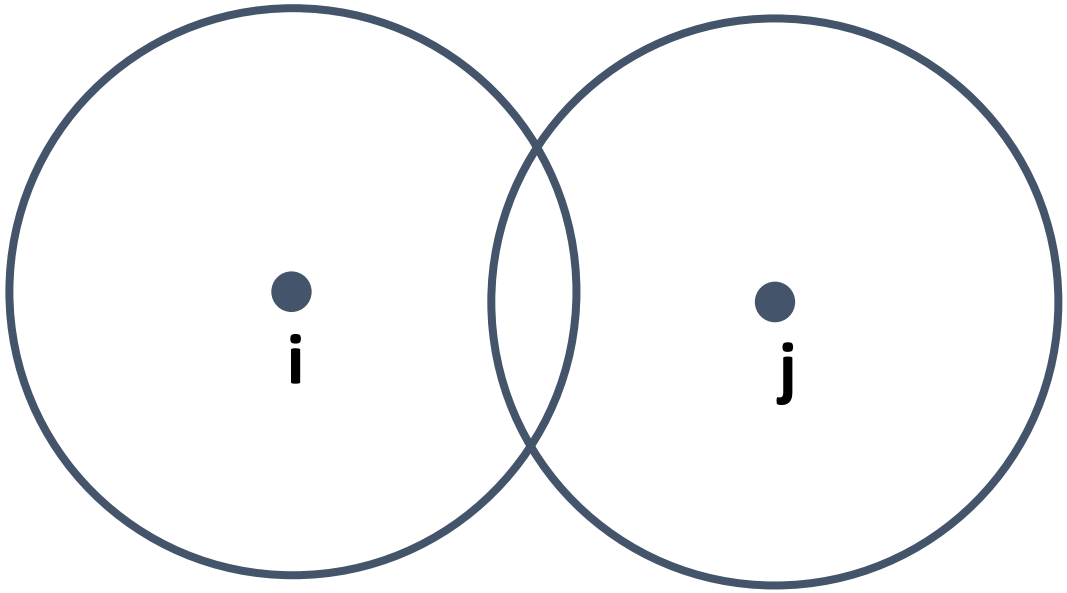
β_{ij} is the feature similarity between node i and j



Structural Only

Structural and Feature

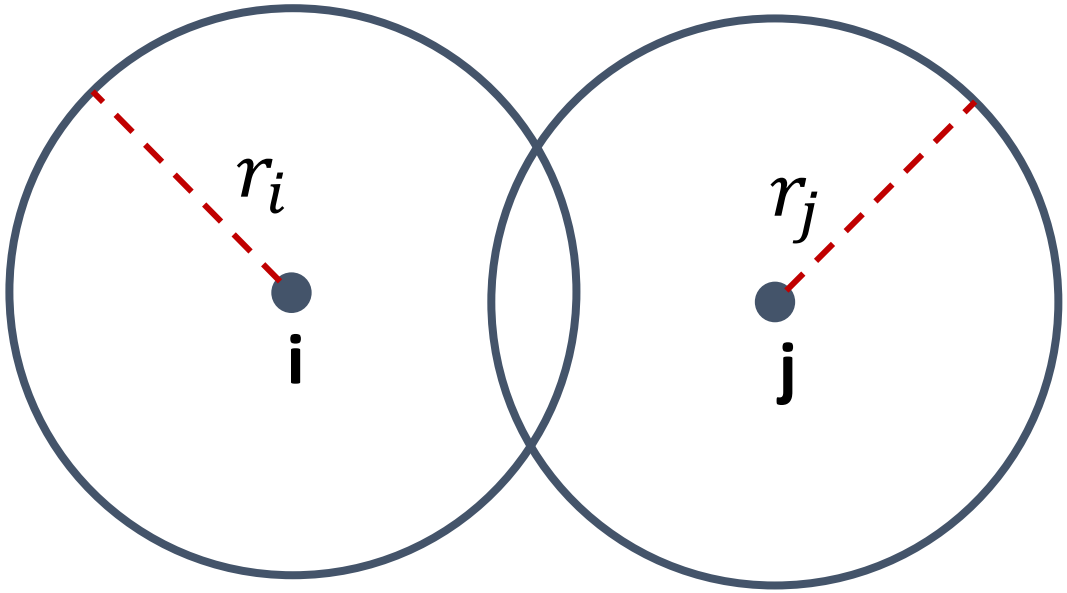
Estimate CN in latent space model



$$\left(\frac{r_i + r_j - d_{ij}}{2}\right)^D \leq \frac{A\left(r_i, r_i, d_{ij}\right)}{V(1)} \leq \left(r_{ij}^{max} - \left(\frac{d_{ij}}{2}\right)^2\right)^{D/2}$$

Common neighbors of two nodes must lie in the intersection $A(r_i, r_j, d_{ij})$.

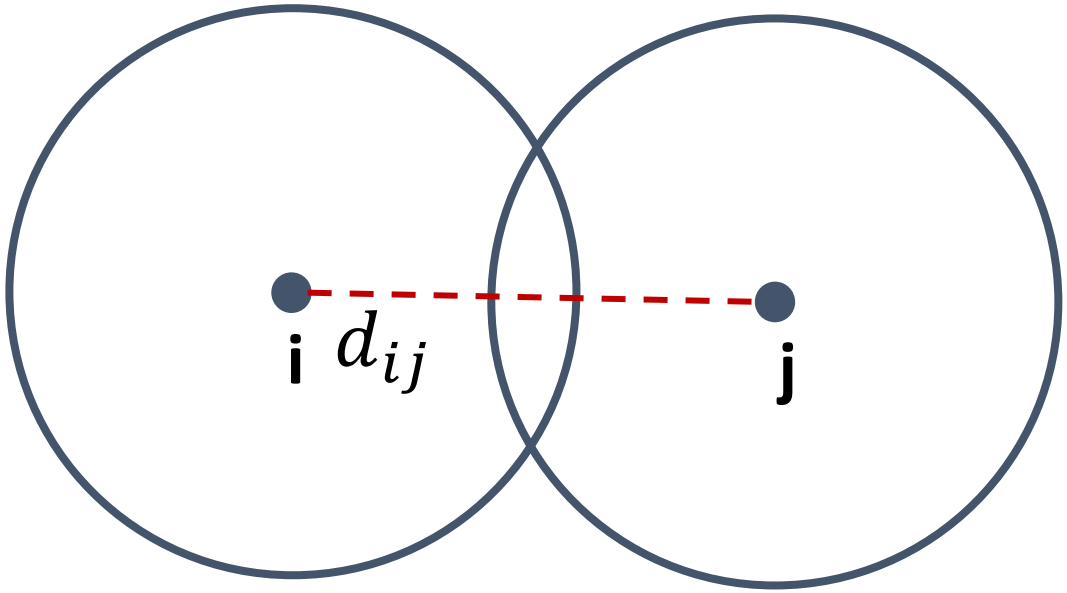
Estimate CN in latent space model



$$\left(\frac{r_i + r_j - d_{ij}}{2}\right)^D \leq \frac{A\left(r_i, r_i, d_{ij}\right)}{V(1)} \leq \left(r_{ij}^{max} - \left(\frac{d_{ij}}{2}\right)^2\right)^{D/2}$$

Common neighbors of two nodes must lie in the intersection $A(r_i, r_j, d_{ij})$.

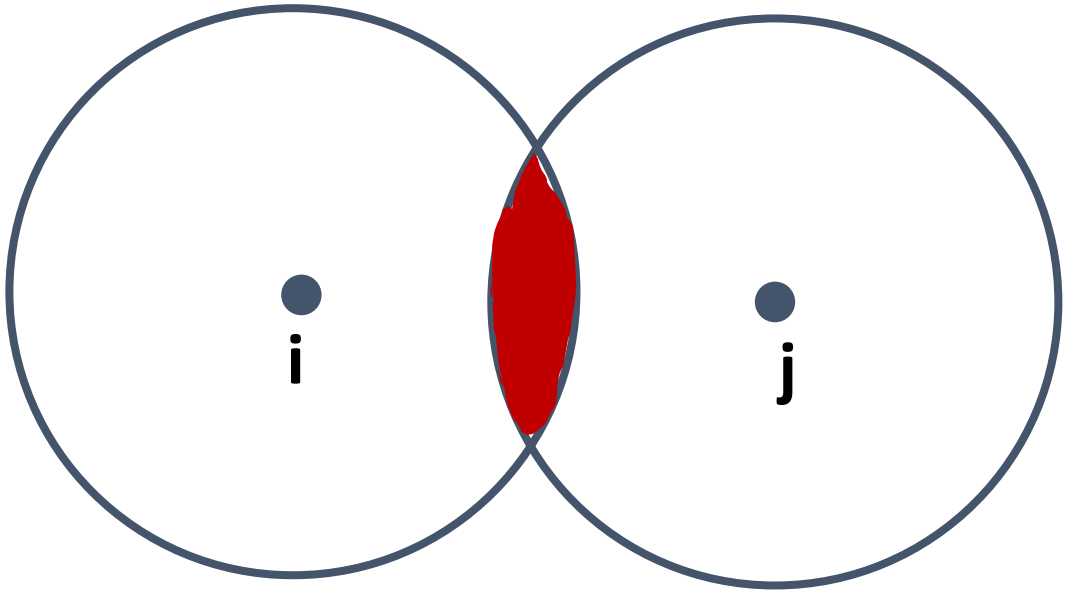
Estimate CN in latent space model



$$\left(\frac{r_i + r_j - d_{ij}}{2}\right)^D \leq \frac{A\left(r_i, r_i, d_{ij}\right)}{V(1)} \leq \left(r_{ij}^{max} - \left(\frac{d_{ij}}{2}\right)^2\right)^{D/2}$$

Common neighbors of two nodes must lie in the intersection $A(r_i, r_j, d_{ij})$.

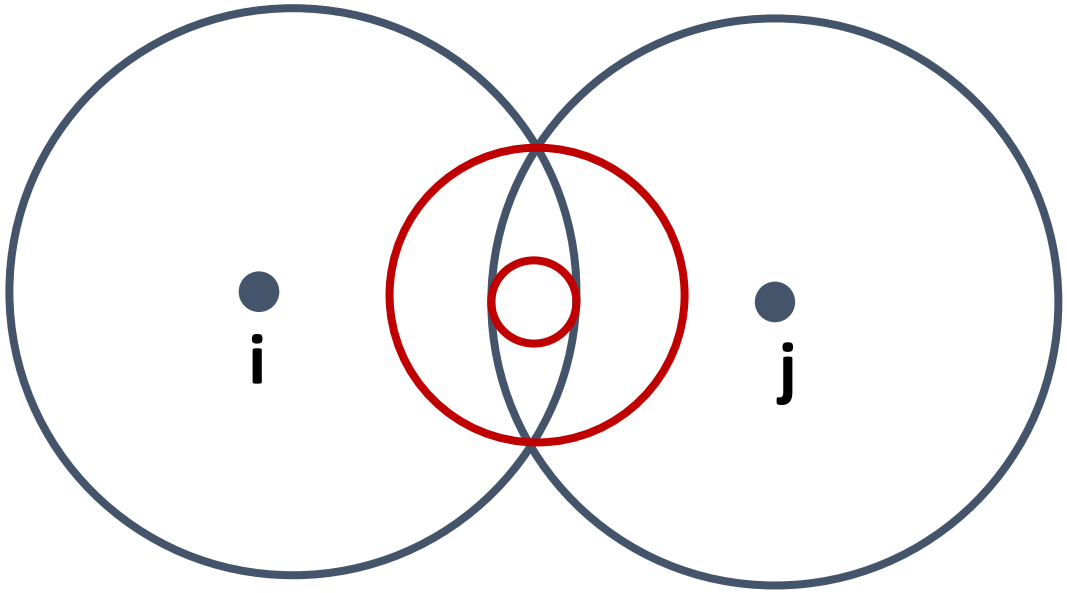
Estimate CN in latent space model



$$\left(\frac{r_i + r_j - d_{ij}}{2}\right)^D \leq \frac{A\left(r_i, r_i, d_{ij}\right)}{V(1)} \leq \left(r_{ij}^{max} - \left(\frac{d_{ij}}{2}\right)^2\right)^{D/2}$$

Common neighbors of two nodes must lie in the intersection $A(r_i, r_j, d_{ij})$.

Estimate CN in latent space model



$$\left(\frac{r_i + r_j - d_{ij}}{2}\right)^D \leq \frac{A\left(r_i, r_i, d_{ij}\right)}{V(1)} \leq \left(r_{ij}^{max} - \left(\frac{d_{ij}}{2}\right)^2\right)^{D/2}$$

Common neighbors of two nodes must lie in the intersection $A(r_i, r_j, d_{ij})$.

The effectiveness of the LSP factor

With the proper latent space model, we can then provide theoretical evidences on the effectiveness of those factors

LSP

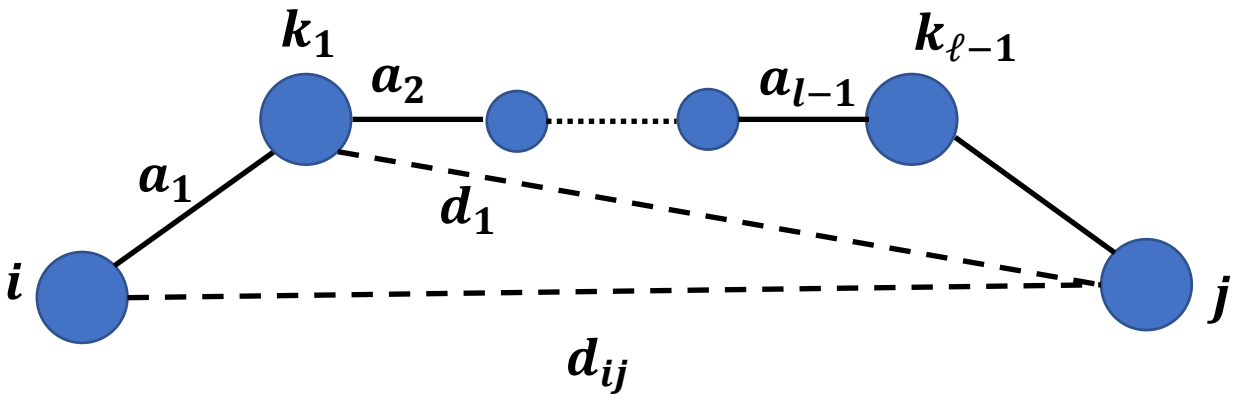
$$d_{ij} \leq 2\sqrt{r_{ij}^{max} - \left(\frac{\eta_{ij}/N - \epsilon}{V(1)}\right)^{2/D}}$$

Structural distance

The number of CNs

When η_{ij} is large, distance d_{ij} is smaller, indicating more likely to be connected

Estimate the number of path in latent space model



Triangulation for bounding d_{ij} using ℓ-hop paths

$$d_{ij} < a_1 + d_1$$

$$d_1 < a_2 + d_2$$

$$\vdots$$

$$d_{\ell-1} < a_{\ell-1} + d_{\ell-1}$$

The effectiveness of the GSP factor

GSP

Structural distance

The number of paths with length ℓ

$$d_{ij} \leq \sum_{n=0}^{M-2} r_n + 2\sqrt{r_M^{\max} - \left(\frac{\eta_\ell(i,j) - b(N,\delta)}{c(N,\delta,\ell)}\right)^{\frac{2}{D(\ell-1)}}}$$

When $\eta_\ell(i,j)$ is large, distance d_{ij} is smaller, indicating more likely to be connected

The effectiveness of the FP factor

FP

Structural distance d_{ij} **The feature similarity** β_{ij}

$$d_{ij} \leq 2\sqrt{r_{ij}^{max} - \left(\frac{\beta_{ij}(1-A(r_i,r_j,d_{ij}))+A(r_i,r_j,d_{ij})}{V(1)}\right)^{2/D}},$$

When β_{ij} is large, distance d_{ij} is smaller,
indicating more likely to be connected

The effectiveness of data factors

Theoretical effectiveness verifies in all data factors

LSP

$$d_{ij} \leq 2\sqrt{r_{ij}^{max} - \left(\frac{\eta_{ij}/N - \epsilon}{V(1)}\right)^{2/D}}$$

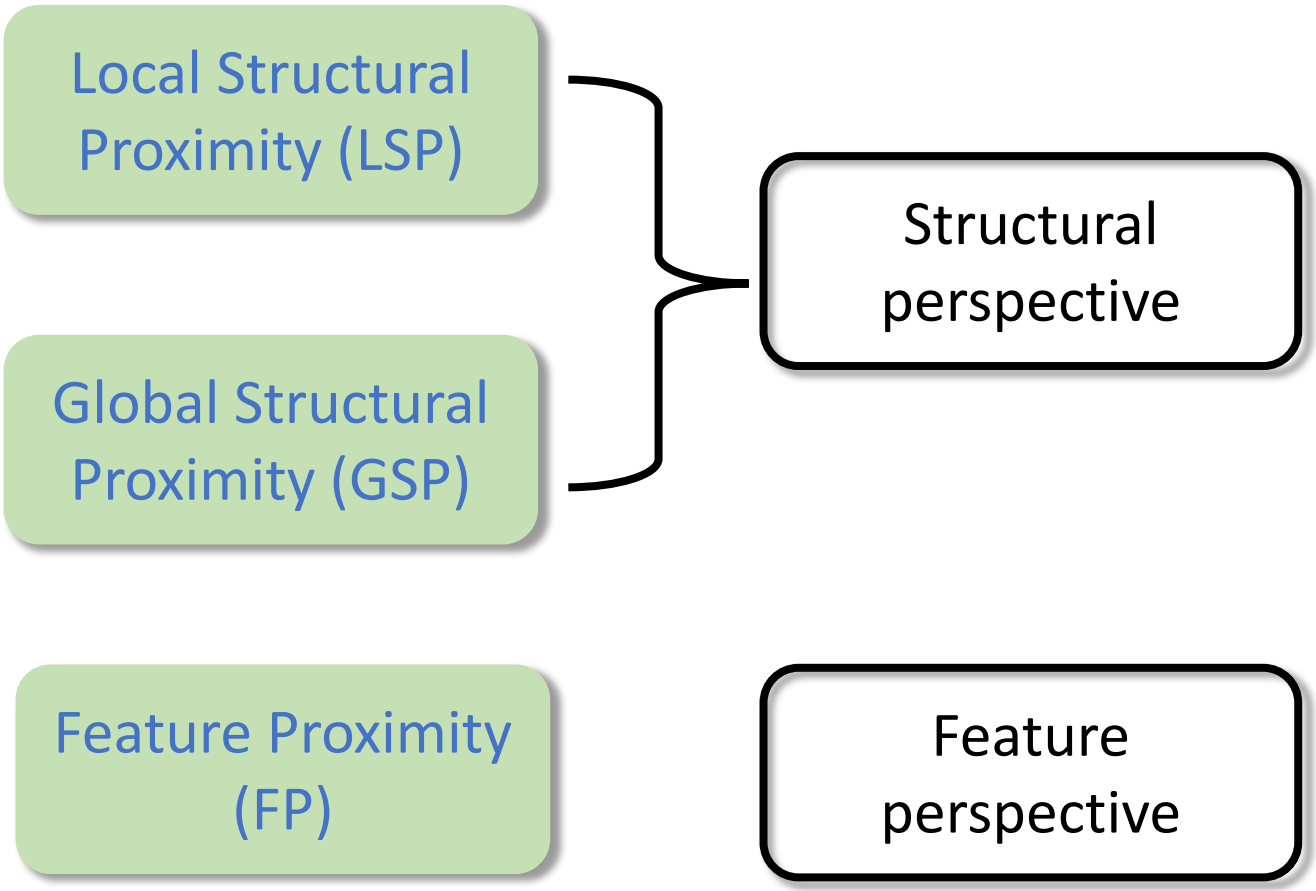
GSP

$$d_{ij} \leq \sum_{n=0}^{M-2} r_n + 2\sqrt{r_M^{\max} - \left(\frac{\eta_\ell(i,j) - b(N,\delta)}{c(N,\delta,\ell)}\right)^{\frac{2}{D(\ell-1)}}}$$

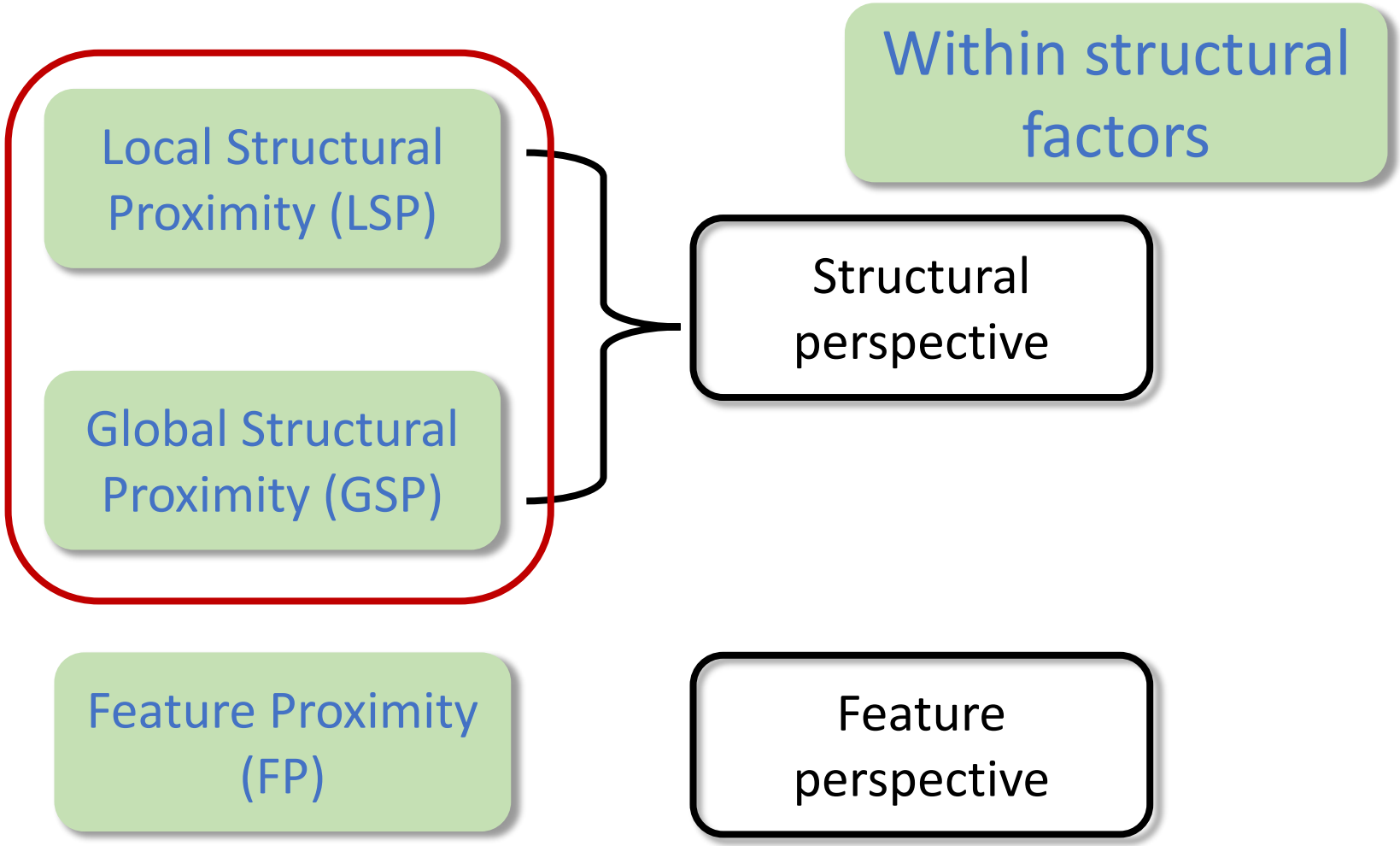
FP

$$d_{ij} \leq 2\sqrt{r_{ij}^{max} - \left(\frac{\beta_{ij}(1 - A(r_i, r_j, d_{ij})) + A(r_i, r_j, d_{ij})}{V(1)}\right)^{2/D}},$$

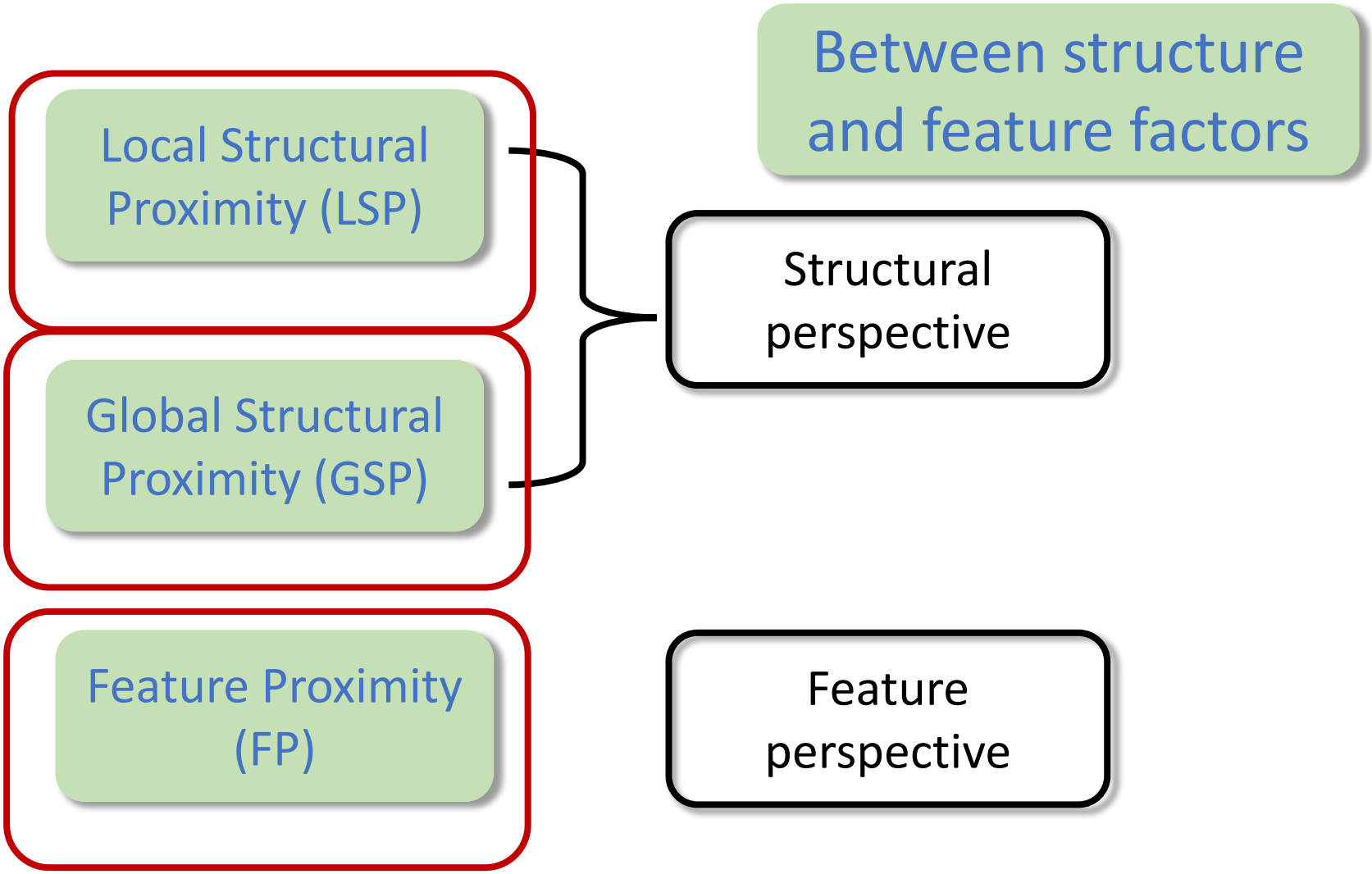
The relationship between different factors



The relationship between different factors



The relationship between different factors



The relationship between different factors

Within structural factors

GSP vs LSP

Between structure and feature factors

LSP vs FP

GSP vs FP

LSP vs GSP

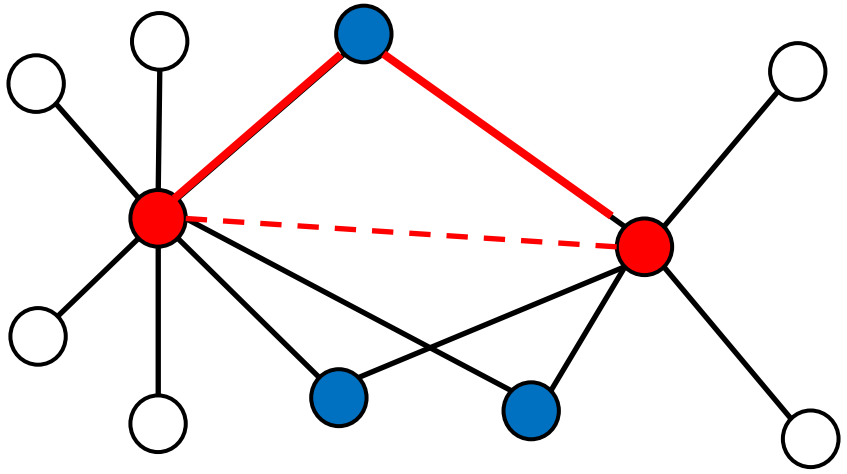
Within structural factors

GSP vs LSP

To achieve this, we compare the effectiveness of LSP and GSP factors via examining which factor can provide a tighter bound on d_{ij}

LSP vs GSP

Common Neighbors (CN)



GSP vs LSP

CN can be viewed as path with length 2

LSP vs GSP

Given the same number of η_{ij}:

$$d_{ij} \leq \sum_{n=0}^{M-2} r_n + 2\sqrt{r_M^{max} - \left(\sqrt{\frac{N \ln(1/\delta)}{2}} - 1\right)^{\frac{2}{D^{(\ell-1)}}}},$$

Structural distance

The length of path

The bound on d_{ij} is exponentially loose with increasing path length ℓ, where LSP corresponding to path length $\ell = 2$

LSP vs GSP

Given the same number of η_{ij}:

The bound on d_{ij} is exponentially loose with increasing path length ℓ, where LSP corresponding to path length $\ell = 2$

When both LSP & GSP exist

LSP provides a tighter bound than GSP

When only GSP exist

GSP is only useful when LSP is absent

The relationship between different factors

Within structural factors

GSP vs LSP

Between structure and feature factors

LSP vs FP

GSP vs FP

LSP vs FP

To achieve this, we examine for the relationship between the number of CNs η_{ij} and the feature similarity β_{ij} in a node pair.

Between structure and feature factors

LSP vs FP

GSP vs FP

LSP vs FP

For a given node pair:

$$\eta_{ij} = \frac{c'}{1-\beta_{ij}} + N(1+\epsilon)$$

$c' < 0$

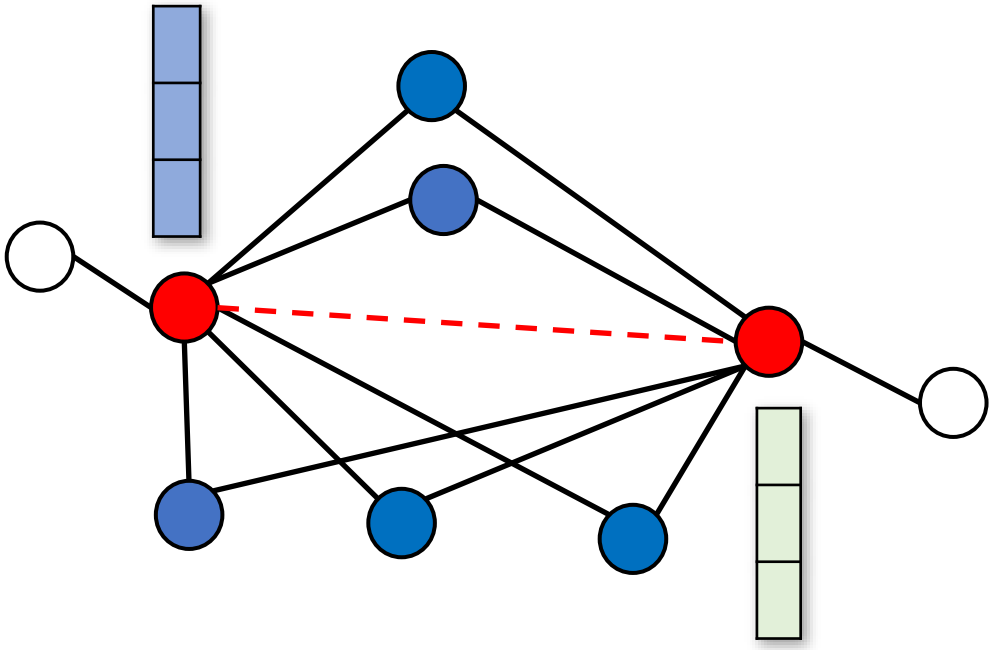
The number of CNs

Feature proximity

Node pairs (i, j) with a large number of CNs η_{ij} is more likely to be with small feature proximity β_{ij}

The incompatibility between LSP and FP

Example of incompatibility between LSP and FP



Two persons with a lot of common neighbors but still not connected is because of the low feature similarity

A recap

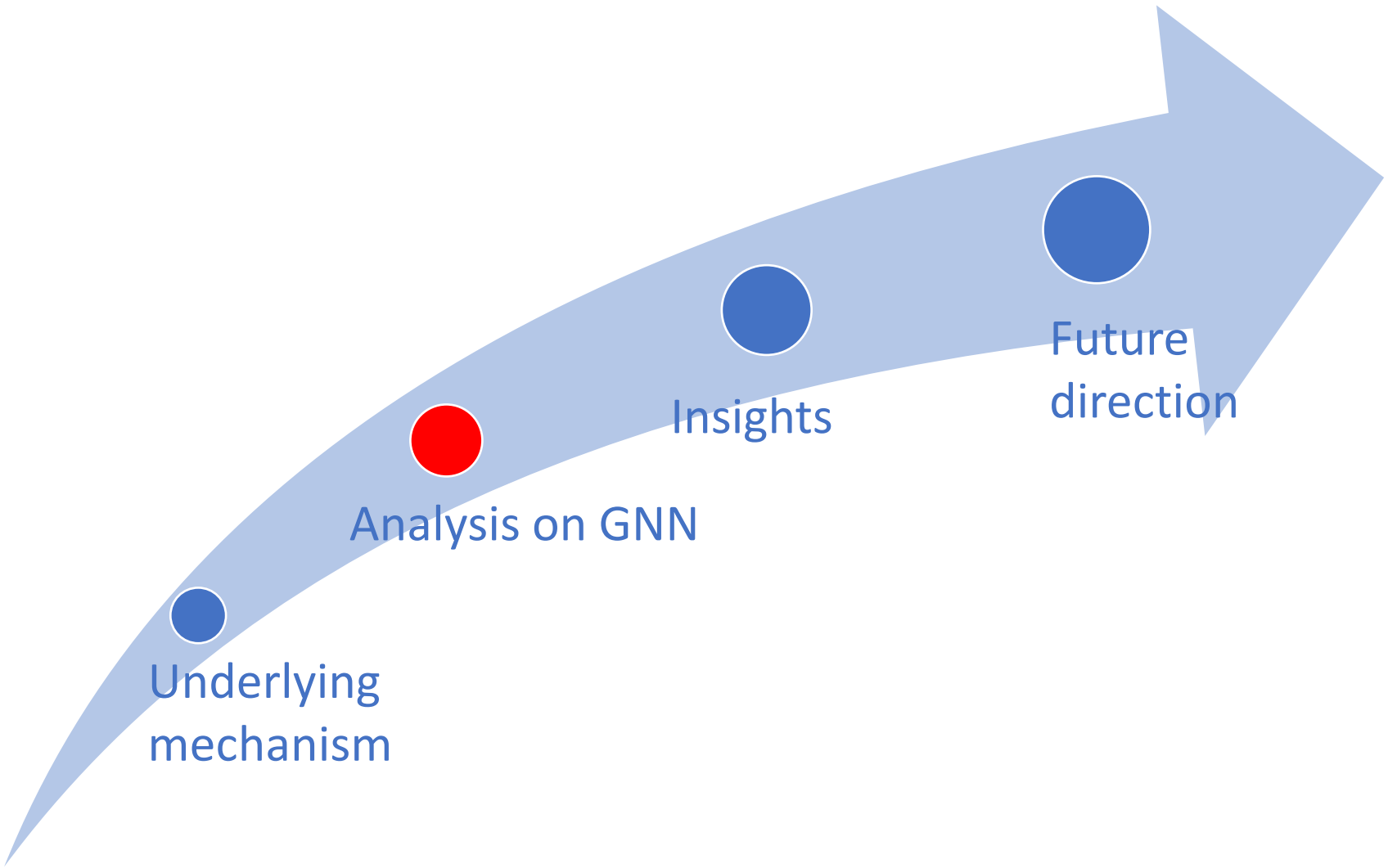
LSP

GSP

FP

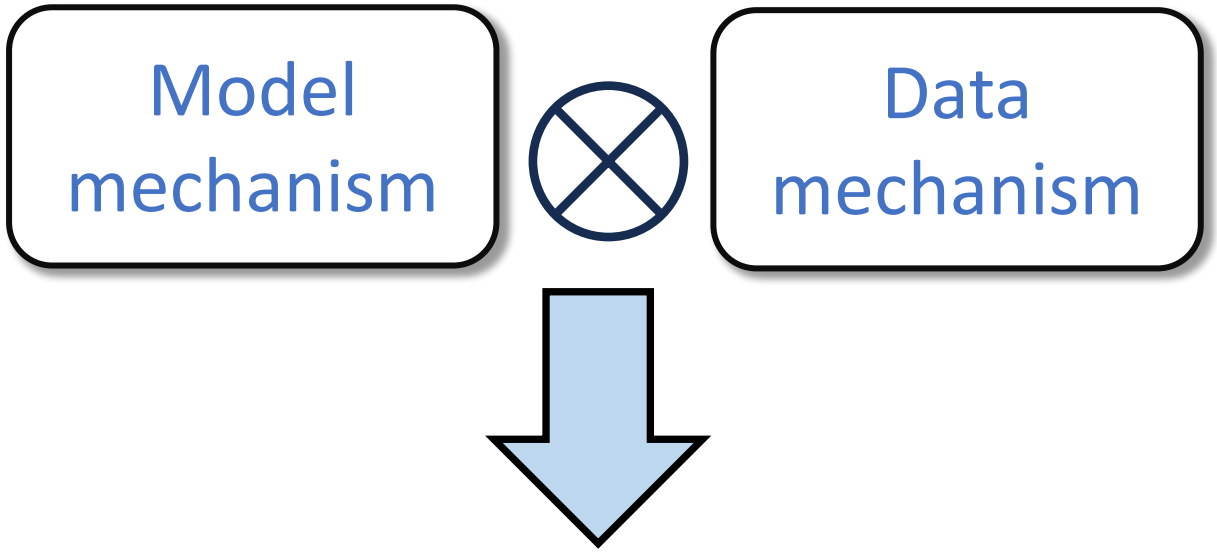
- Each data factor is essential
- Global structural proximity only shows effectiveness when local structural proximity is deficient.
- The incompatibility can be found between feature and structural proximity.

Outline


Underlying mechanism
Analysis on GNN
Insights
Future direction

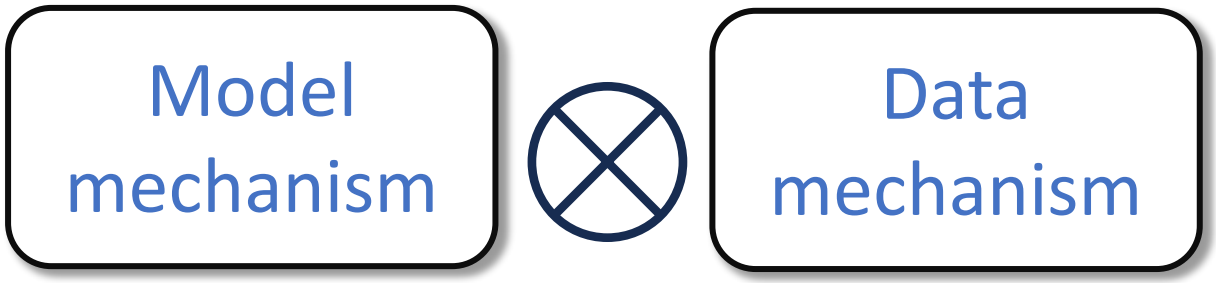
Analysis on GNNs

When do GNNs learn a good representation for different node pairs and when not?



How models behave with such incompatibility?

How models behave with such incompatibility?



GNN4LP is more capable on structural proximity

The incompatibility between feature and structure

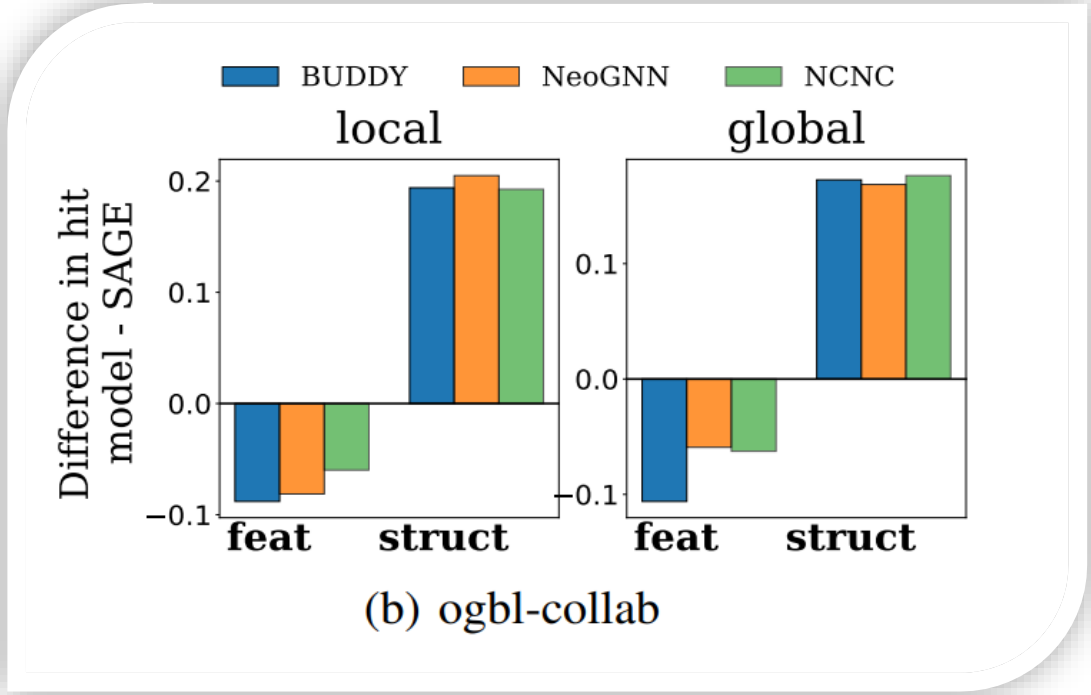
There is no free lunch with incompatibility!

Better structure modeling

Worse feature modeling

Empirical evidence

Performance comparison between GNN4LP models and GraphSAGE.



(b) ogbl-collab

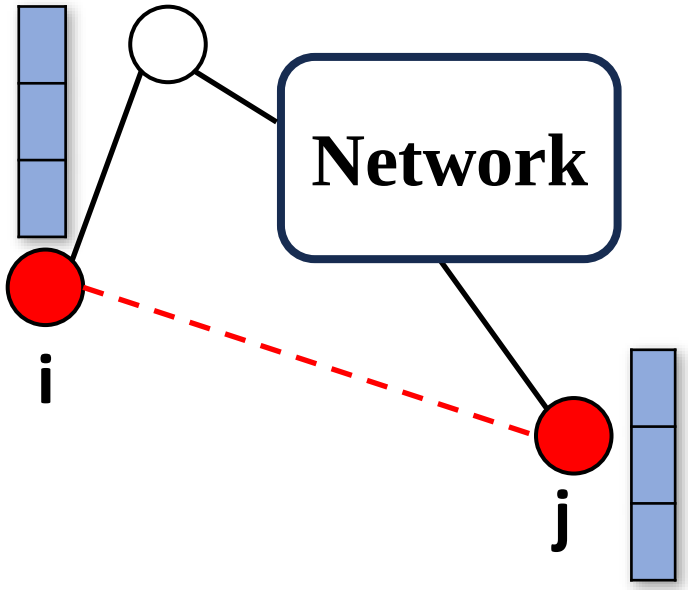
GNN4LP models outperform on node pairs with structure proximity, but fail on the feature ones

Broader impact

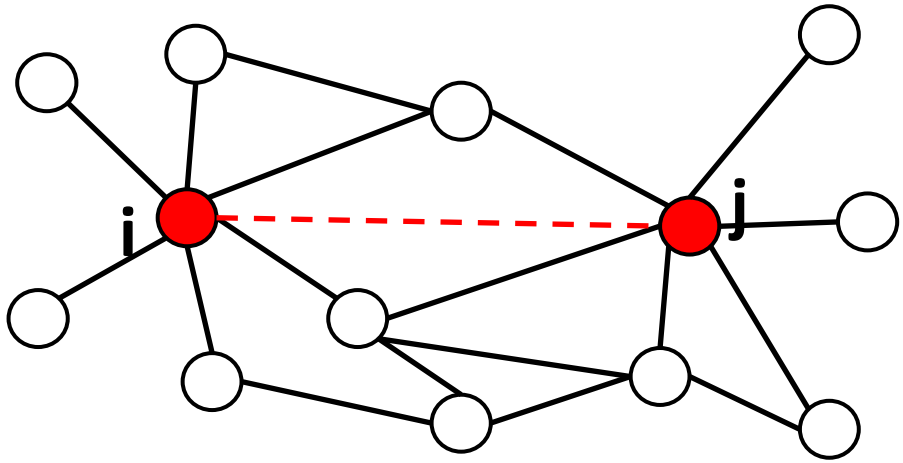
GNN4LP models outperform on node pairs with structure proximity, but fail on the feature ones

Leads to the essential fairness issue

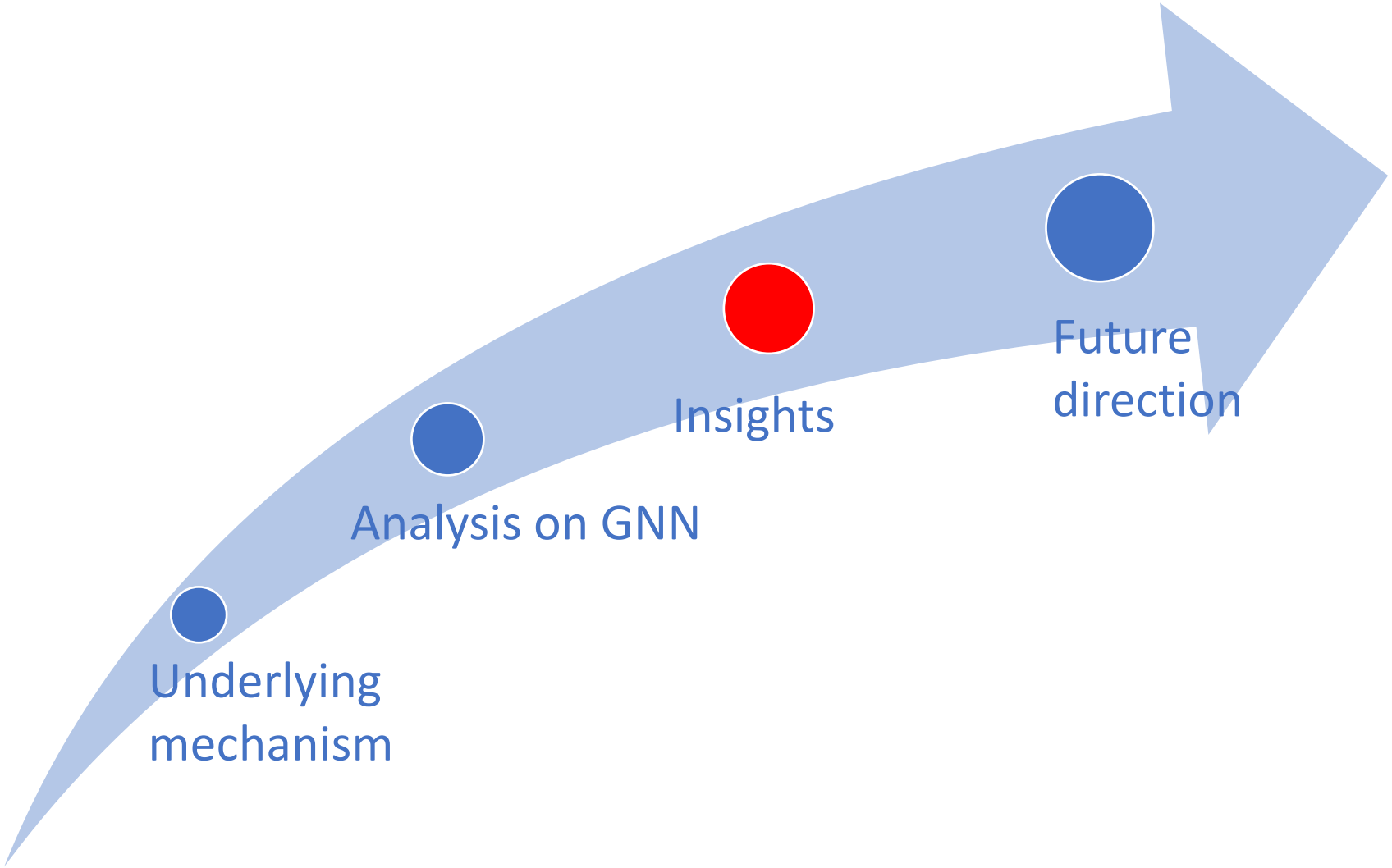
Two nervous guys



Two popular guys



Outline


Underlying mechanism
Analysis on GNN
Insights
Future direction

Insights for building foundation model

Multiple insights:

- What is transferable across different datasets in link prediction?
- What is the essential difficulty for the model design?
- Basics and instructions for building the Graph Foundation Model

Insights for building foundation model

Multiple insights:

- What is transferable across different datasets in link prediction?
- What is the essential difficulty for the model design?
- Basics and instructions for building the Graph Foundation Model

Insights for building foundation model

What is transferable across different datasets?

LSP

- Three data factors are key concepts across different datasets

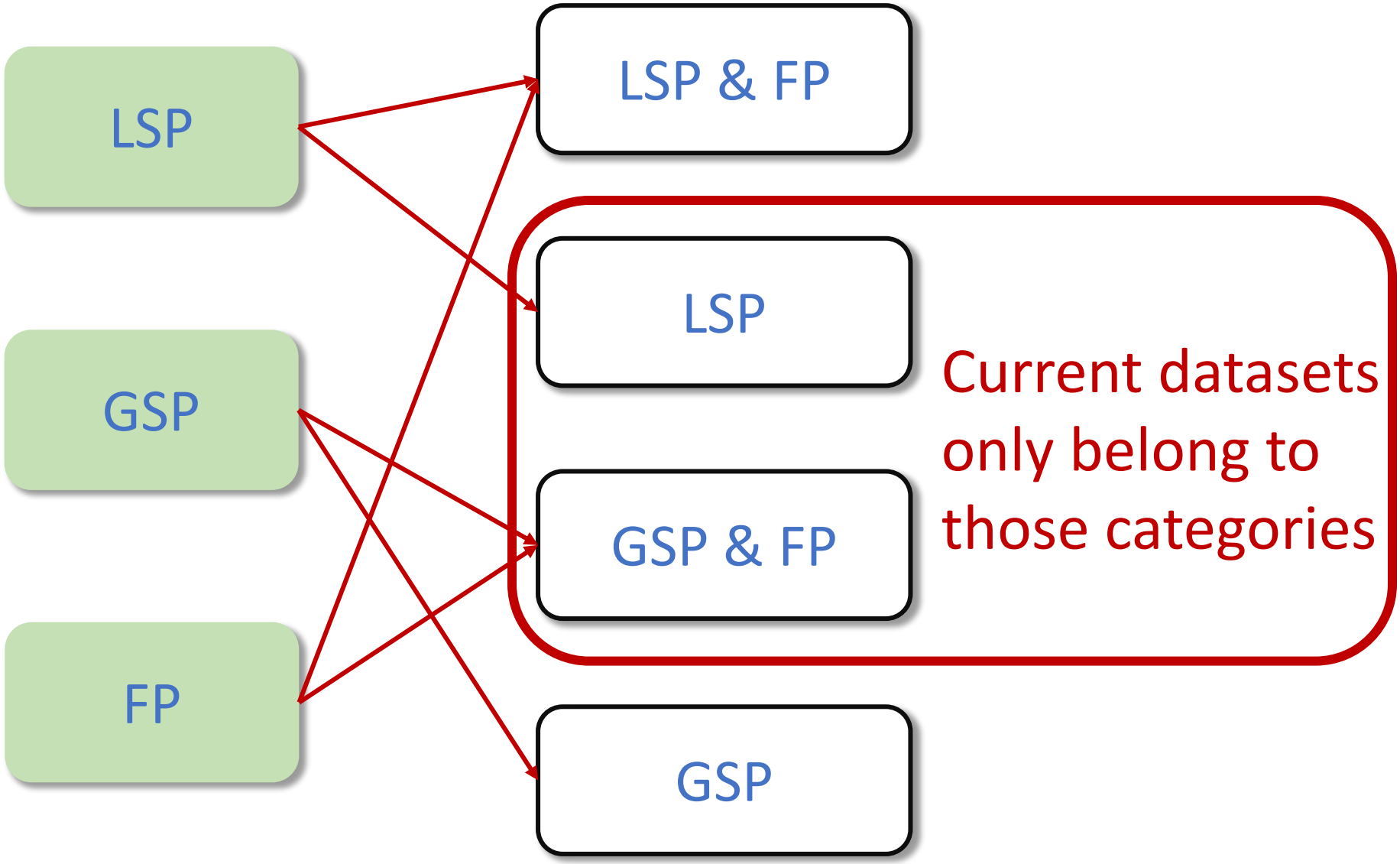
GSP

- Different factors combinations lead to different dataset properties

FP

- One rule: LSP and GSP factors cannot co-occur in one dataset

Dataset categories



New Benchmarking Dataset

GSP

LSP & FP

Performance on new datasets.

	CN	Katz	FH	SAGE	BUDDY
POWER	**12.88**	**29.85**	**NA**	**6.99**	**19.88**

The GSP heuristic (Katz) even outperforms the GNN4LP model (Buddy)

New Benchmarking Dataset

GSP

LSP & FP

Performance on new datasets.

	CN	Katz	FH	SAGE	BUDDY
POWER	12.88	29.85	NA	6.99	19.88
PHOTO	18.34	7.07	13.78	18.61	18.09

The vanilla GNN (SAGE) even outperforms the GNN4LP model (Buddy)

Insights for building foundation model

Multiple insights:

- What is transferable across different datasets in link prediction?
- What is the essential difficulty for the model design?
- Basics and instructions for building the Graph Foundation Model

Instruction for building GFM

- Three data factors, LSP, GSP, and FP, are key concepts across different datasets
- Current benchmark are not sufficient for building GFM as they lose certain combination
- We should have a better dataset set covering all categories to build the Graph Foundation Model

Instruction for building GFM

More datasets are collected with diverse categories and domains

We provide additional 21 datasets for link prediction

- Biology
- Transportation
- Web
- Academia
- Social science

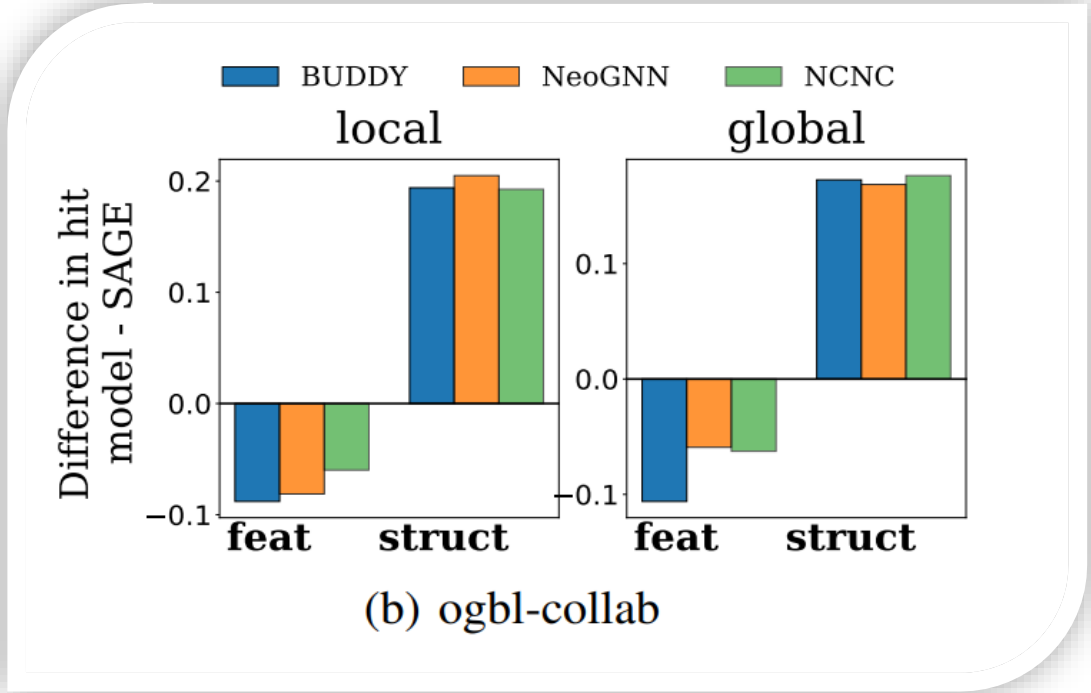
Insights for building foundation model

Multiple insights:

- What is transferable across different datasets in link prediction?
- What is the essential difficulty for the model design?
- Basics and instructions for building the Graph Foundation Model

Incompatibility between feature and structure

Performance comparison between GNN4LP models and GraphSAGE.



(b) ogbl-collab

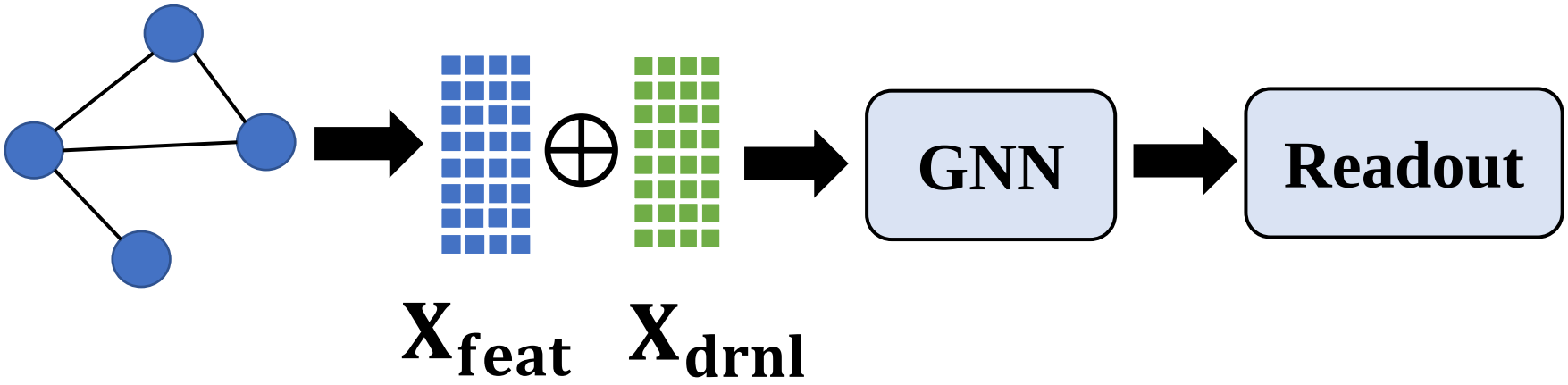
GNN4LP models outperform on node pairs with structure proximity, but fail on the feature ones

Insights for building foundation model

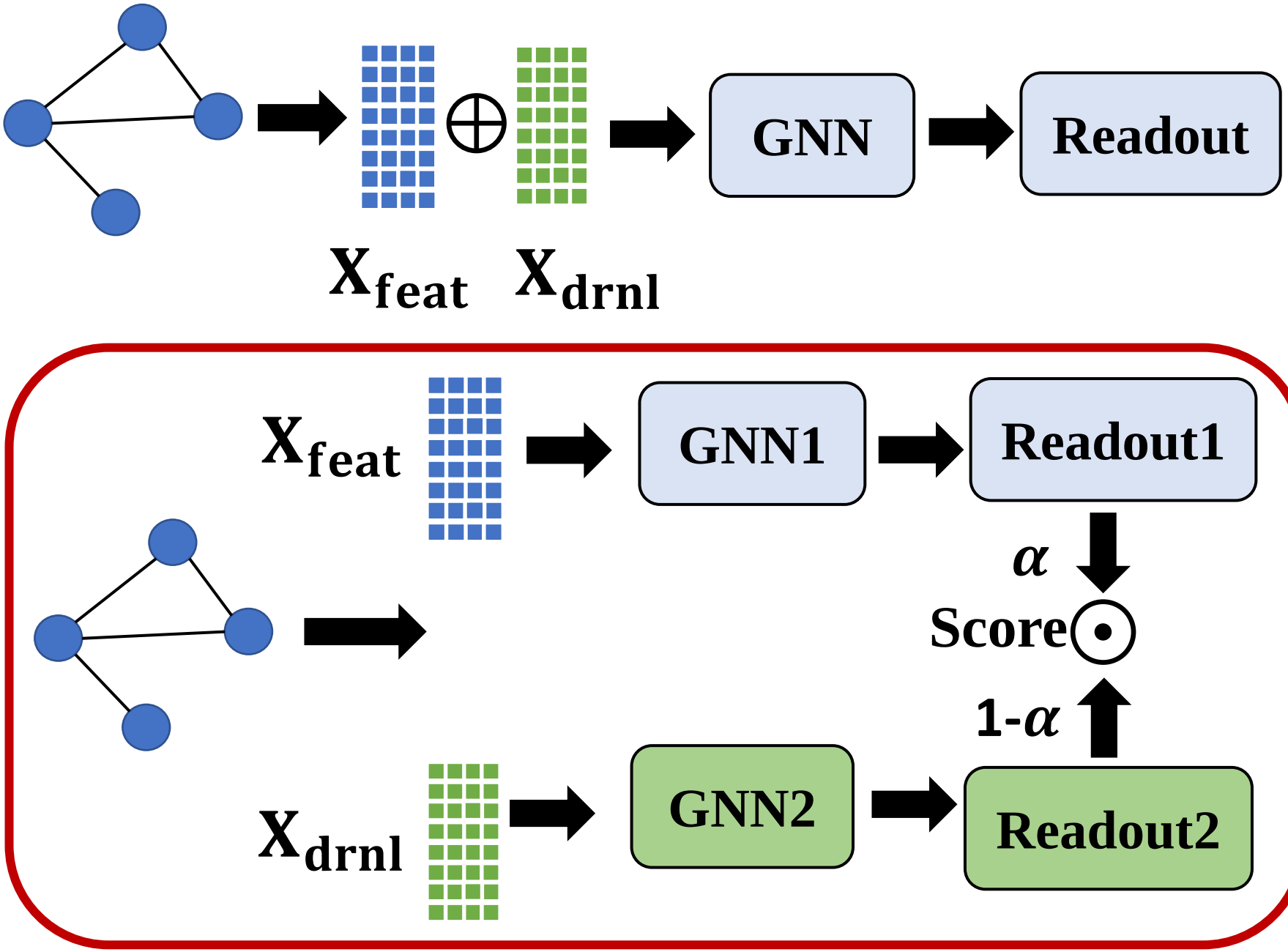
Multiple insights:

- What is transferable across different datasets in link prediction?
- What is the essential difficulty for the model design?
- Basics and instructions for building the Graph Foundation Model

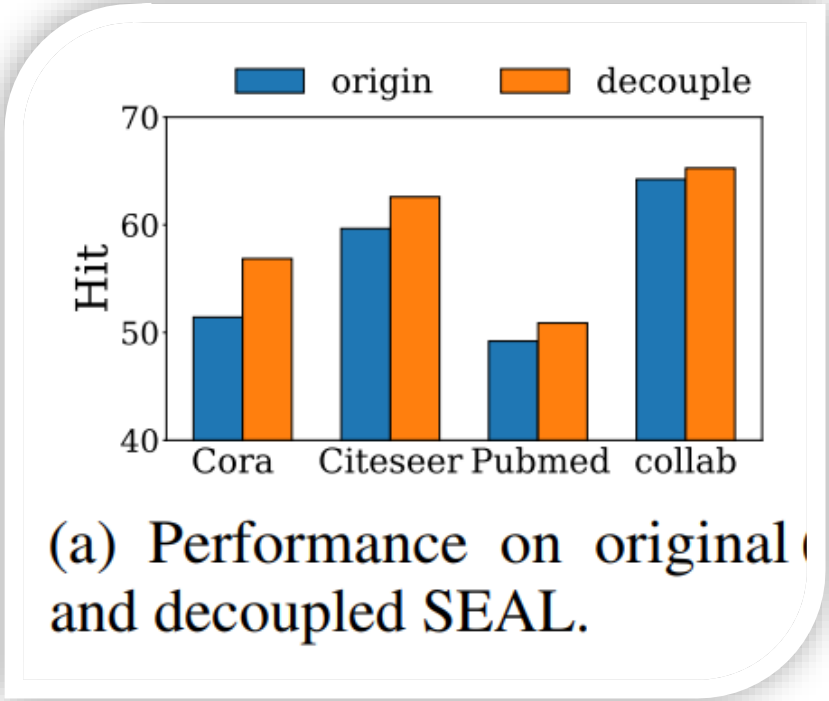
Avoid modeling feature and structural proximity simultaneously



Combine feature and structure into one GNN


GNN
Readout
X_{feat}
X_{drnl}
X_{feat}
GNN1
Readout1
α
Score
1-α
X_{drnl}
GNN2
Readout2

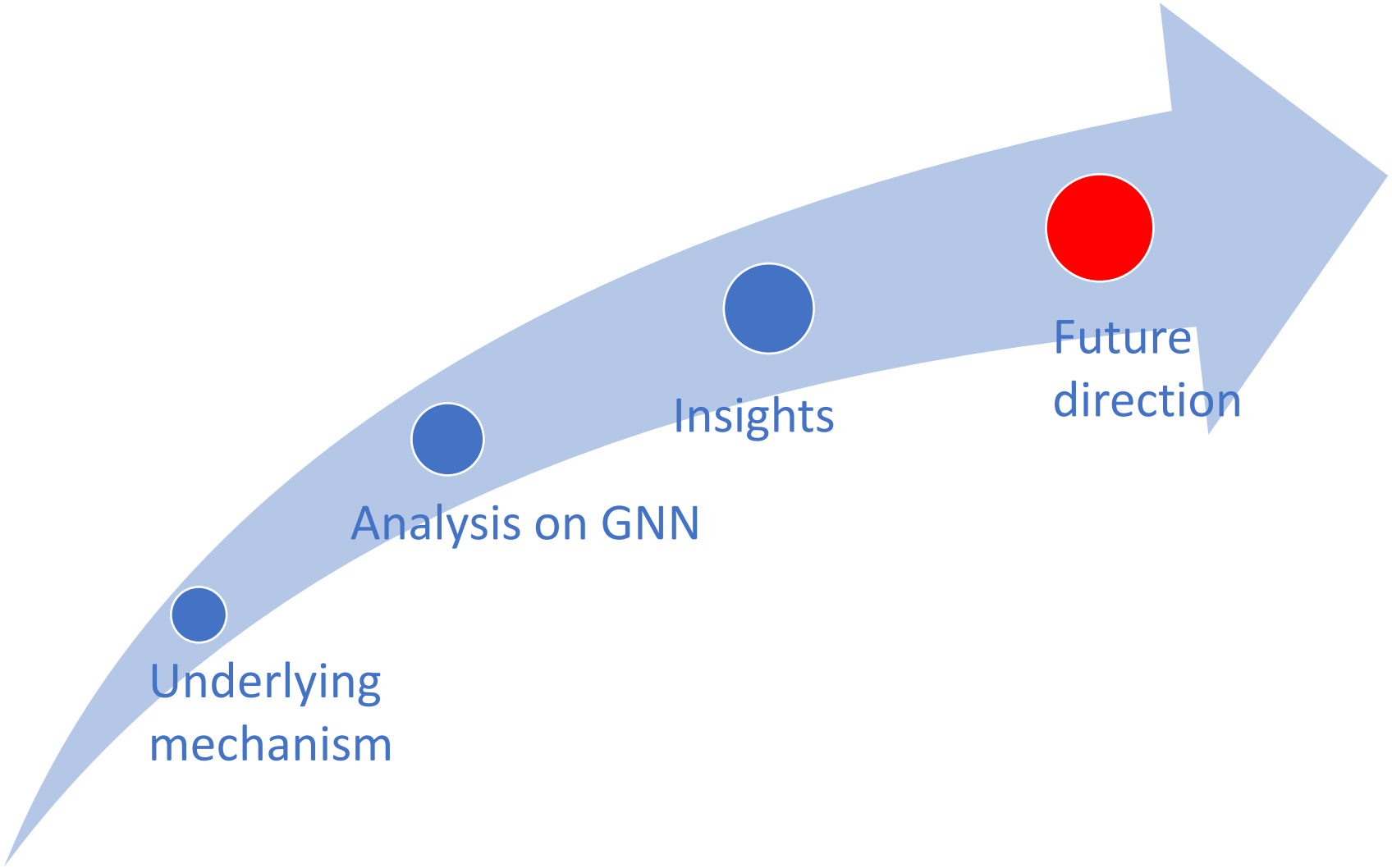
Empirical evidence



(a) Performance on original and decoupled SEAL.

Decoupling model leads to the performance gain

Outline


Underlying mechanism
Analysis on GNN
Insights
Future direction

Main takeaways

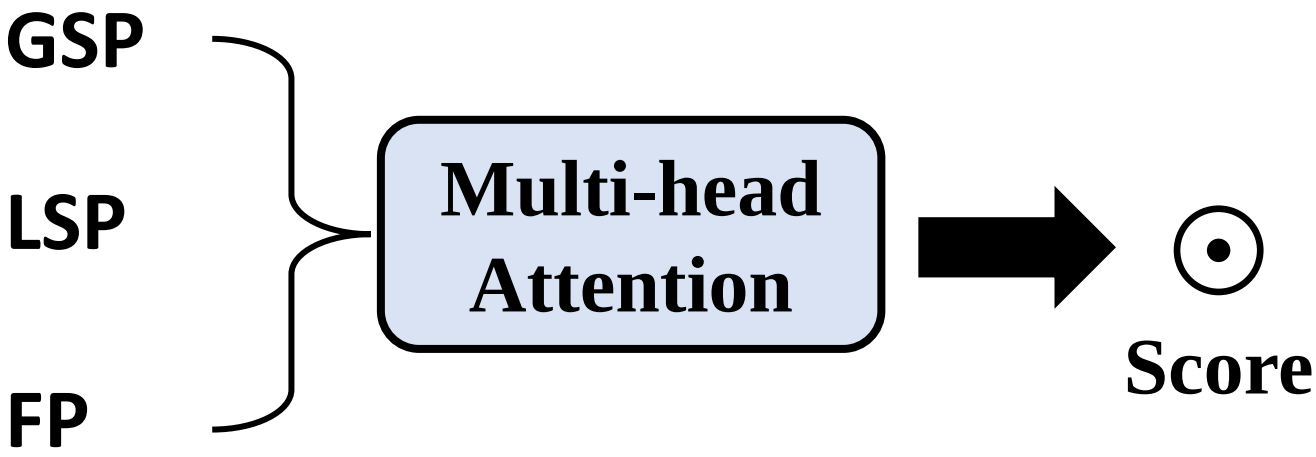
LSP, GSP, FP are three data factors in Link Prediction

Use more datasets covering all categories for train and test

Carefully handle incompatibility between FP and LSP

What are we doing now?

A more fair model



Paper

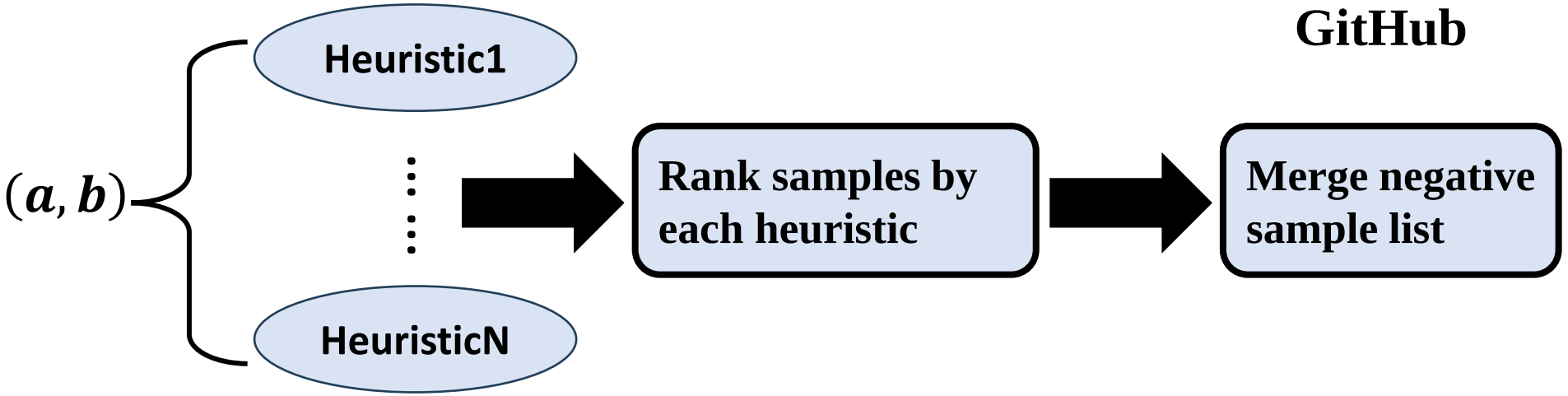
Utilize transformer to capture different data factors together

What are we doing now?

Link Prediction benchmark

- Provide new fixed split
- Benchmarking results (Better baseline tunning)
- Guidance: No AUC & AP for evaluation
- New hard evaluation setting

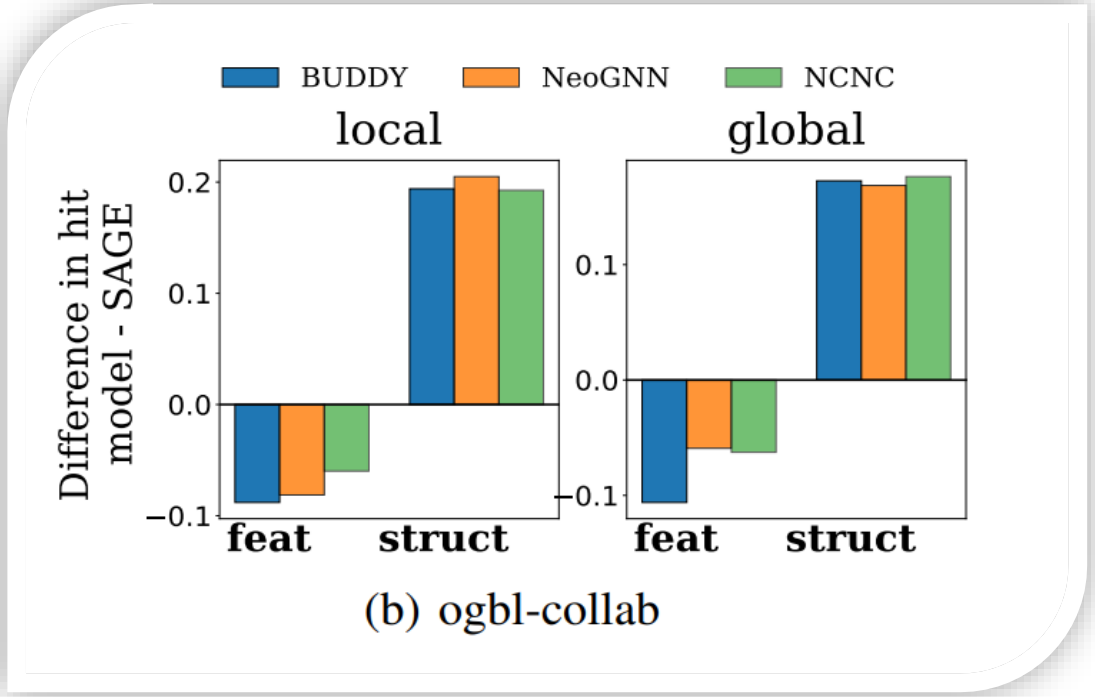
GitHub



A pause & QA!

Empirical evidence

GNN4LP models outperform on the node pairs with structural proximity, but fails on the feature ones



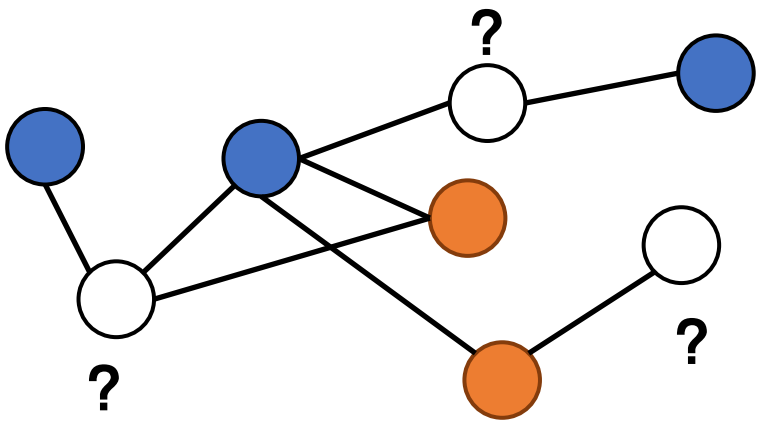
How about the node classification?

Demystifying Structural Disparity in Graph Neural Networks: Can one size Fit ALL?

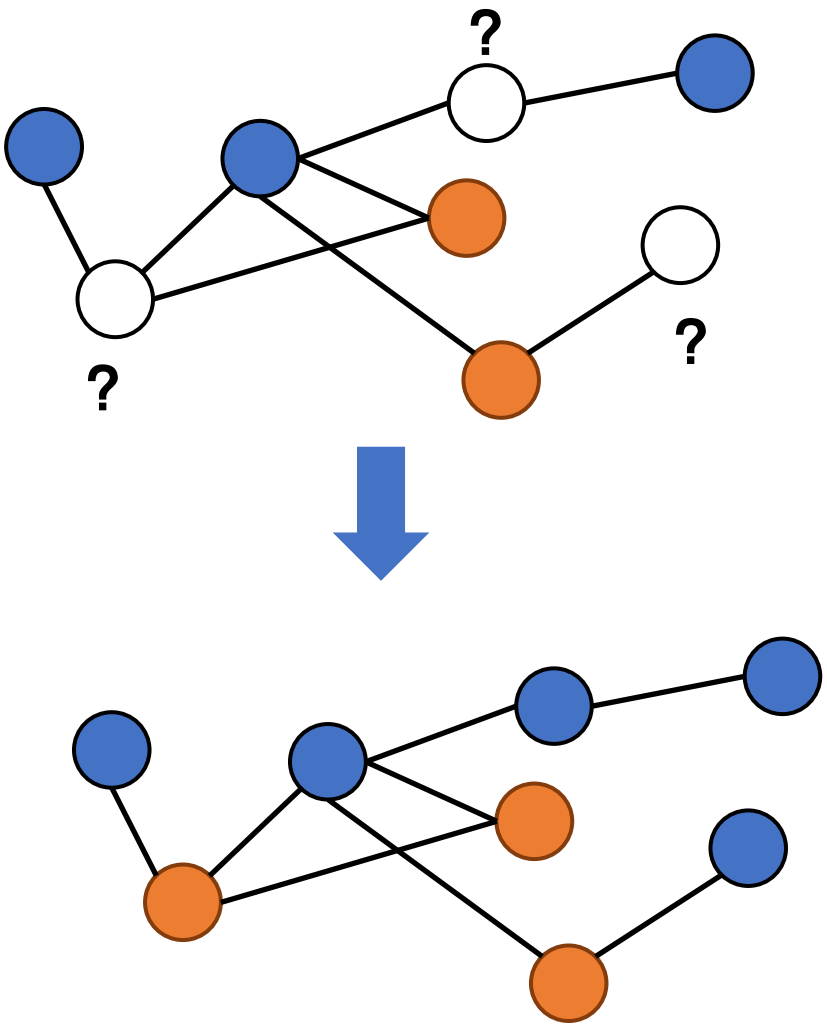
Haitao Mao, Zhikai Chen, Wei Jin, Haoyu Han
Yao Ma, Tong Zhao, Neil Shah, Jiliang Tang

MICHIGAN STATE
UNIVERSITY

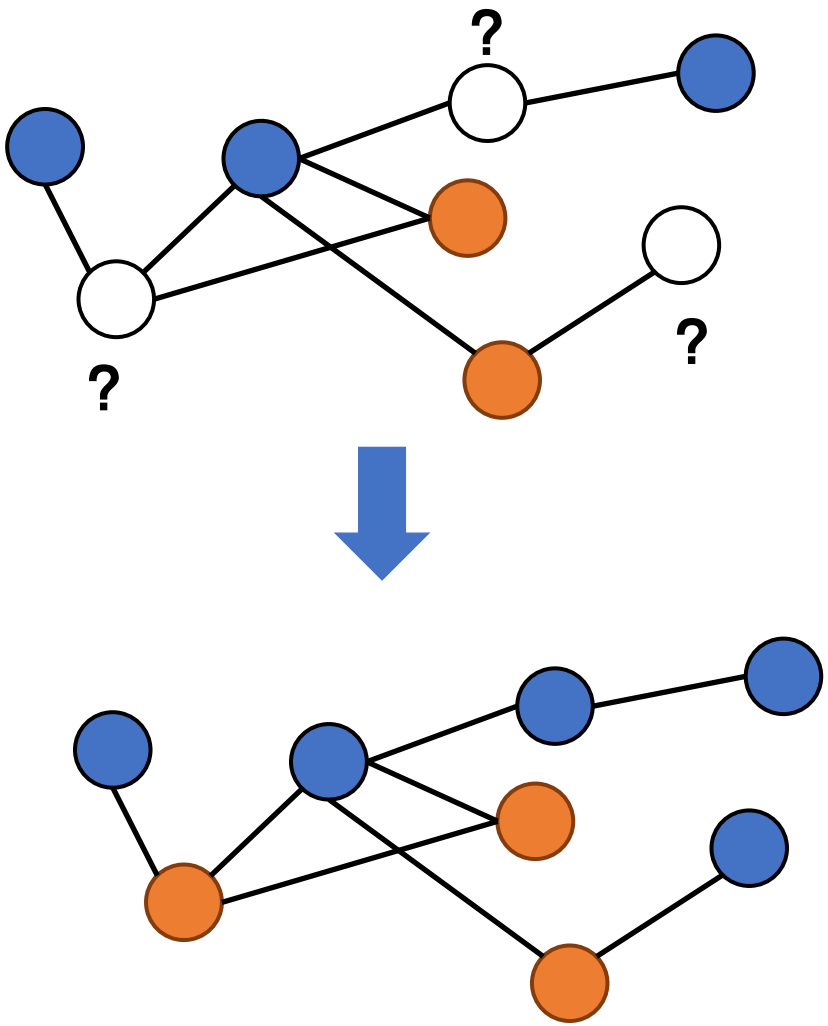
Node Classification task



Node Classification task



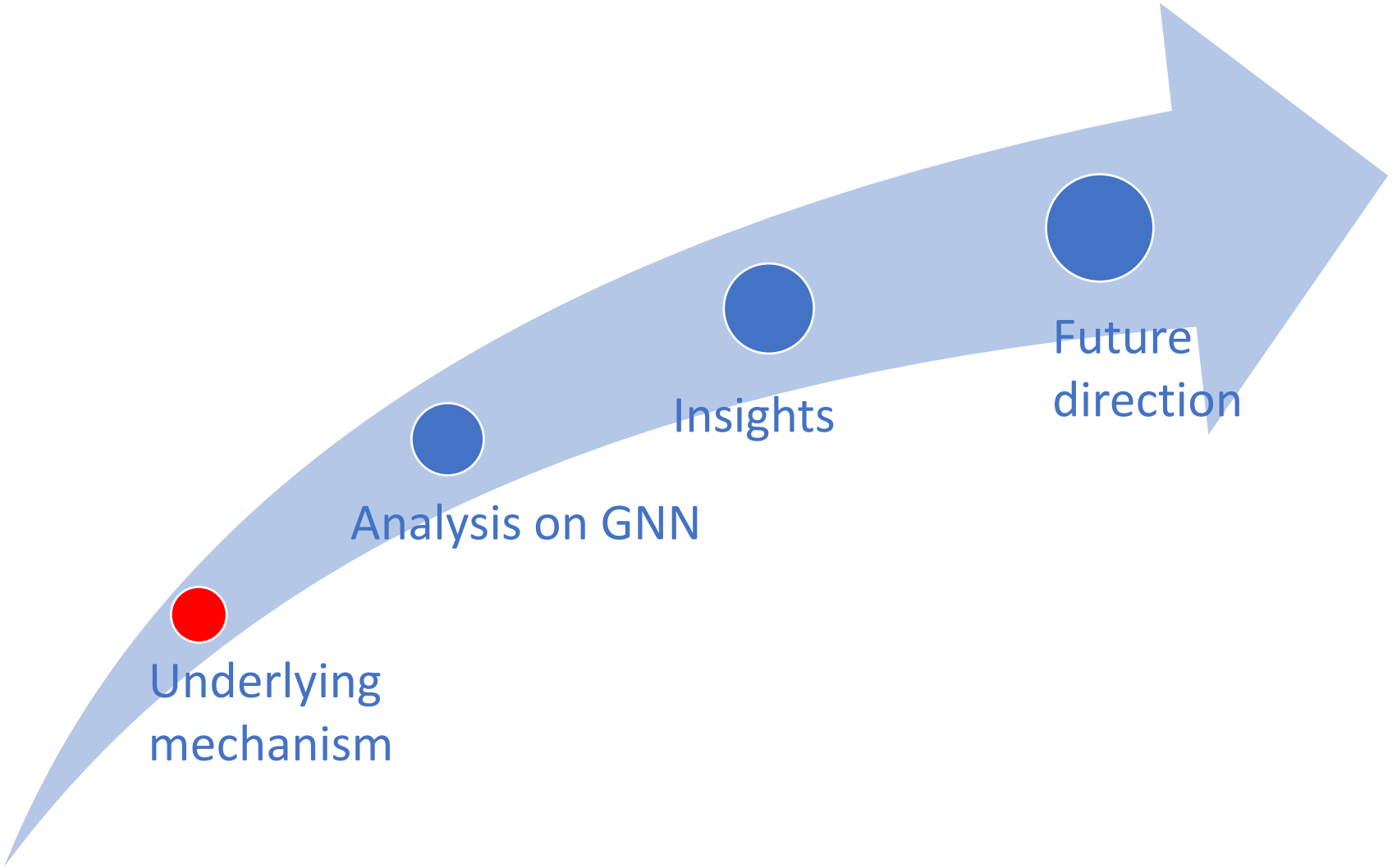
Node Classification task



- Inferring Node Attributes
- Social Influence Prediction
- Traffic Prediction
- Air Quality Prediction

⋮

Outline


Underlying mechanism
Analysis on GNN
Insights
Future direction

Data mechanism

When do GNNs perform well on different nodes and when not?

Model mechanism

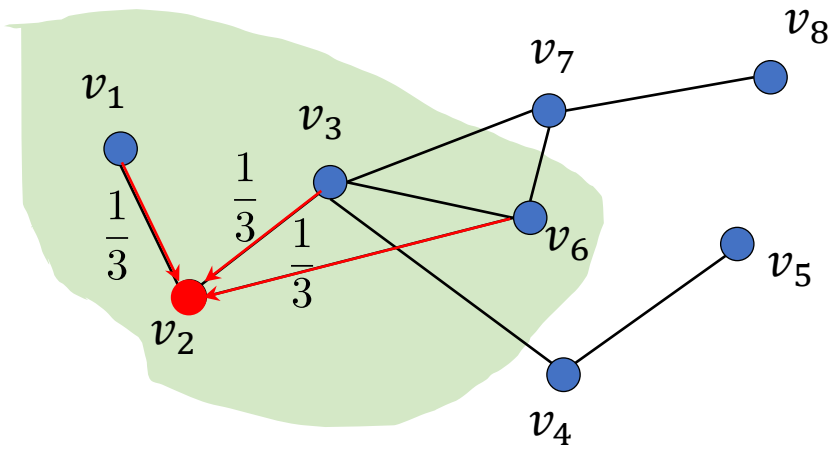
Data mechanism

Graph Neural Networks

Neighbors of node v_2

$$\mathcal{N}(2)$$

$$\{v_1, v_3, v_6\}$$

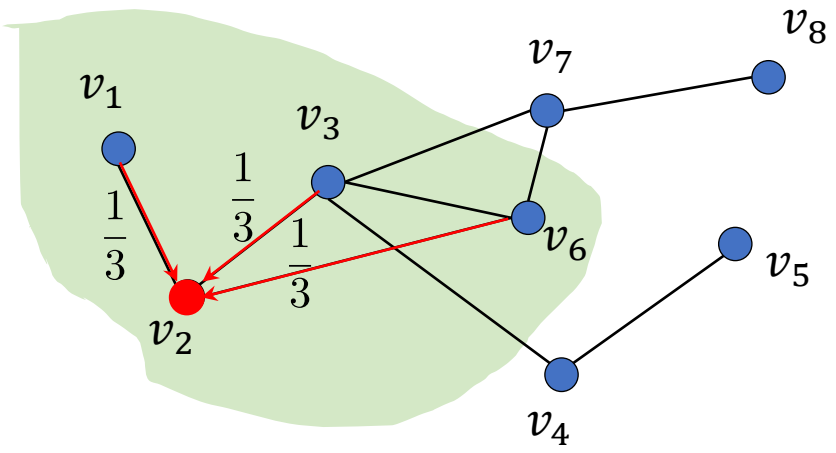


Graph Neural Networks

Neighbors of node v_2

$\mathcal{N}(2)$

$\{v_1, v_3, v_6\}$



Feature Transformation

GNN:

$$\mathbf{F}_i = \sum_{j \in \mathcal{N}(i)} \frac{1}{|\mathcal{N}(i)|} \mathbf{X}_j \mathbf{W}$$

$$\mathbf{X}'_j = \mathbf{X}_j \mathbf{W}$$

Neighborhood Aggregation

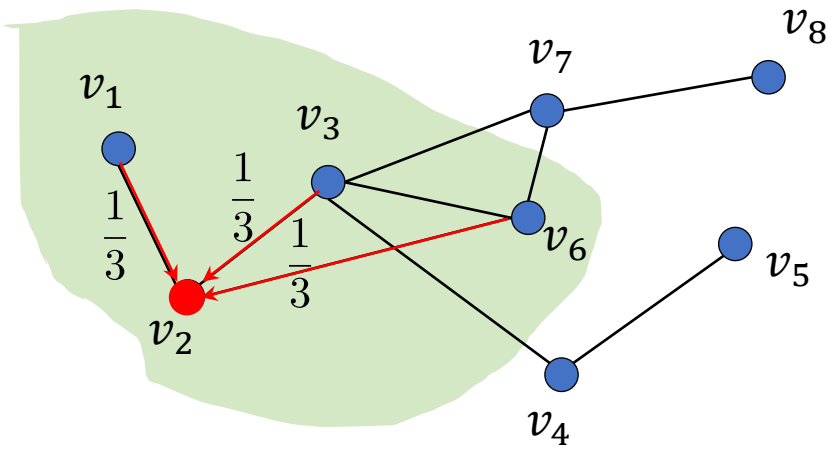
11
0

Graph Neural Networks

Neighbors of node v_2

$$\mathcal{N}(2)$$

$$\{v_1, v_3, v_6\}$$



GNN:

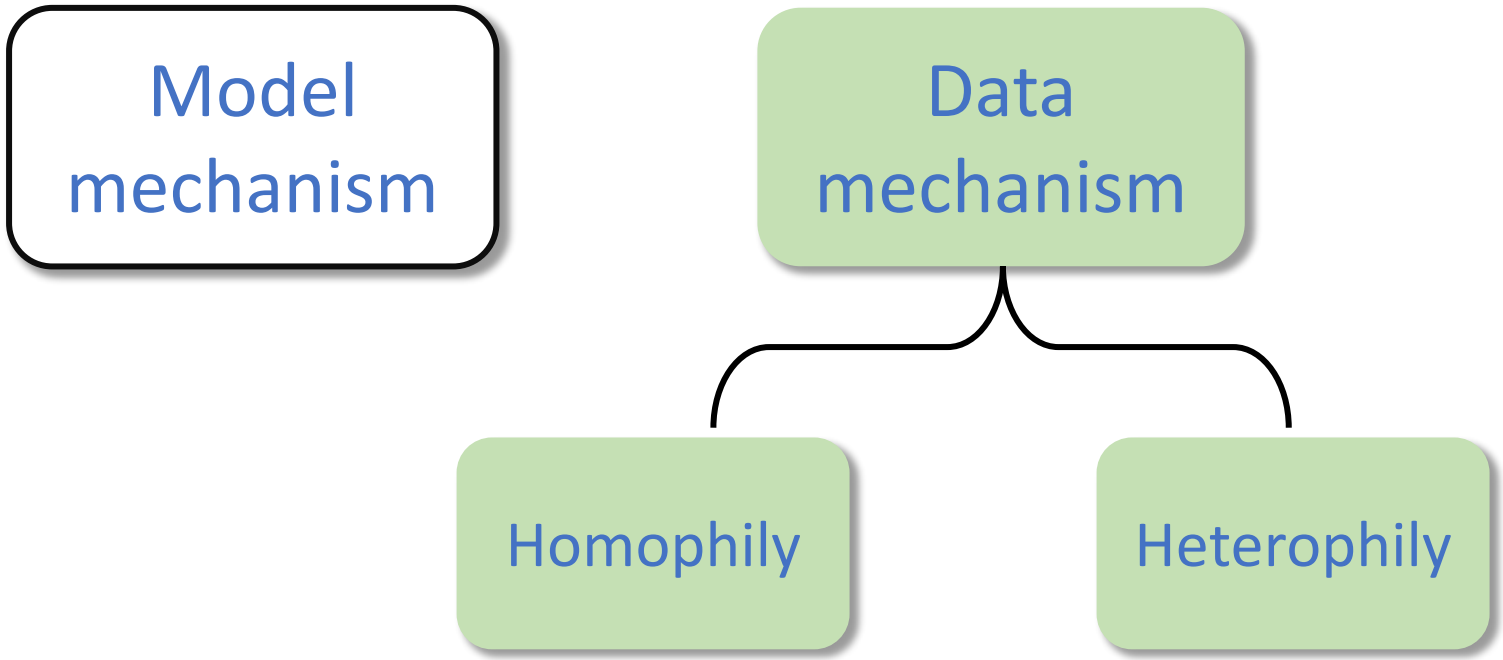
$$\mathbf{F}_i = \sum_{j \in \mathcal{N}(i)} \frac{1}{|\mathcal{N}(i)|} \mathbf{X}'_j$$

Majorly focusing on Aggregation

Neighborhood Aggregation

Data mechanism

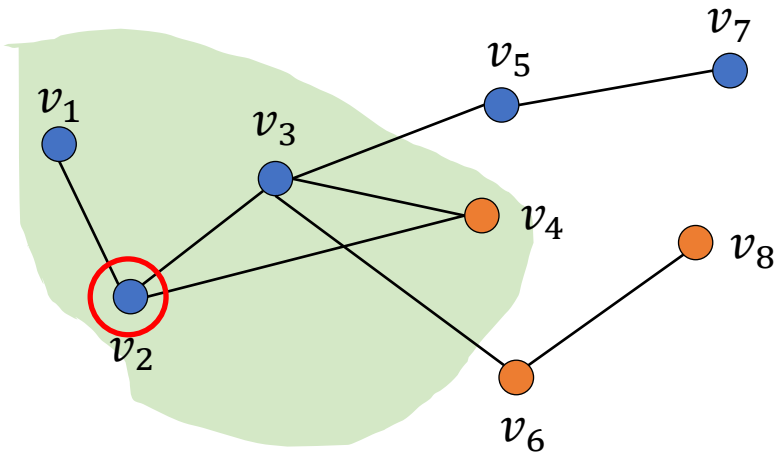
When do GNNs perform well on different nodes and when not?



Homophily & Heterophily

Homophily: "nodes tend to connect with "similar" or "alike" others"

$$\mathrm{h}_i = \frac{|\{u \in \mathcal{N}(v_i): y_u = y_v\}|}{|\mathcal{N}(v_i)|}$$



Center node: v_2

Neighbor set: v_1 v_3 v_4

Node homophily ratio of v_2:

$$h_2 = \frac{2}{3}$$

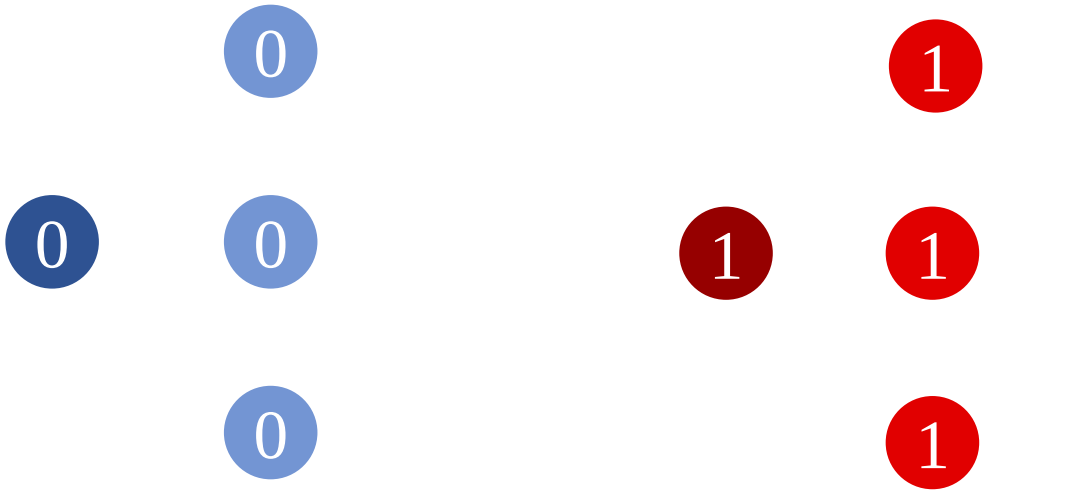
Data assumption for homophily & heterophily

CSBM(μ_0, μ_1, p, q)

Feature generation

$x_i \sim \mathcal{N}(\mu_0, I)$ For c_j = 0

$x_j \sim \mathcal{N}(\mu_1, I)$ For c_j = 1



Data assumption for homophily & heterophily

CSBM(μ_0, μ_1, p, q)

Edge generation

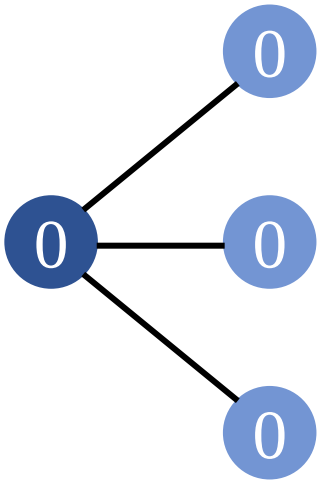
$e_{ij} \sim B(1, p)$ when $c_i = c_j$

$e_{ij} \sim B(1, q)$ when $c_i \neq c_j$

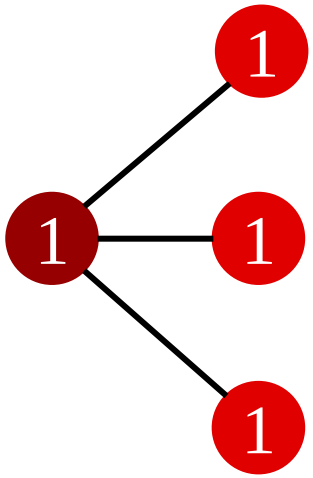
Intra-class probability: $p = 0.8$

Inter-class probability: $q = 0$

Nodes in class 0



Nodes in class 1



A homophily case

Data assumption for homophily & heterophily

CSBM(μ_0, μ_1, p, q)

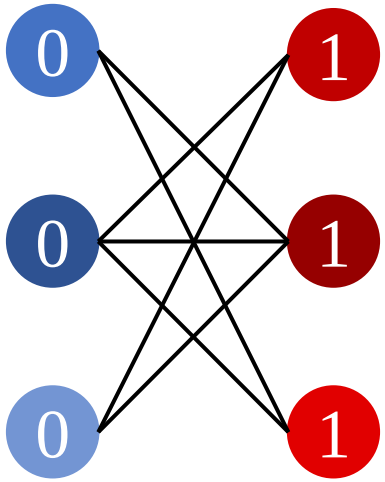
Edge generation

$e_{ij} \sim B(1, p)$ when $c_i = c_j$

$e_{ij} \sim B(1, q)$ when $c_i \neq c_j$

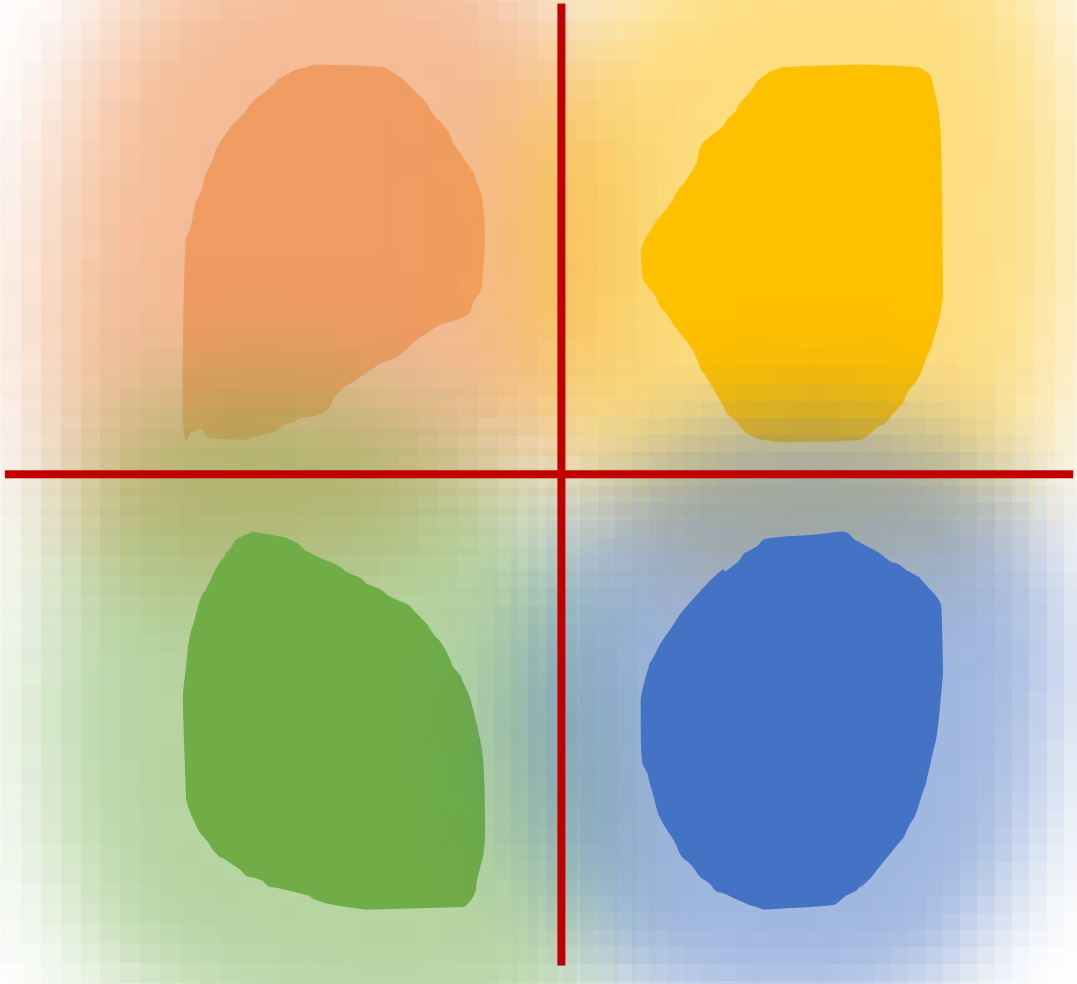
Intra-class probability: $p = 0$

Inter-class probability: $q = 0.8$



A heterophily case

What is a good node representation?



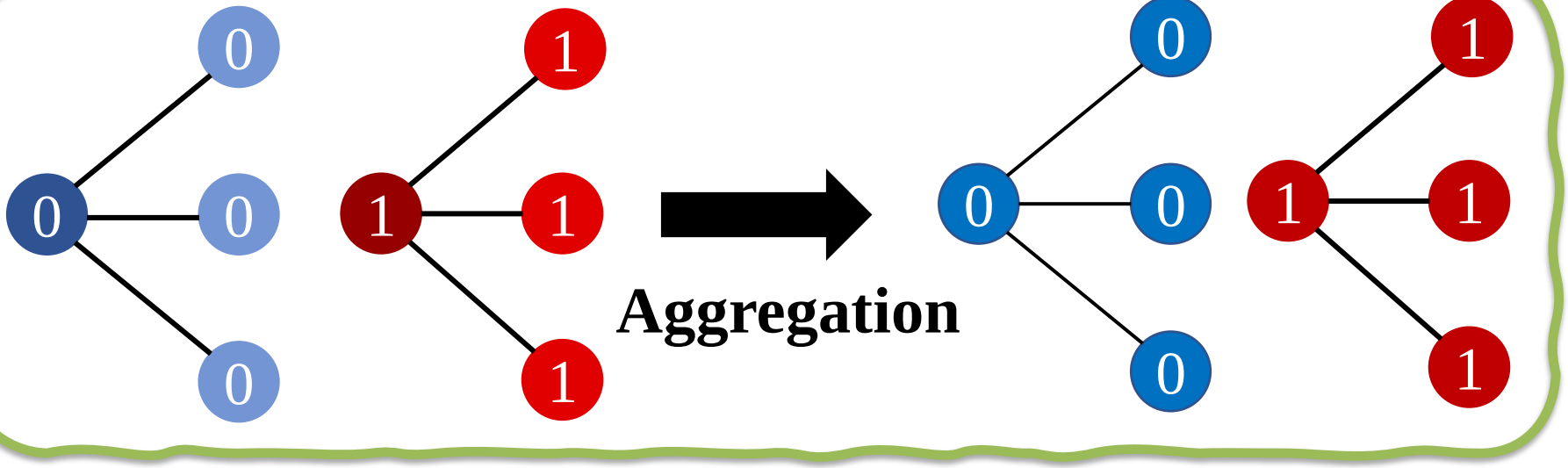
Cohesion:
Intra-class similar

Separation:
Inter-class dissimilar

How can GNN work well on homophily?

The Homophily Case

$$\mathbf{F}_i = \sum_{j \in \mathcal{N}(i)} \frac{1}{|\mathcal{N}(i)|} \mathbf{X}_j$$

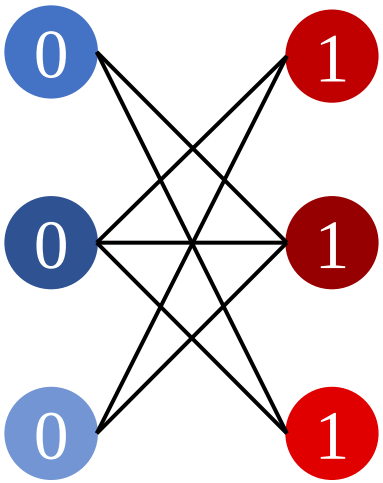


Good feature separability can be observed after aggregation

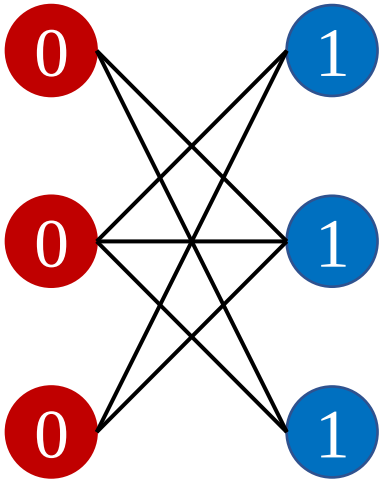
How can GNN work well on heterophily?

The Heterophily Case

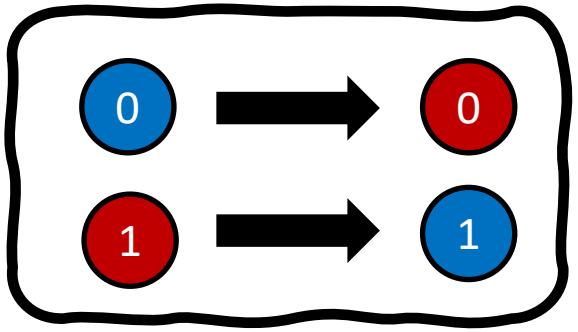
$$\mathbf{F}_i = \sum_{j \in \mathcal{N}(i)} \frac{1}{|\mathcal{N}(i)|} \mathbf{X}_j$$



Aggregation



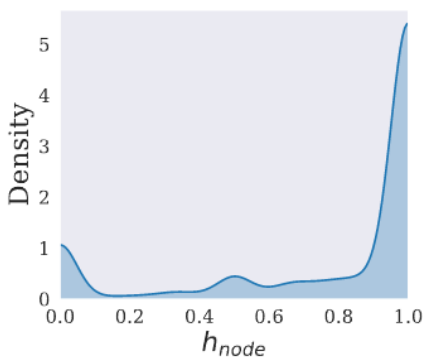
Good separability can still be observed despite alteration



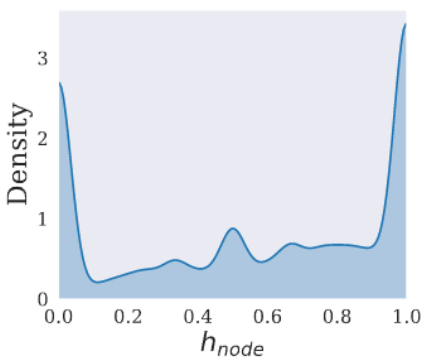
Perfect Separation

Misalign with real-world scenario

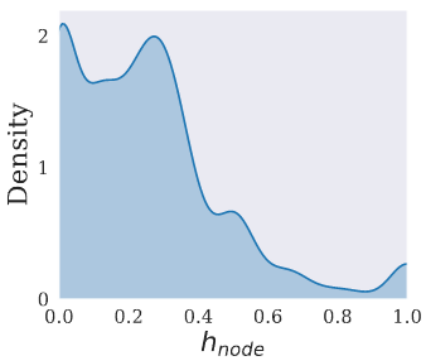
Node homophily ratio distribution



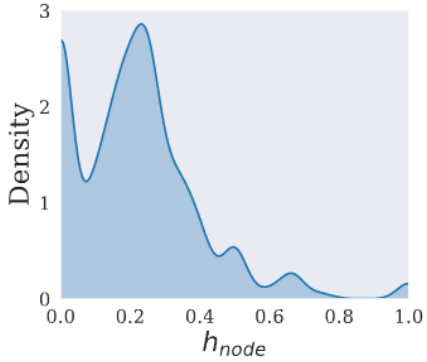
(a) PubMed (h=0.79)



(b) Ogbn-arxiv (h=0.63)



(c) Chameleon (h=0.22)



(d) Squirrel (h=0.25)

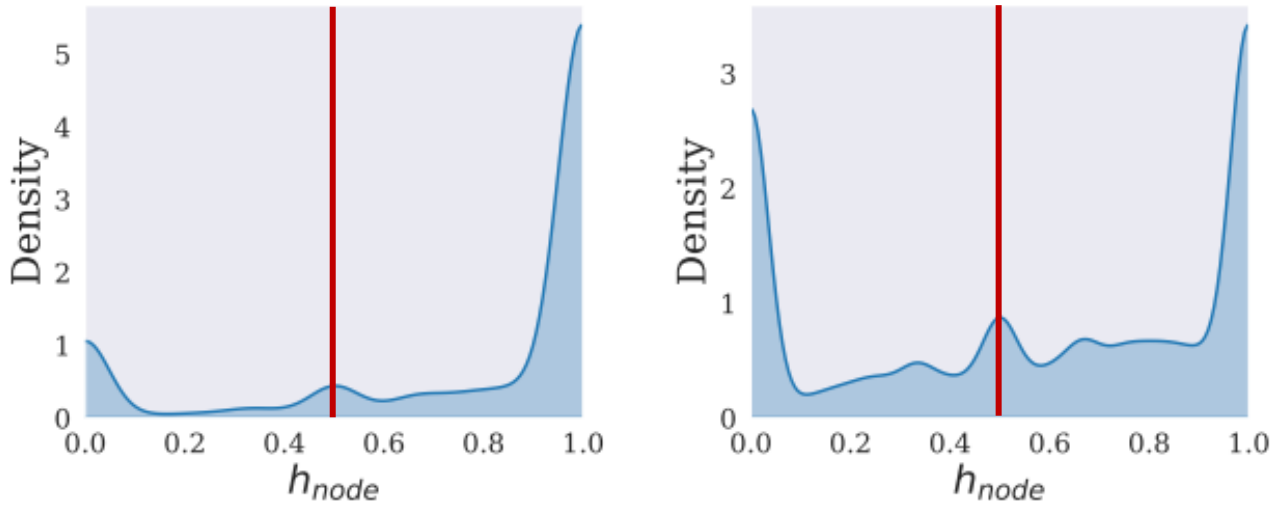
Both homophily and heterophily nodes appears across all real-world graphs

We can not consider homophily or heterophily solely!

Structure disparity often happens in real world scenario!

Can one GNN fit all nodes?

Node homophily ratio distribution



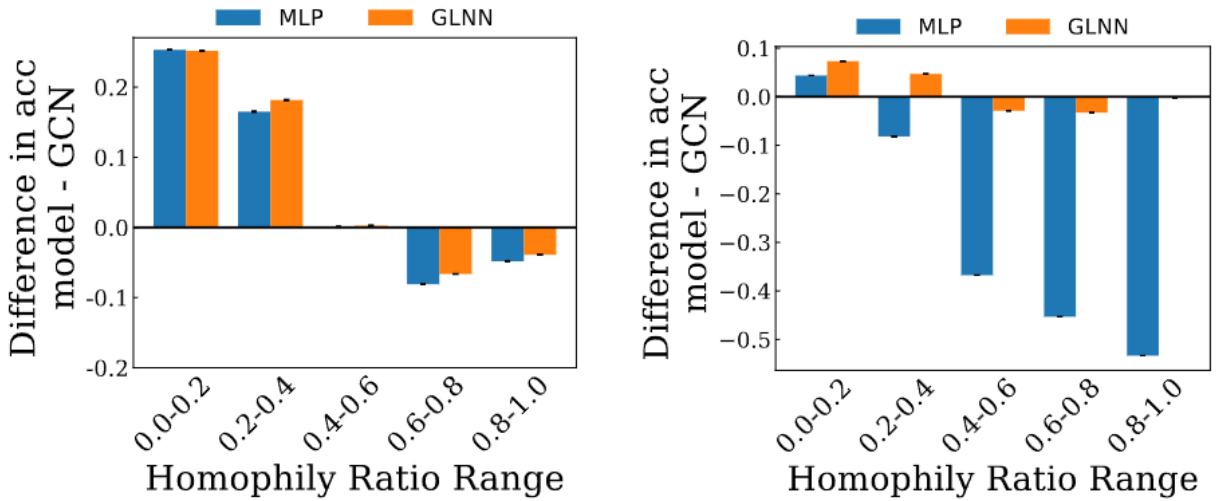
(a) PubMed (h=0.79) (b) Ogbn-arxiv (h=0.63)

Majority pattern : Homophily nodes in a homophily graph

Minority pattern : Heterophily nodes in a homophily graph

Can one GNN fits all nodes?

Performance comparison between GCN and MLP-based models

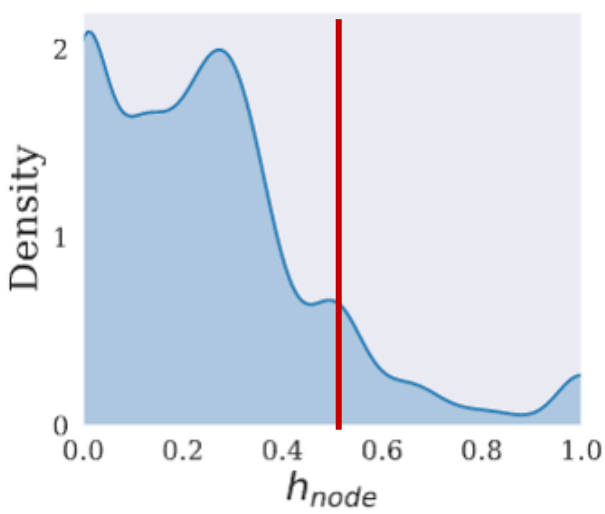


(a) PubMed (h=0.79) (b) Ogbn-arxiv (h=0.63)

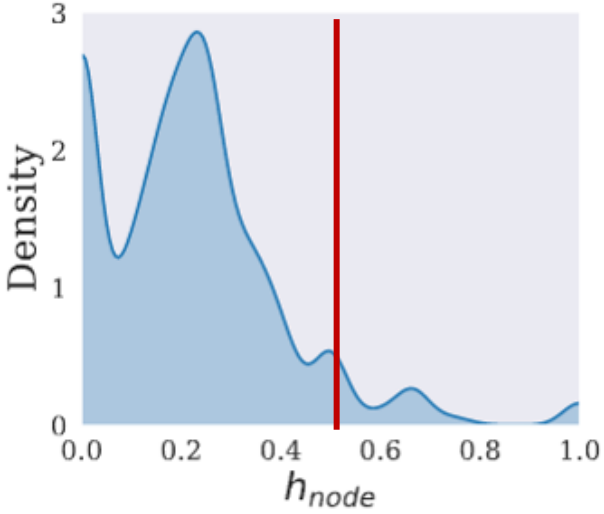
GCN outperforms on the majority pattern, but fails in the minor pattern

Can one GNN fit all nodes?

Node homophily ratio distribution



(c) Chameleon (h=0.22)



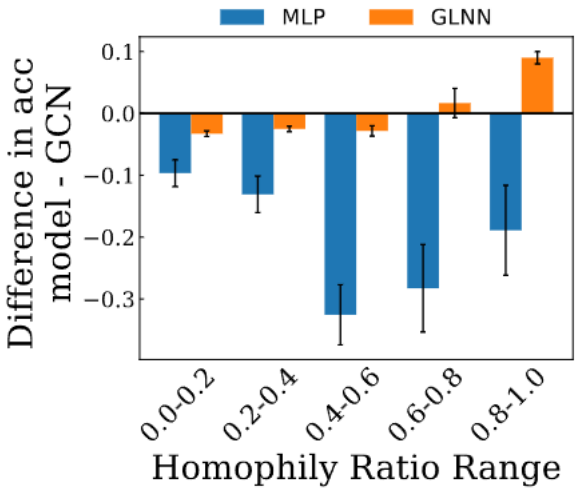
(d) Squirrel (h=0.25)

Majority pattern : Heterophily nodes in a heterophily graph

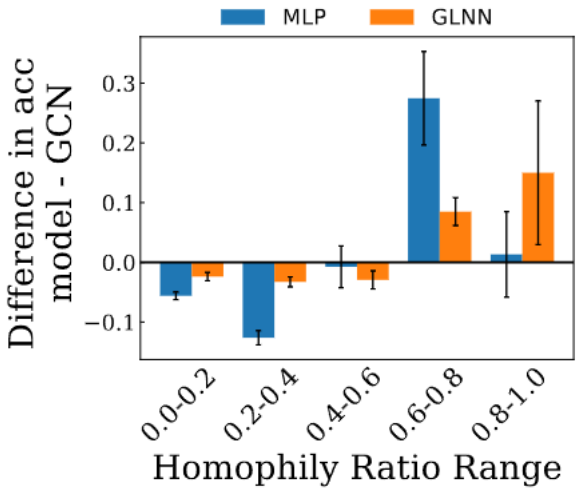
Minority pattern : Homophily nodes in a heterophily graph

Can one GNN fits all nodes?

Performance comparison between GCN and MLP-based models



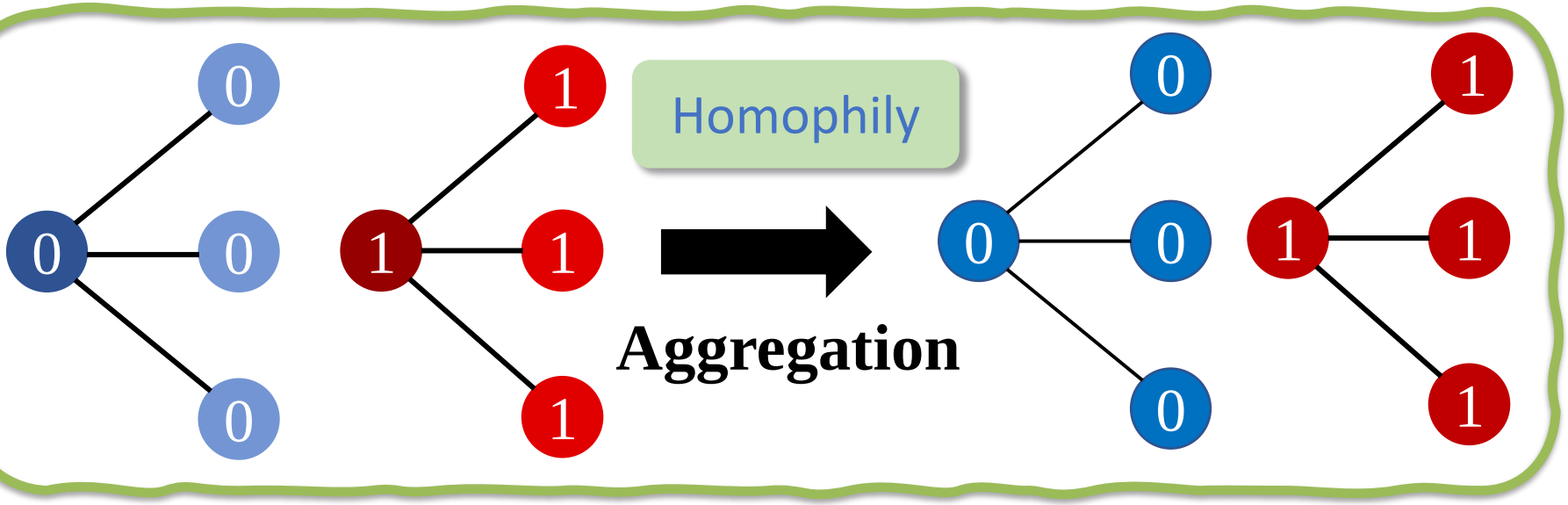
(c) Chameleon (h=0.22)

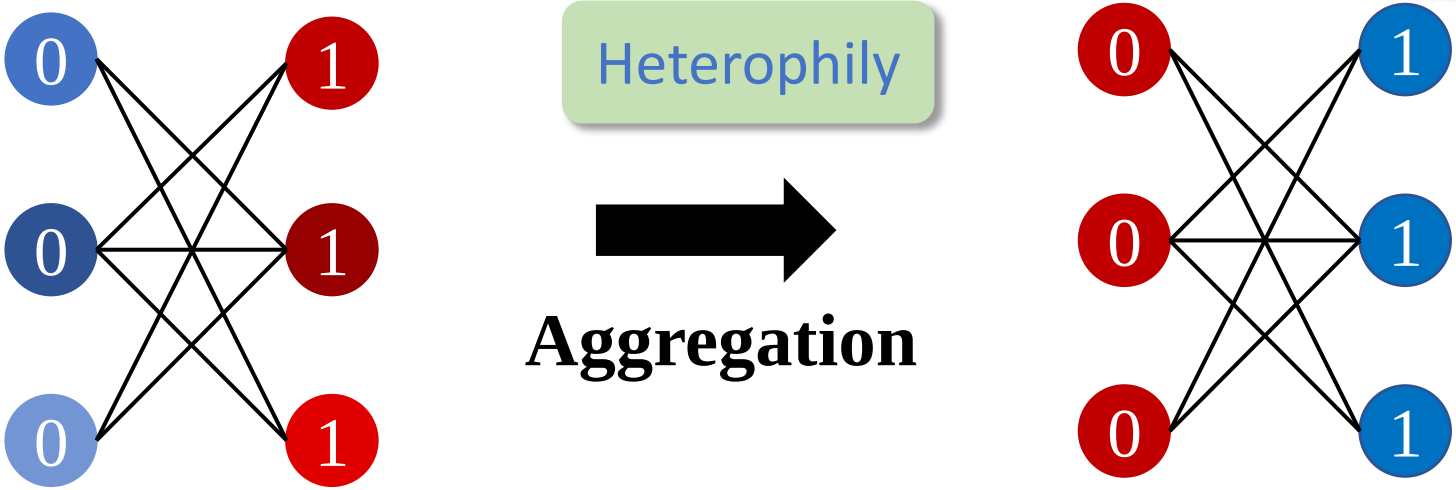


(d) Squirrel (h=0.25)

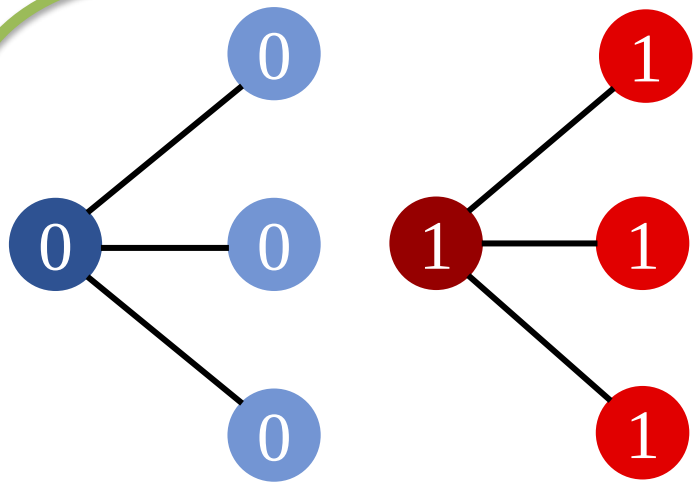
GCN outperforms on the majority pattern, but fails in the minor pattern

How can GNN work well on homophily & heterophily?

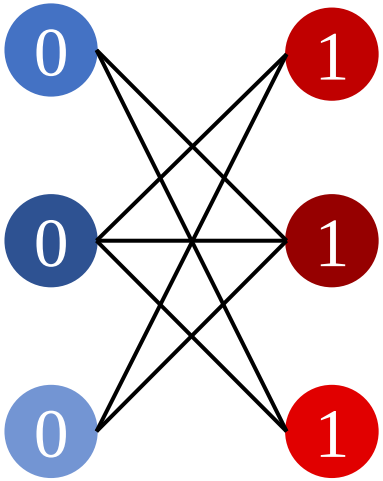




How can GNN work well on homophily & heterophily?



Homophily



Heterophily

Before aggregation

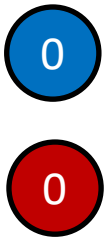
Nodes with class 0 in Blue 0

Nodes with class 1 in Red 1

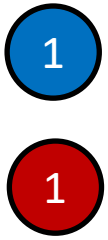
How can GNN work well on homophily & heterophily?

After aggregation

Nodes with class 0 in both Blue and red

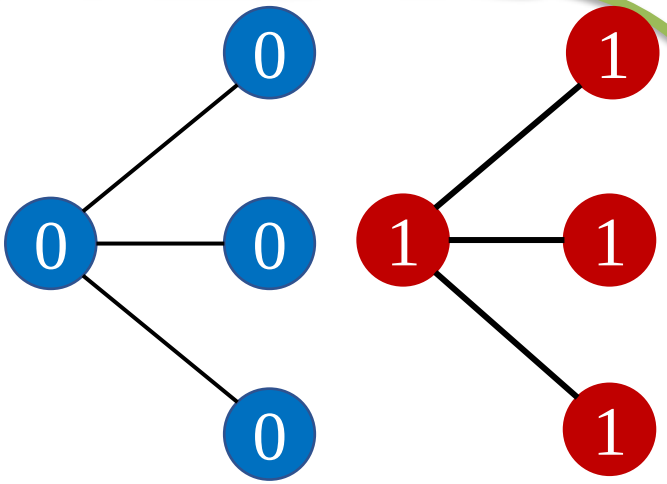


Nodes with class 1 in both Blue and red

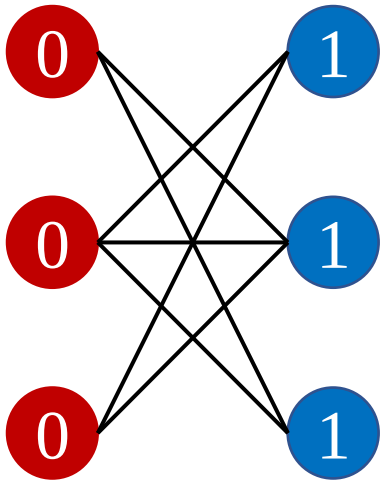


Hard to classify!

Homophily



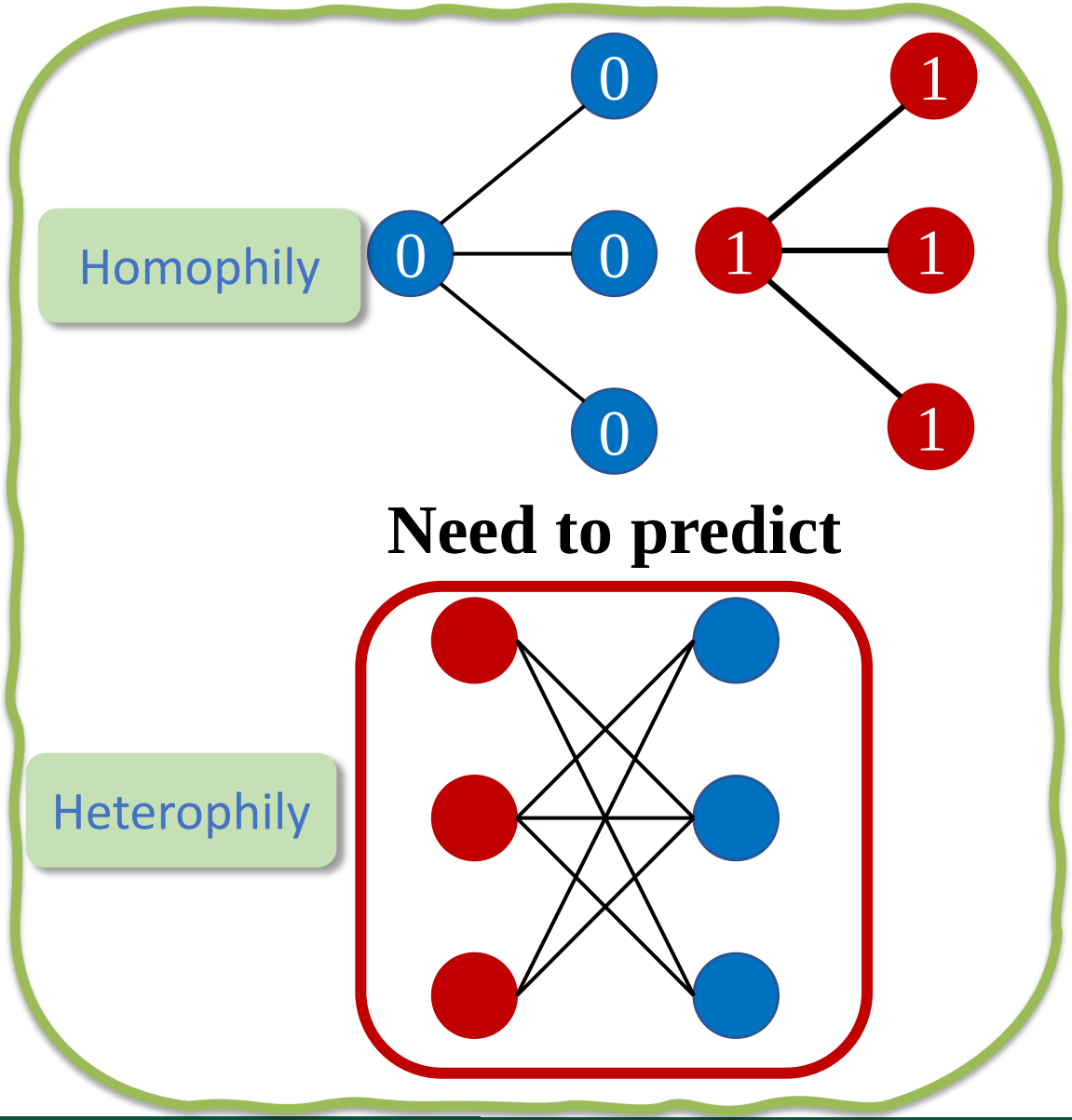
Heterophily



How can GNN work well on homophily & heterophily?

After aggregation

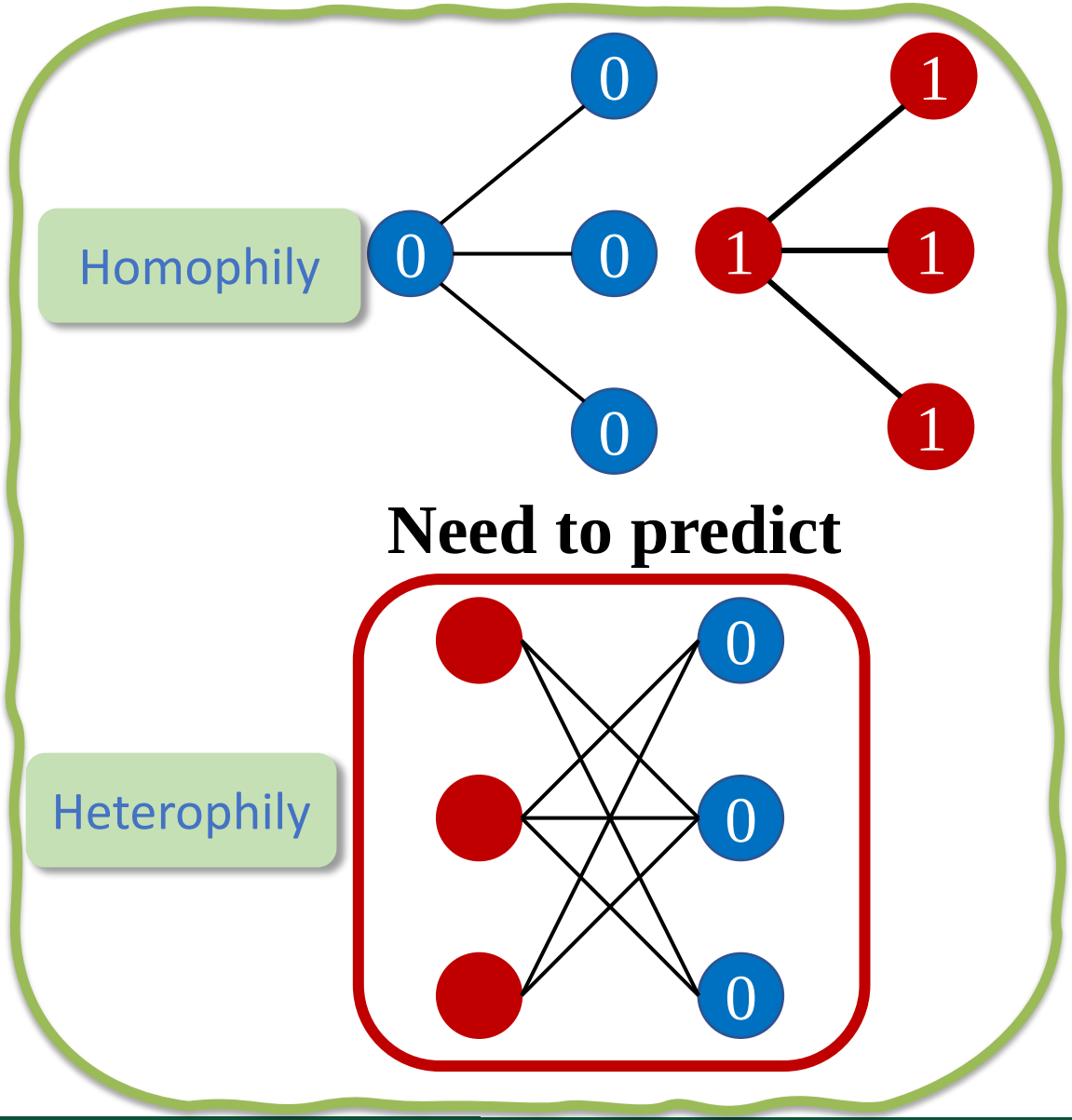
If all the homophily nodes are labeled, all the heterophily nodes are unlabeled



How can GNN work well on homophily & heterophily?

After aggregation

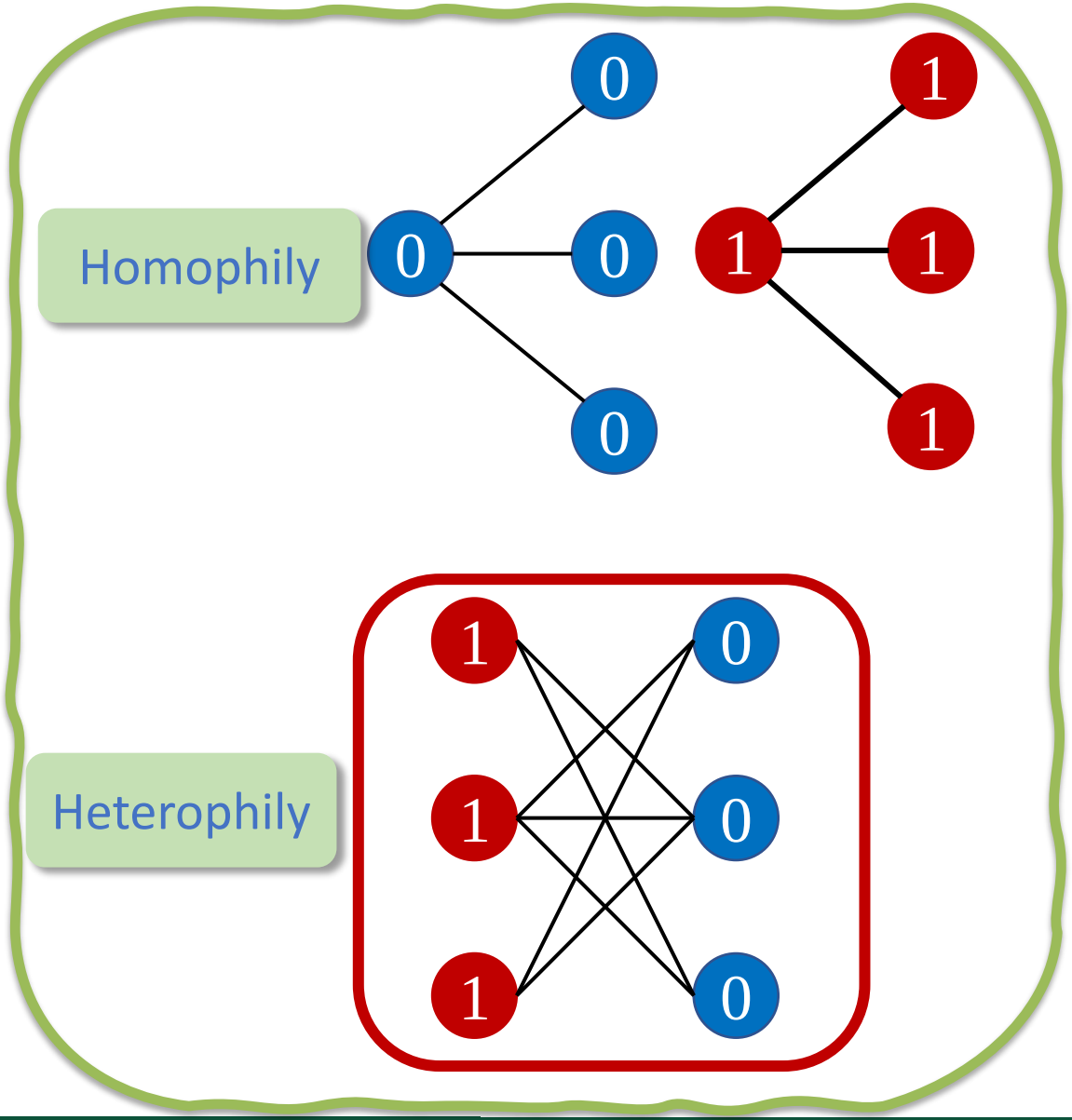
If all the homophily nodes are labeled, all the heterophily nodes are unlabeled



How can GNN work well on homophily & heterophily?

After aggregation

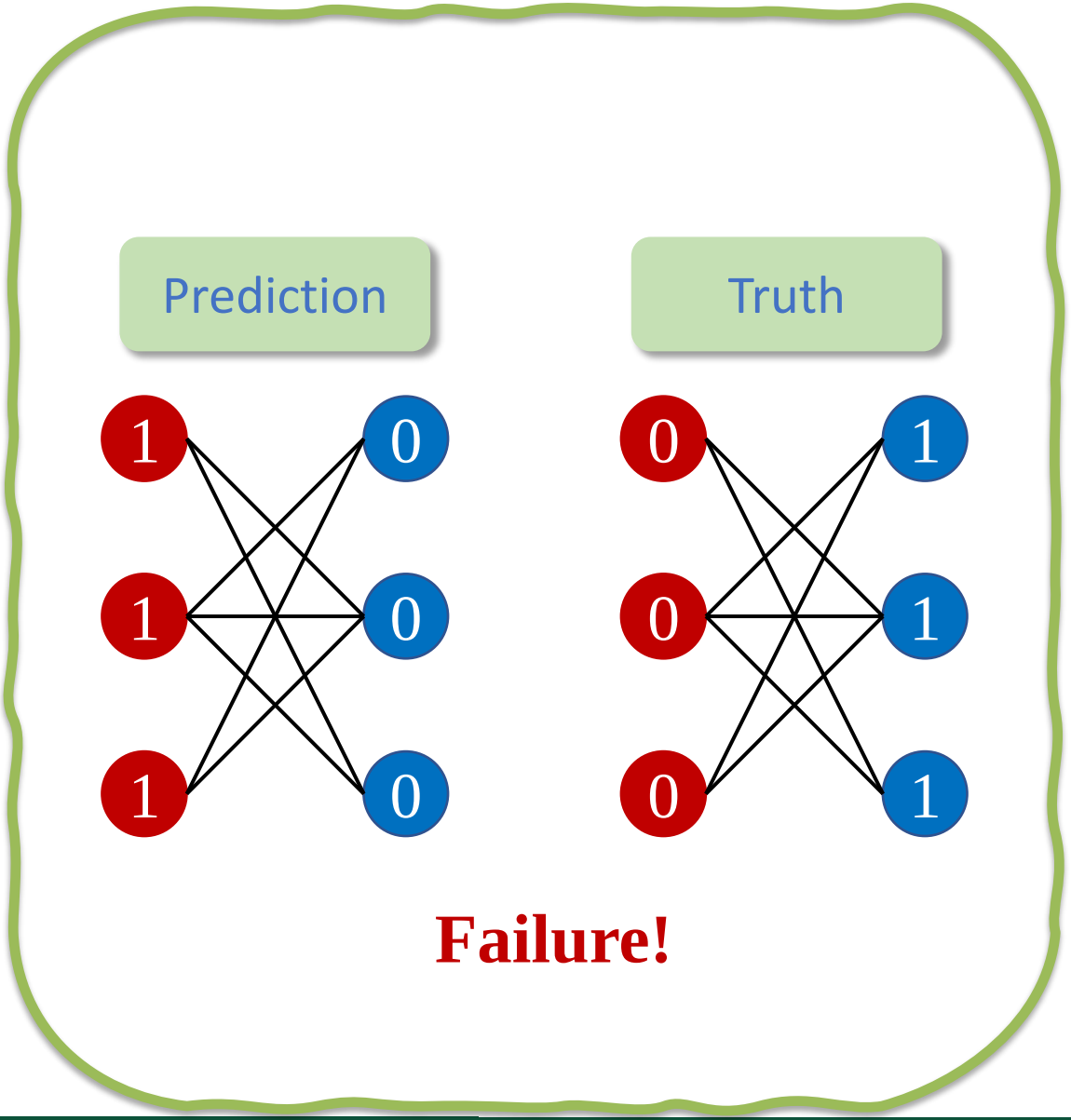
If all the homophily nodes are labeled, all the heterophily nodes are unlabeled



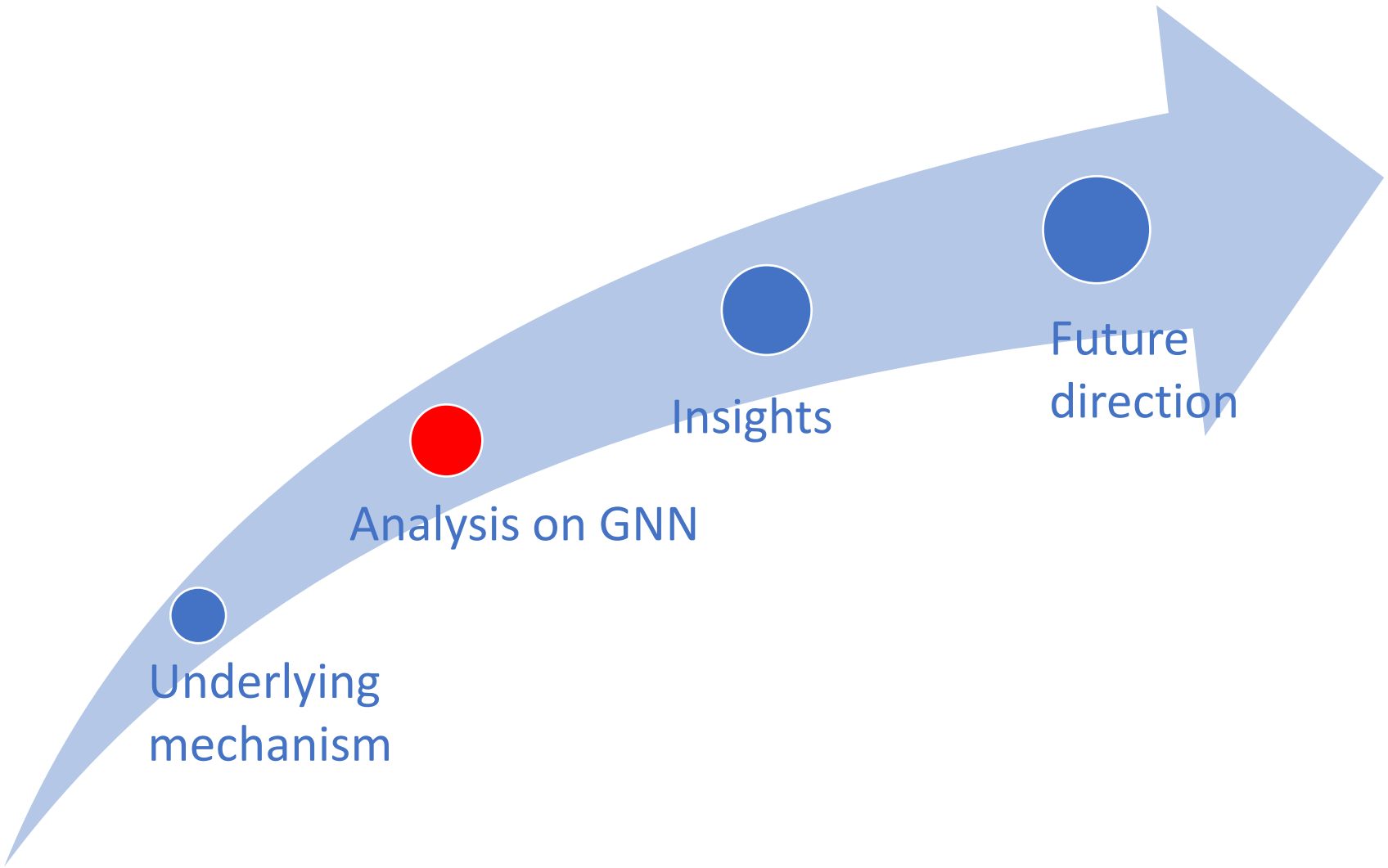
How can GNN work well on homophily & heterophily?

After aggregation

If all homophily nodes are labeled, failures in heterophily nodes.



Outline


Underlying mechanism
Analysis on GNN
Insights
Future direction

How Aggregation affects nodes differently?

Lemma 1. When nodes u and v have the same aggregated features $\mathbf{f}_u = \mathbf{f}_v$ but different structural patterns $h_u \neq h_v$

$$|\mathbf{P}_1(y_u = c_1|\mathbf{f}_u) - \mathbf{P}_2(y_v = c_1|\mathbf{f}_v)| \leq \frac{\rho^2}{\sqrt{2\pi}\sigma}|h_u - h_v|$$

The probability difference on nodes sharing the same class

Homophily ratio Difference (Structure disparity)

Nodes with a small homophily ratio difference are likely to share the same class

Why performance disparity happen on GNN?

Theorem 1 (Subgroup Generalization Bound for GNNs). *Let $\tilde{h}$ be any classifier in the classifier family $\mathcal{H}$ with parameters $\{\widetilde{W}_l\}_{l=1}^{L}$. for any $0 < m \leq M$, $\gamma \geq 0$, and large enough number of the training nodes $N_{tr} = |V_{tr}|$, there exist $0 < \alpha < \frac{1}{4}$ with probability at least $1 - \delta$ over the sample of $y^{tr} := \{y_i\}_{i \in V_{tr}}$, we have:*

Train loss

$$\mathcal{L}_m^0(\tilde{h}) \leq \widehat{\mathcal{L}}_{tr}^{\gamma}(\tilde{h}) + O\left(\underbrace{\frac{K\rho}{\sqrt{2\pi}\sigma}(\epsilon_m + |h_{tr} - h_m| \cdot \rho)}_{(a)} + \underbrace{\frac{b\sum_{l=1}^{L}\|\widetilde{W}_l\|_F^2}{(\gamma/8)^{2/L}N_{tr}^{\alpha}}(\epsilon_m)^{2/L}}_{(b)} + \mathbf{R} \right)$$

Test node subgroup with homophily ratio h_m

Small gap indicates better generalization performance

Why performance disparity happen on GNN?

Theorem 1 (Subgroup Generalization Bound for GNNs). *Let $\tilde{h}$ be any classifier in the classifier family $\mathcal{H}$ with parameters $\{\widetilde{W}_l\}_{l=1}^{L}$. for any $0 < m \leq M$, $\gamma \geq 0$, and large enough number of the training nodes $N_{tr} = |V_{tr}|$, there exist $0 < \alpha < \frac{1}{4}$ with probability at least $1 - \delta$ over the sample of $y^{tr} := \{y_i\}_{i \in V_{tr}}$, we have:*

$$\mathcal{L}_m^0(\tilde{h}) \leq \widehat{\mathcal{L}}_{tr}^{\gamma}(\tilde{h}) + O\left(\underbrace{\frac{K\rho}{\sqrt{2\pi}\sigma}\left(\epsilon_m + |h_{tr} - h_m| \cdot \rho\right)}_{(a)} + \underbrace{\frac{b\sum_{l=1}^{L}\|\widetilde{W}_l\|_F^2}{(\gamma/8)^{2/L} N_{tr}^{\alpha}}\left(\epsilon_m\right)^{2/L}}_{(b)} + \mathbf{R} \right)$$

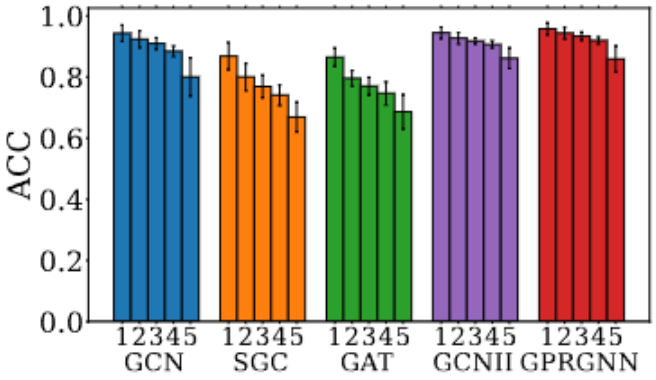
$\epsilon_m = \left|f_i - f_j\right|_F^2$ is the aggregated feature distance between train and test subgroup(s).

$|h_{tr} - h_m|$ is the homophily ratio difference between train and test subgroup(s).

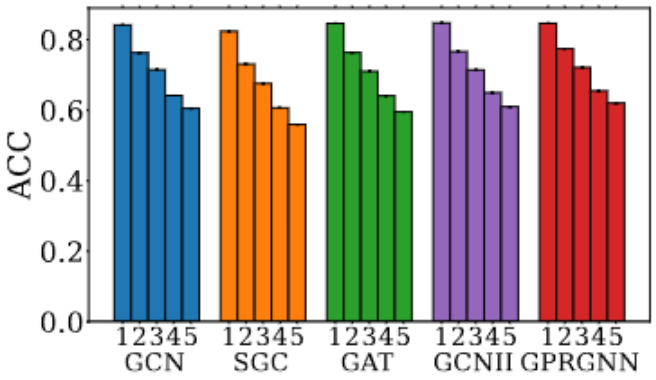
Structure disparity

Empirical verification

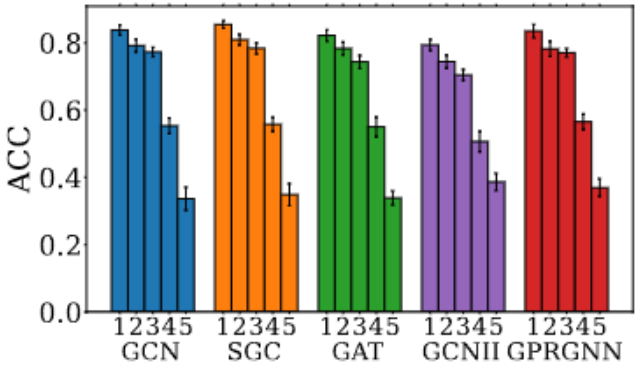
$$s = \epsilon_m + |h_{\text{tr}} - h_m|$$



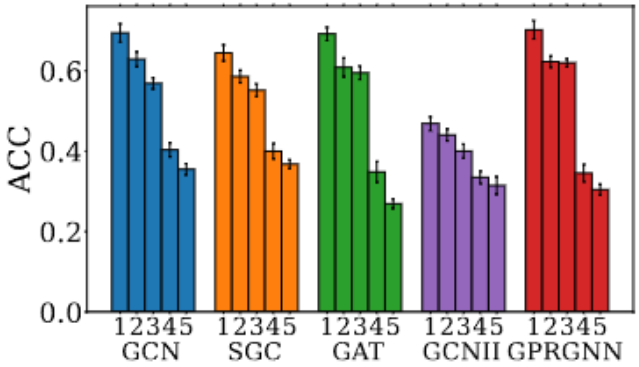
(a) PubMed (h=0.79)



(b) Ogbn-arxiv (h=0.63)



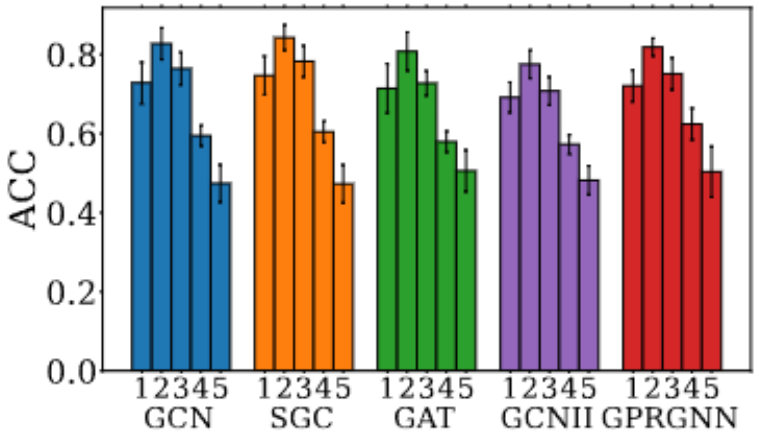
(c) Chameleon (h=0.22)



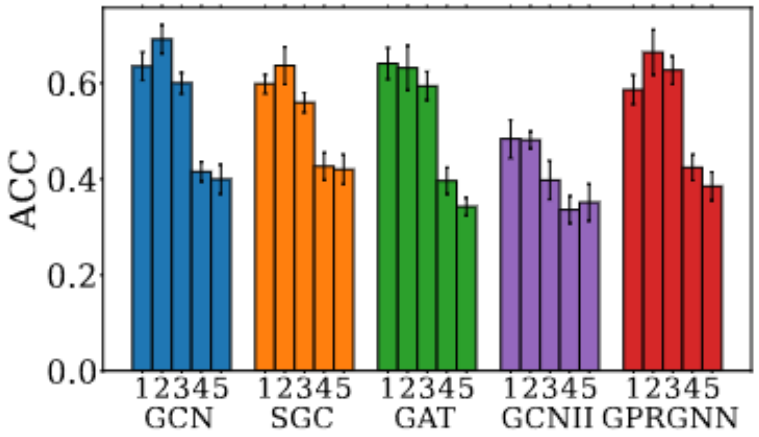
(d) Squirrel (h=0.25)

Empirical verification

$$s = \epsilon_m$$



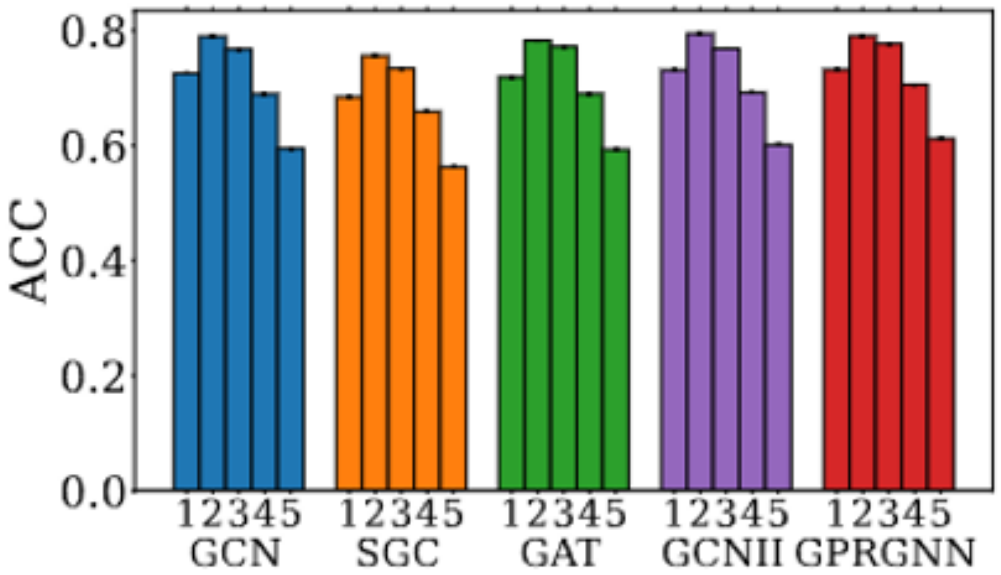
(c) Chameleon (h=0.22)



(d) Squirrel (h=0.25)

Empirical verification

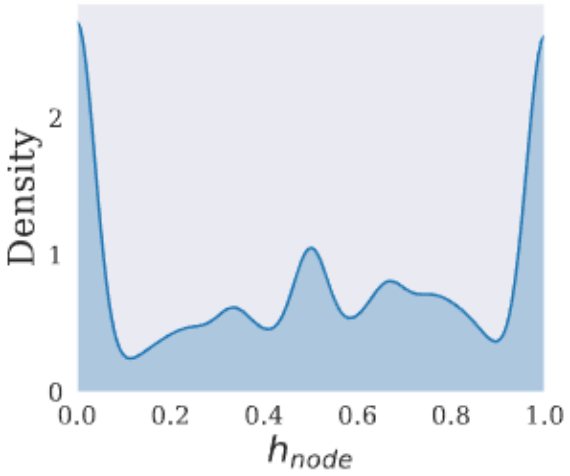
$$s = |h_{\text{tr}} - h_m|$$



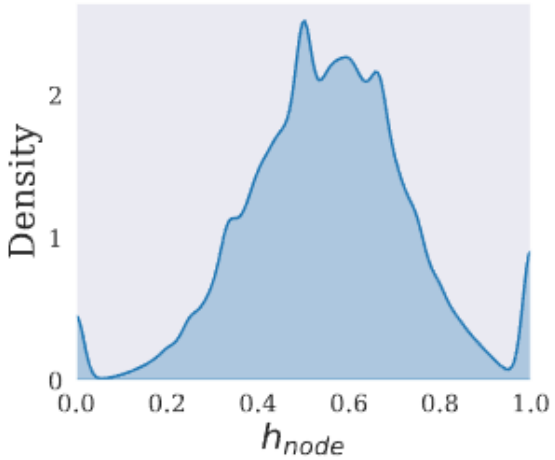
(b) Ogbn-arxiv (h=0.63)

Empirical success on more datasets

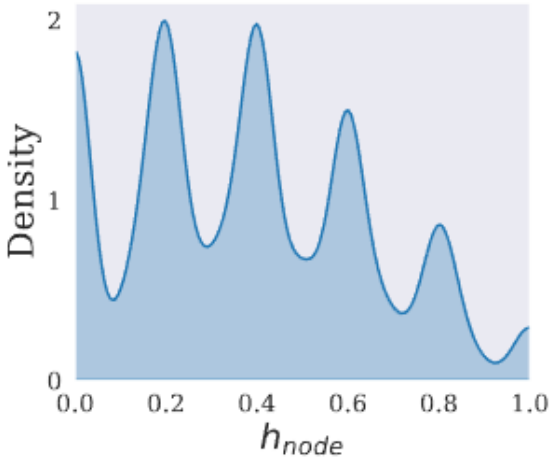
$$s = \epsilon_m + |h_{\text{tr}} - h_m|$$



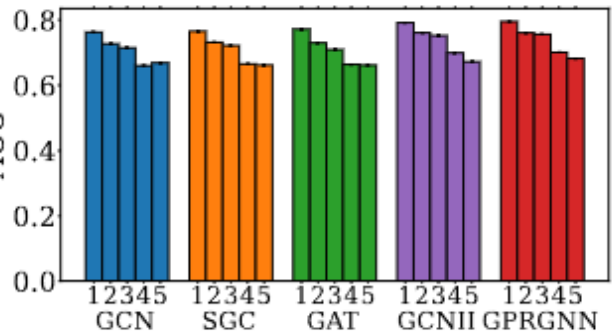
(e) IGB-tiny (h=0.57)



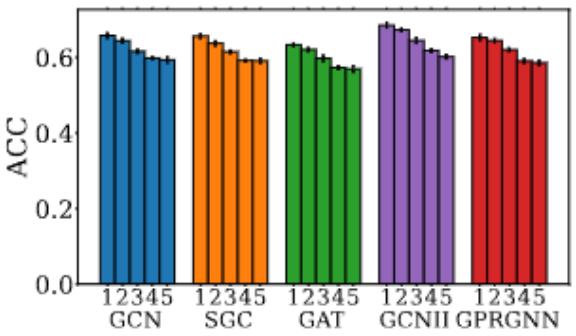
(i) Twitch-gamers (h=0.56)



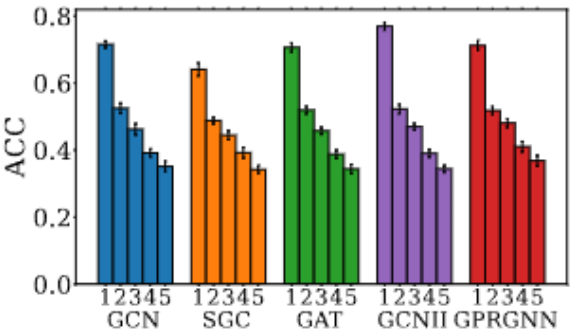
(j) Amazon-ratings (h=0.38)



(c) IGB-Tiny (h=0.57)

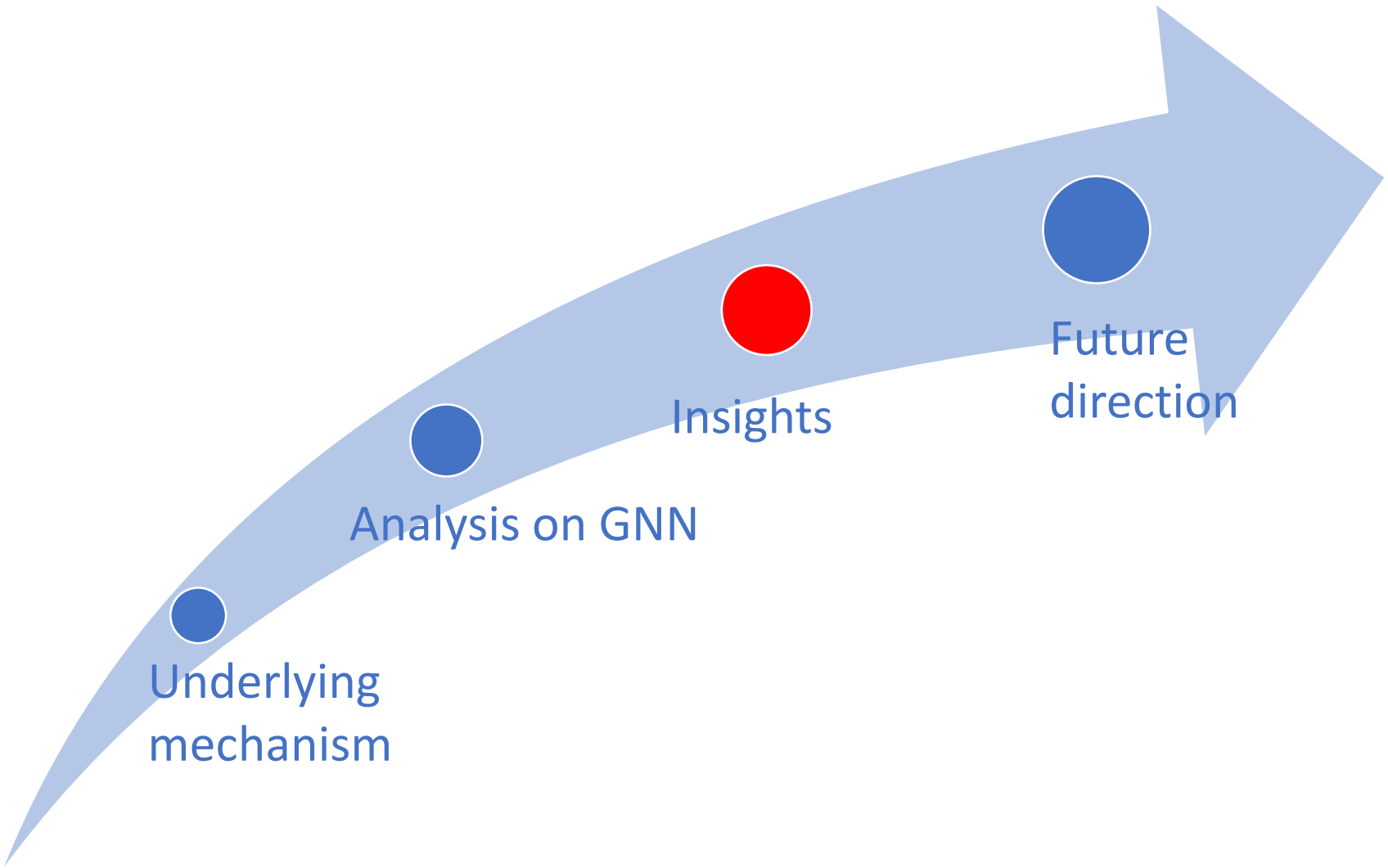


(e) Twitch-gamers (h=0.56)



(f) Amazon-ratings (h=0.38)

Outline


Underlying mechanism
Analysis on GNN
Insights
Future direction

Insights for building foundation model

Multiple insights:

- What is transferable across different datasets in node classification?
- What is the essential difficulty for the model design?
- Instructions for building the Graph Foundation Model

Insights for building foundation model

Multiple insights:

- What is transferable across different datasets in node classification?
- What is the essential difficulty for the model design?
- Instructions for building the Graph Foundation Model

Insights for building foundation model

What is transferable across different datasets?

Homophily

- Two data factors are key concepts across different datasets

Heterophily

- Factors are not independent but co-exist across all the datasets

Insights for building foundation model

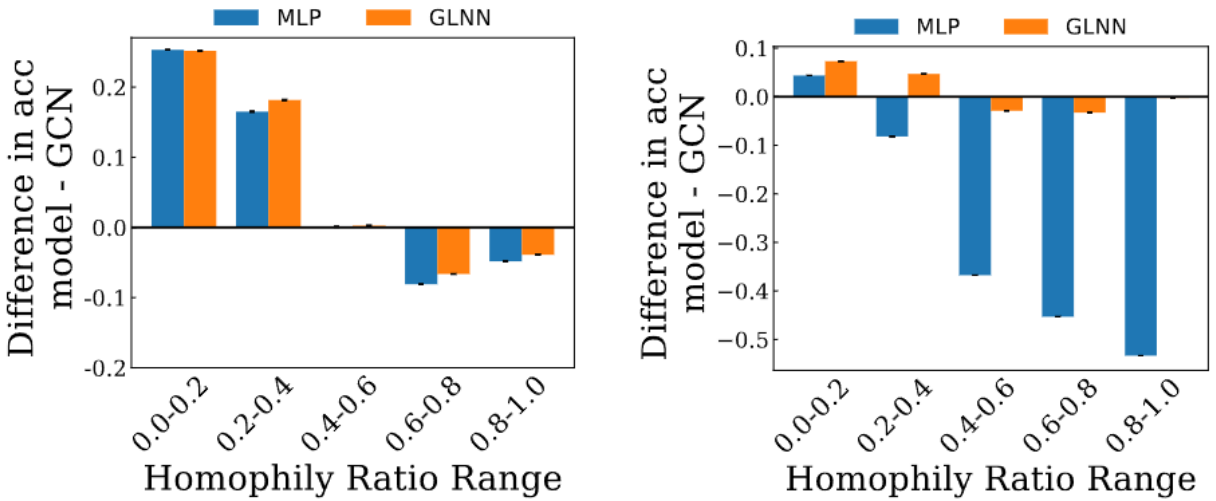
Multiple insights:

- What is transferable across different datasets in node classification?
- What is the essential difficulty for the model design?
- Instructions for building the Graph Foundation Model

The incompatibility between homophily and heterophily

The incompatibility between homo and hete

Performance comparison between
GCN and MLP-based models



(a) PubMed (h=0.79) (b) Ogbn-arxiv (h=0.63)

GCN outperforms on the majority pattern,
but fails in the minor pattern

Insights for building foundation model

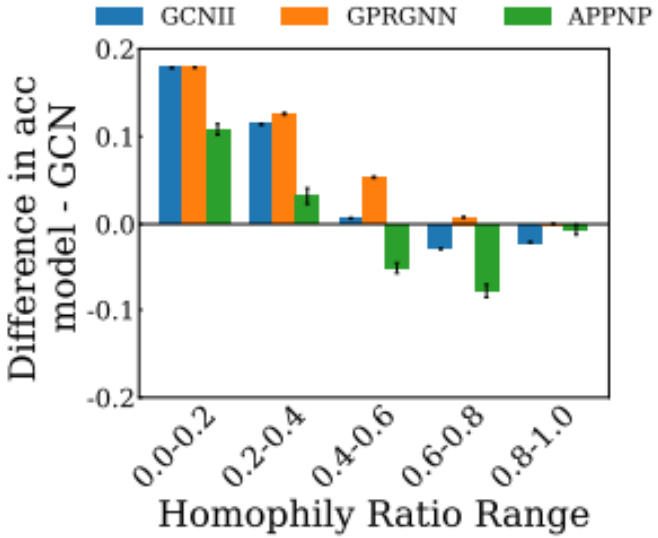
Multiple insights:

- What is transferable across different datasets in node classification?
- What is the essential difficulty for the model design?
- Instructions for building the Graph Foundation Model

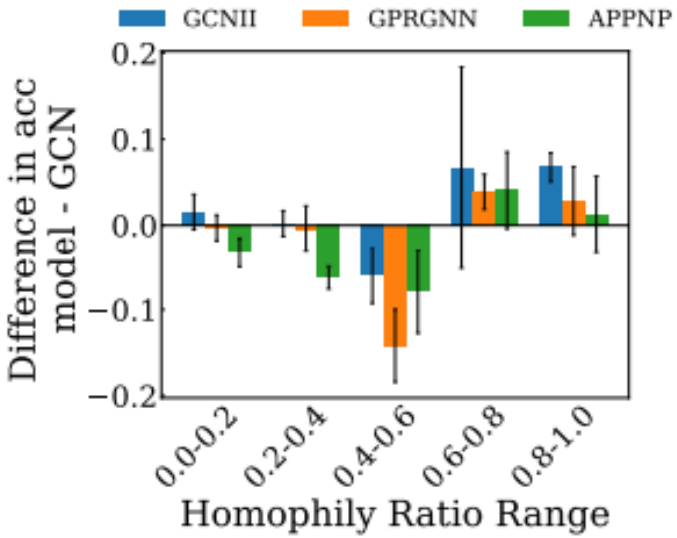
Deeper model is the potential solution

Elucidating the effectiveness of Deeper GNNs

Comparison between GCN and Deeper GNNs



(a) PubMed (h=0.79)



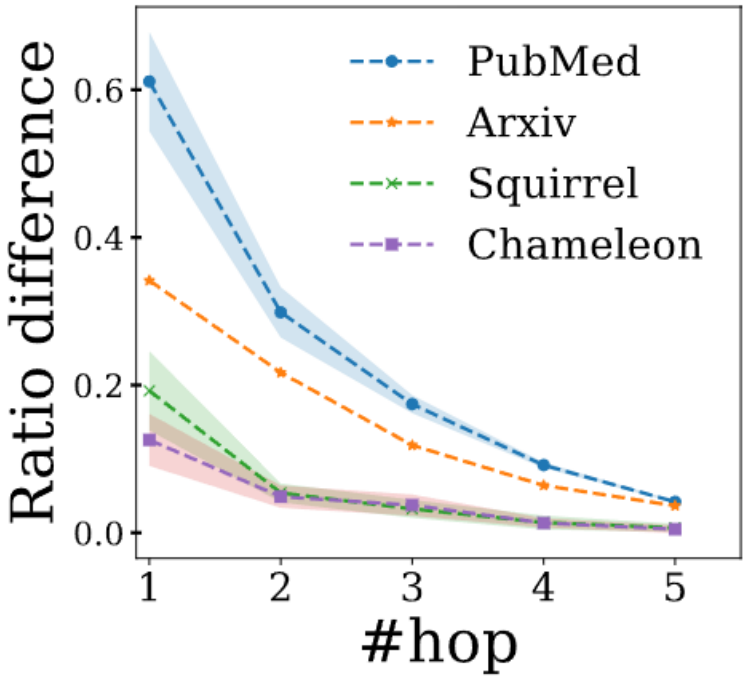
(c) Chameleon (h=0.22)

The performance improvement from Deeper GNNs is from the minority nodes

Elucidating the effectiveness of Deeper GNNs

Multiple-hop homophily ratio differences between training and minority test nodes

$$|h_{\text{tr}} - h_m|$$



The homophily ratio difference on minority pattern decreased in higher-order

Main takeaways!

Main takeaway

Model mechanism

Data mechanism

Reveal the inner work of GNNs

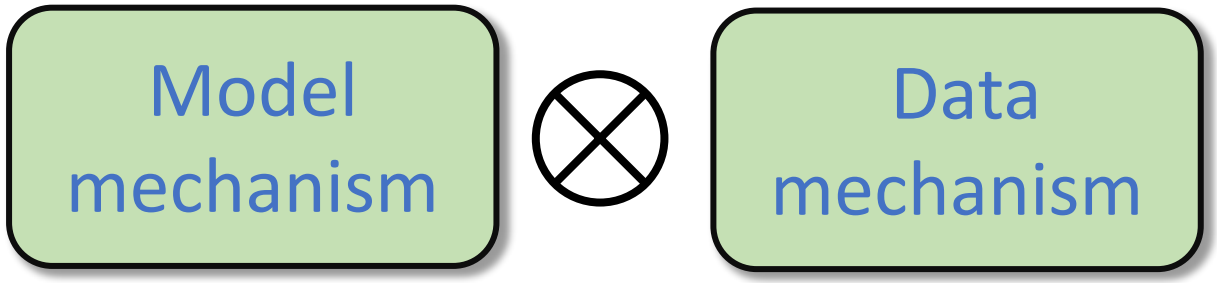
Main takeaway

Model mechanism

Data mechanism

Reveal how data principle plays an important role

Main takeaway



When they fit, GNNs can perform well

When not, GNNs underperform

Important data factors on different tasks

Link Prediction

- LSP
- GSP
- FP

Node classification

- Homophily
- Heterophily

Is there any shared concepts between different tasks?

Shared knowledge between tasks

Link Prediction

Node classification

FP

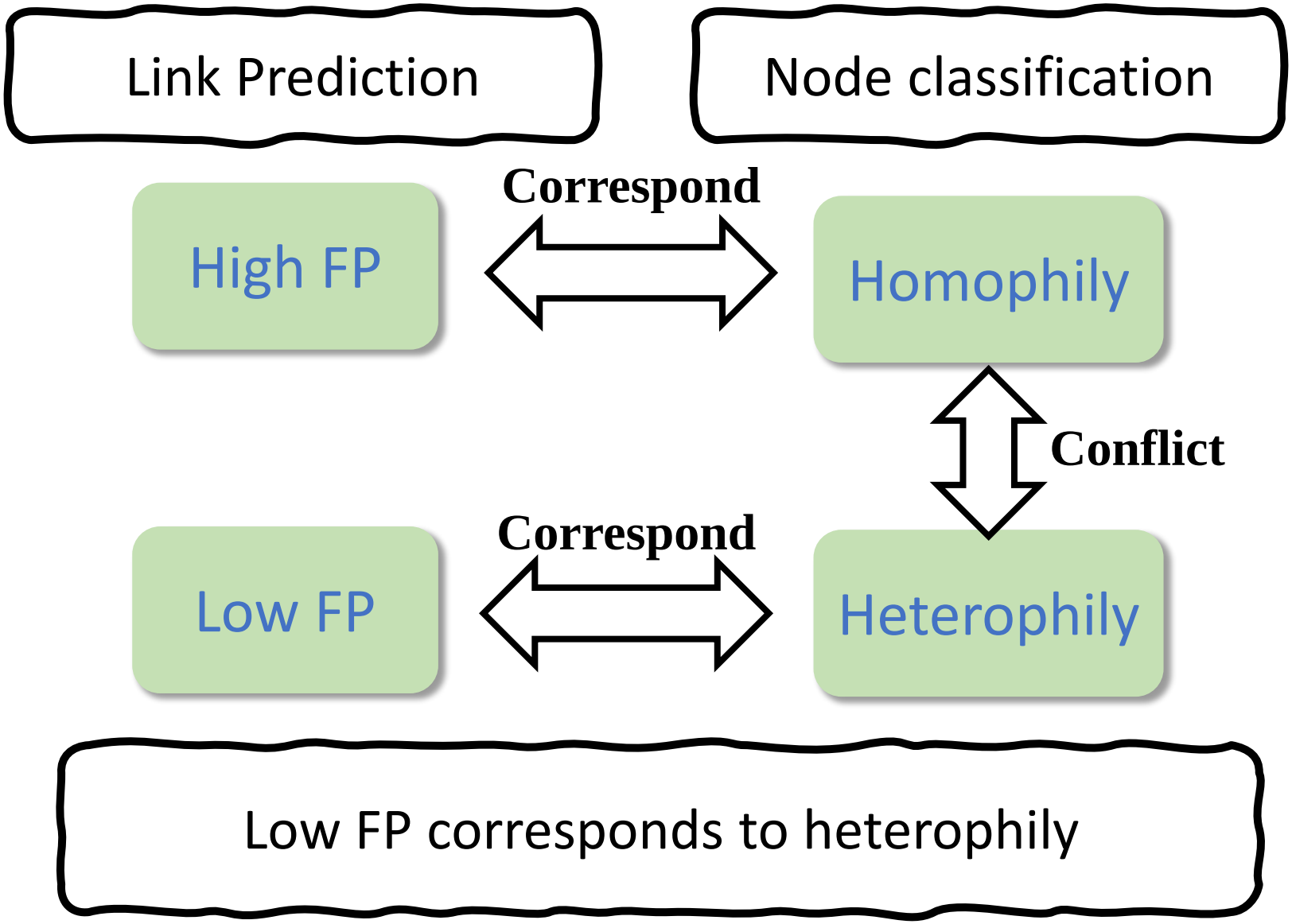
Homophily

Nodes with high feature similarity are more likely to be connected

Connected nodes are more likely to be with high feature similarity

FP and homophily are chicken-egg problem

Shared knowledge between tasks



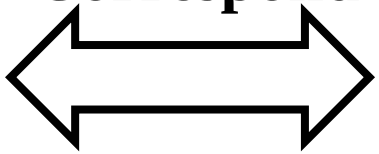
Shared knowledge between tasks

Link Prediction

Node classification

Low FP

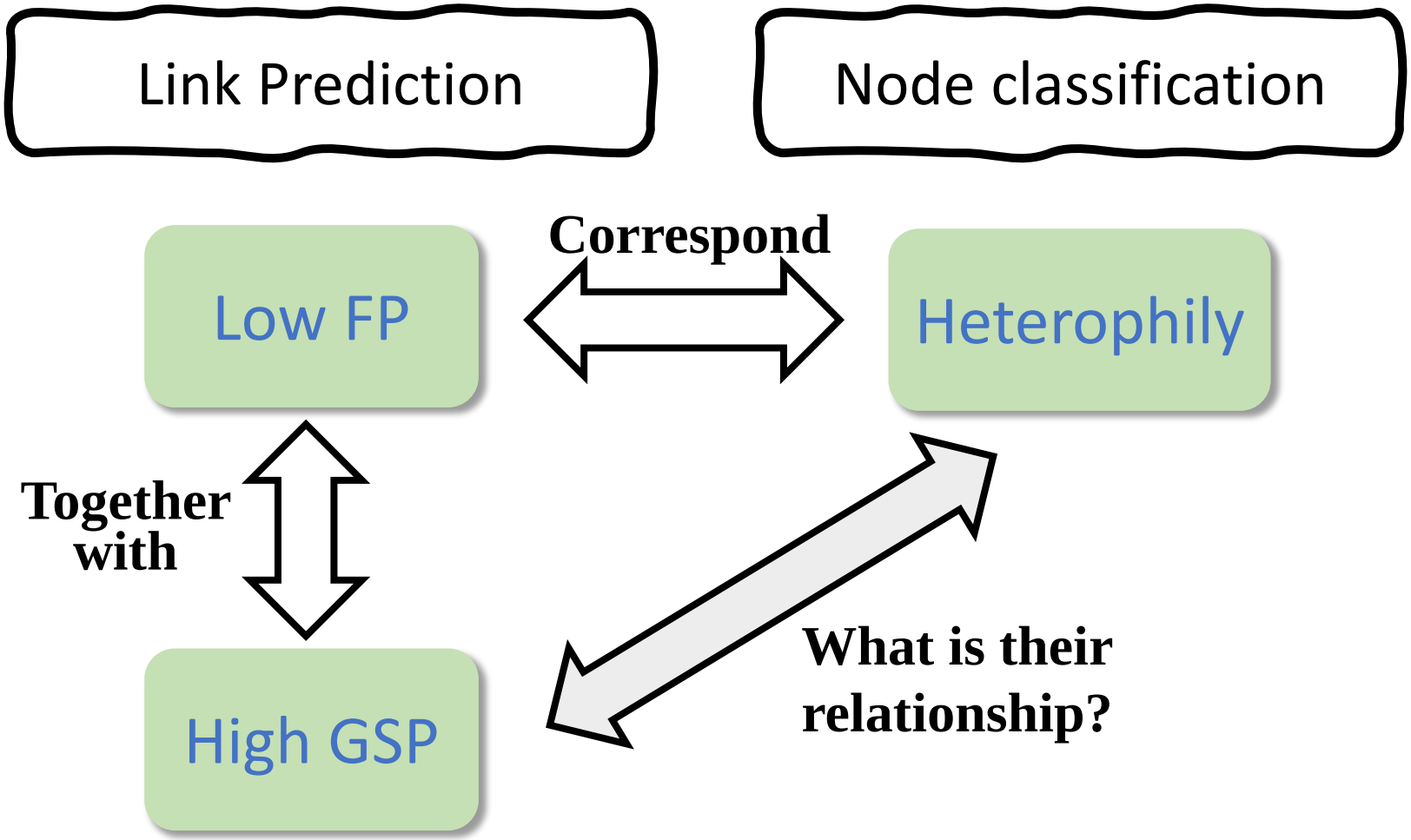
Correspond



Heterophily

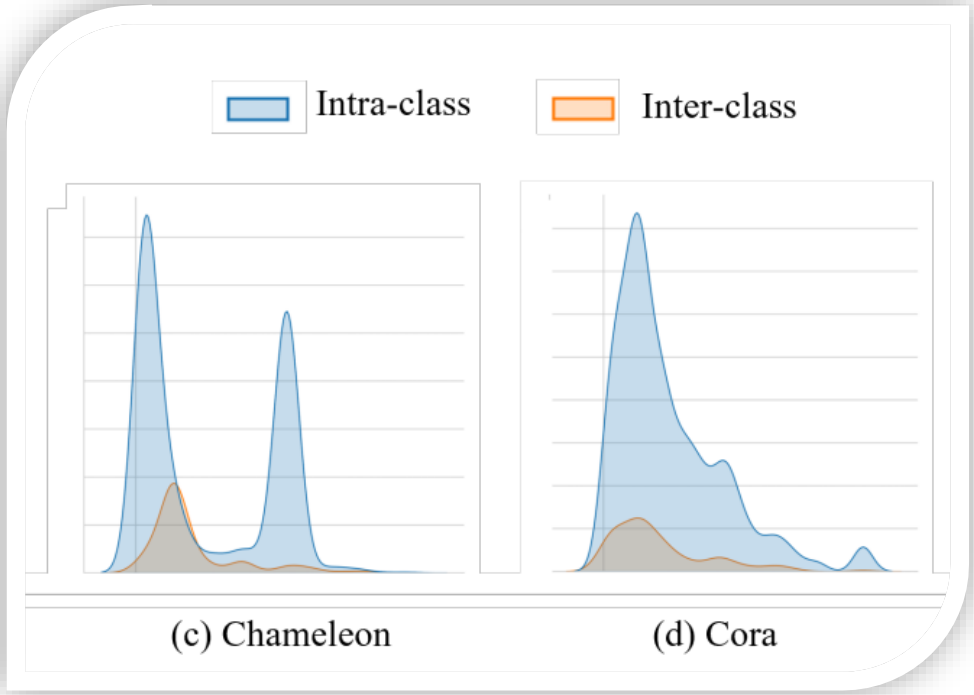
The incompatibility can be found between feature and structural proximity.

Shared knowledge between tasks



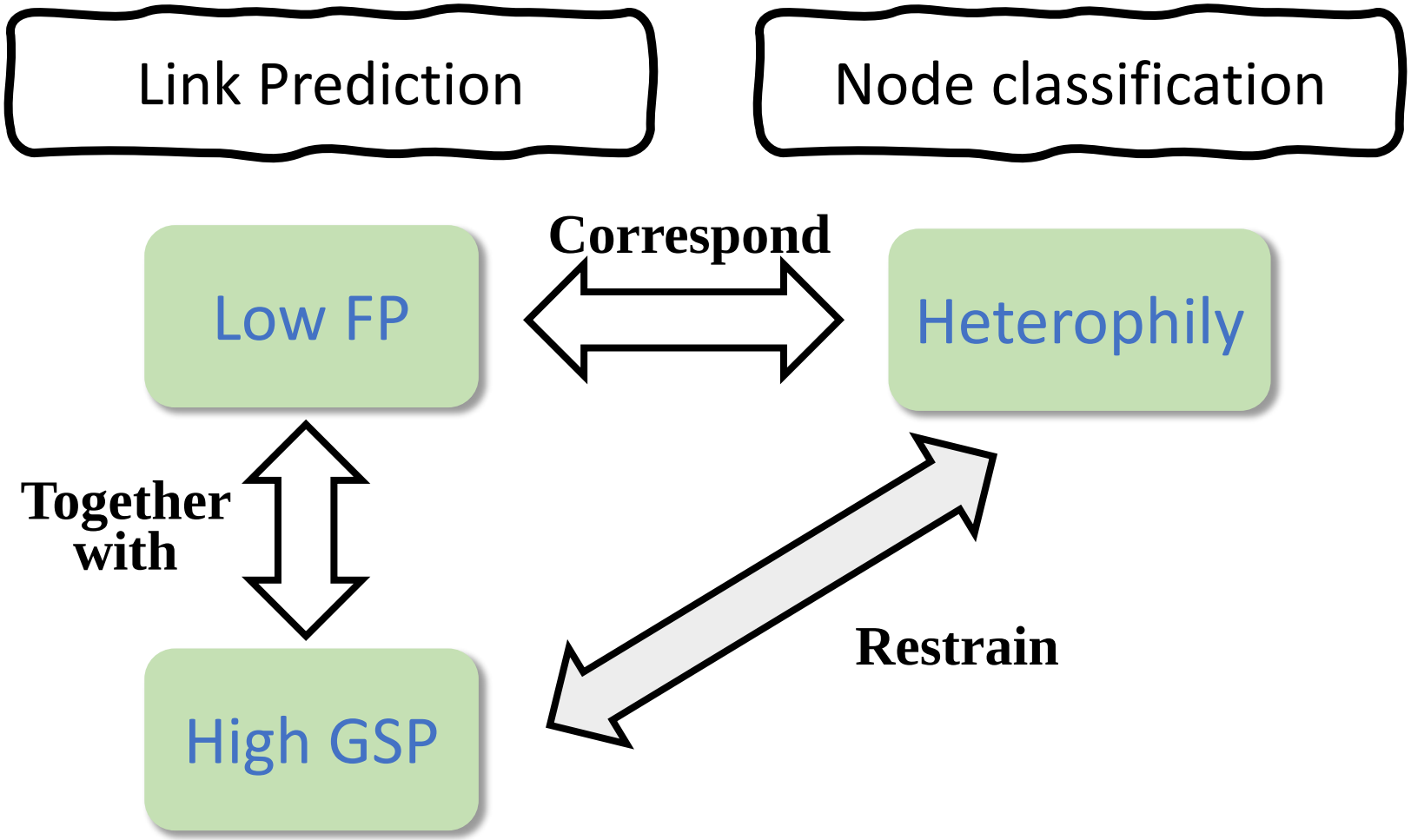
Shared knowledge between tasks

Distribution of SimRank scores over intra-class and inter-class node pairs.

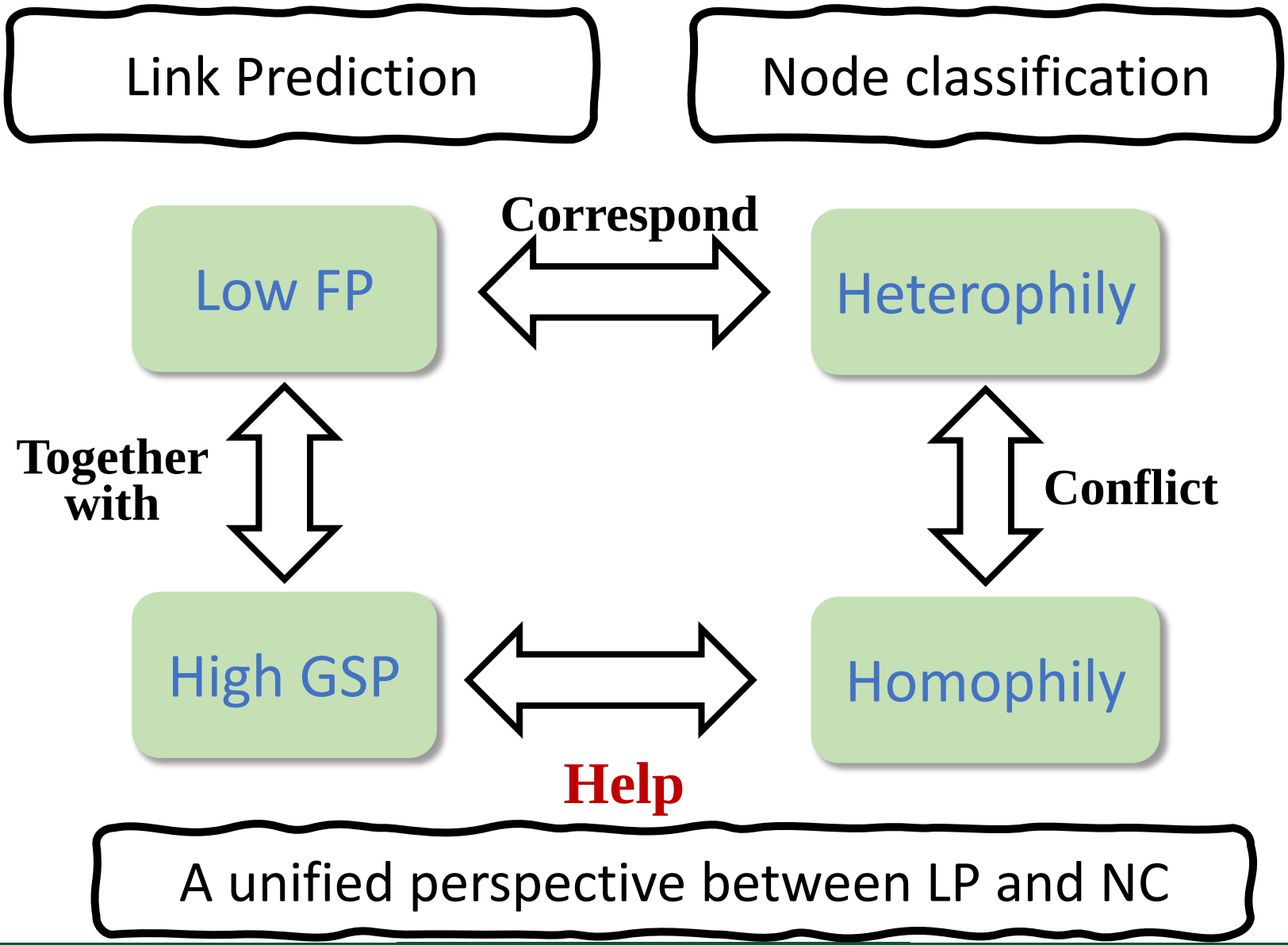


High GSP leads to homophily

Shared knowledge between tasks



Shared knowledge between tasks



Shared knowledge between tasks

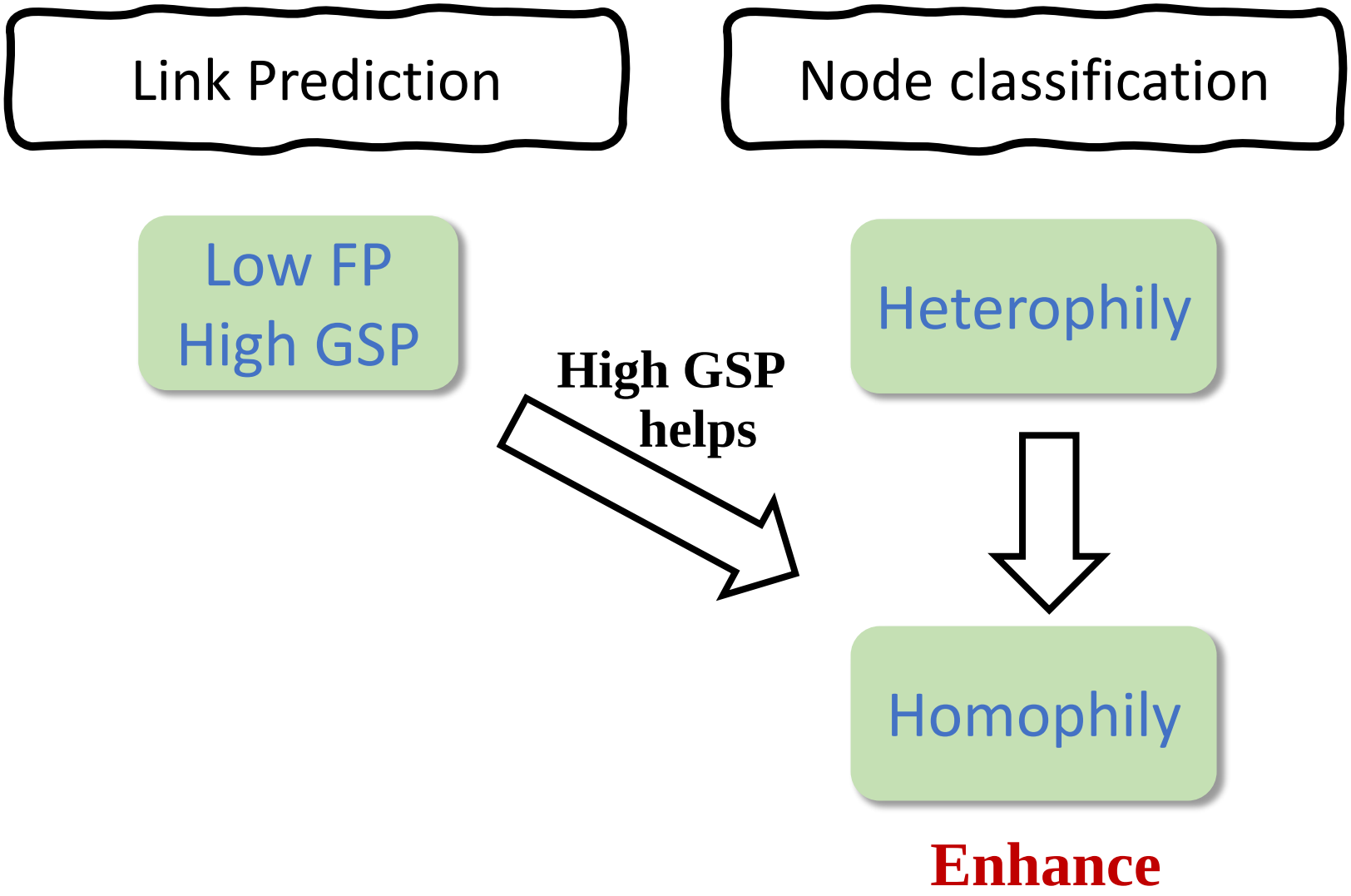
Link Prediction

Low FP
High GSP

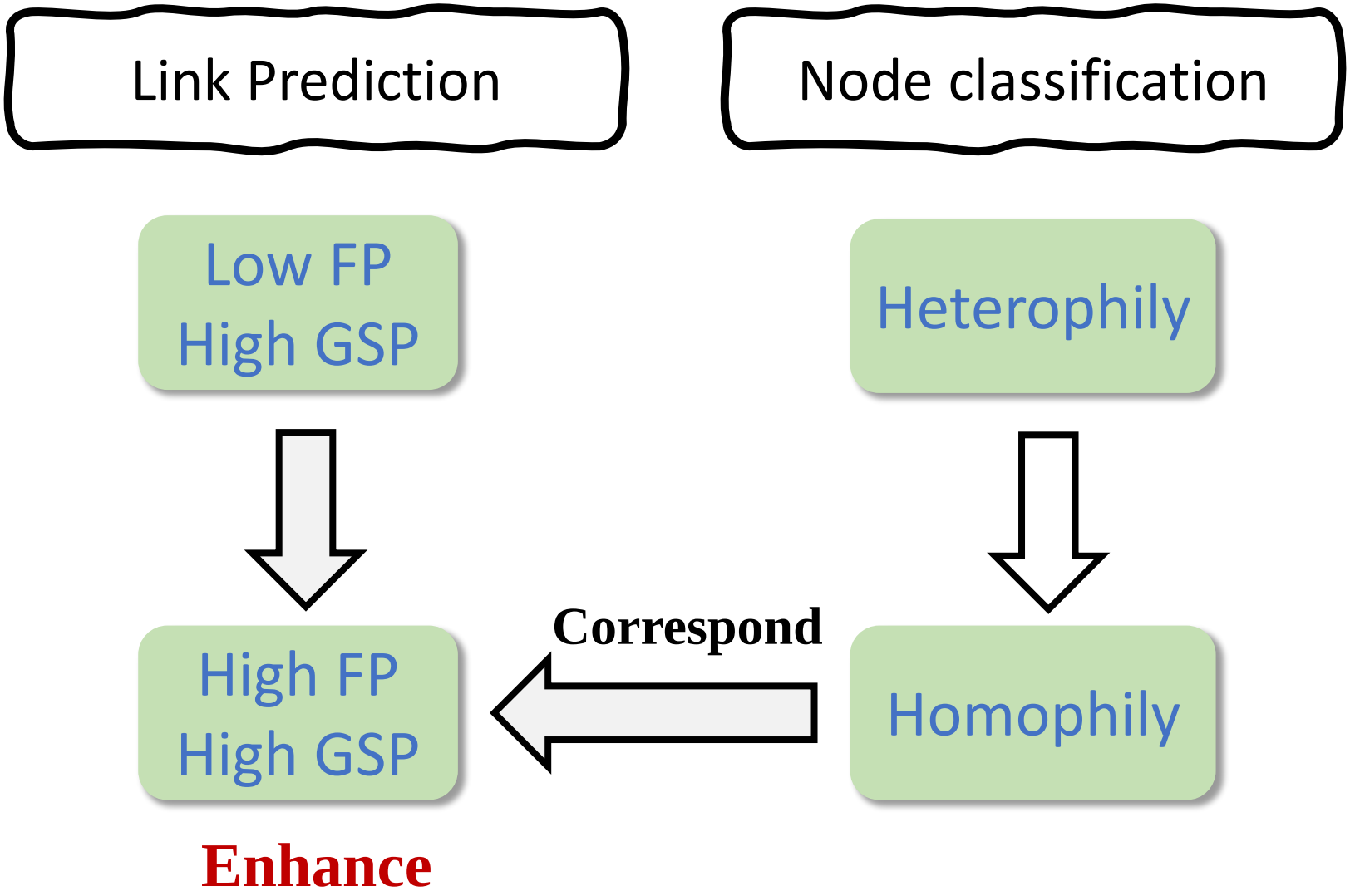
Node classification

Heterophily

Shared knowledge between tasks



Shared knowledge between tasks



Some Suggestions & look ahead

Graph Foundation model is encouraging

Feature alignment problem still remains

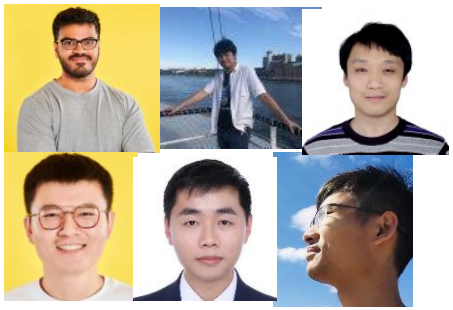
Graph classification still lacks understanding

After capturing data factors, how to use them

Acknowledgements

DSE Lab @ MSU

Collaborators



Funding Sources

Thanks & QA!

Appendix for Link Prediction

Concrete form for latent model

$$P(i \sim j | d_{ij}) = \begin{cases} \frac{1}{1 + e^{\alpha(d_{ij} - \max\{r_i, r_j\})}} \cdot (1 - \beta_{ij}) & d_{ij} \leq \max\{r_i, r_j\} \\ \beta_{ij} & d_{ij} > \max\{r_i, r_j\} \end{cases}$$

where $P(i \sim j | d_{ij})$ depicts the probability of forming an undirected link between i and j ($i \sim j$), predicated on both the features and structure. The latent distance d_{ij} indicates the structural likelihood of link formation between i and j. The feature proximity parameter $\beta_{ij} \in [0, 1]$ additionally introduces the influence from the feature perspective. Moreover, the model has two parameters α and r. $\alpha > 0$ controls the sharpness of the function. To ease the analysis, we set $\alpha = +\infty$. Discussions on when $\alpha \neq +\infty$ are in Appendix C.6. r_i is a connecting threshold parameter corresponding to node i. With $\alpha = +\infty$, $\frac{1}{1+e^{\alpha(d_{ij} - \max\{r_i, r_j\})}} = 0$ if $d_{ij} > \max\{r_i, r_j\}$, otherwise it equals to 1. Therefore, a large r_i indicates node i is more likely to form edges, leading to a potentially larger degree. Nodes in the graph can be associated with different r values, allowing us to model graphs with various degree distributions. Such flexibility enables our theoretical model to be applicable to more real-world graphs. We then identify how the model can reveal different important data factors in link prediction. To achieve this goal, we **(i)** derive heuristic scores revolving around each factor in the latent space and **(ii)** provide a theoretical foundation suggesting that each score can offer a suitable bound for the probability of link formation. Theoretical results underscore the effectiveness of each factor, e.g., LSP, GSP, and FP.

Research Question

Does each data factor indeed play a key role?

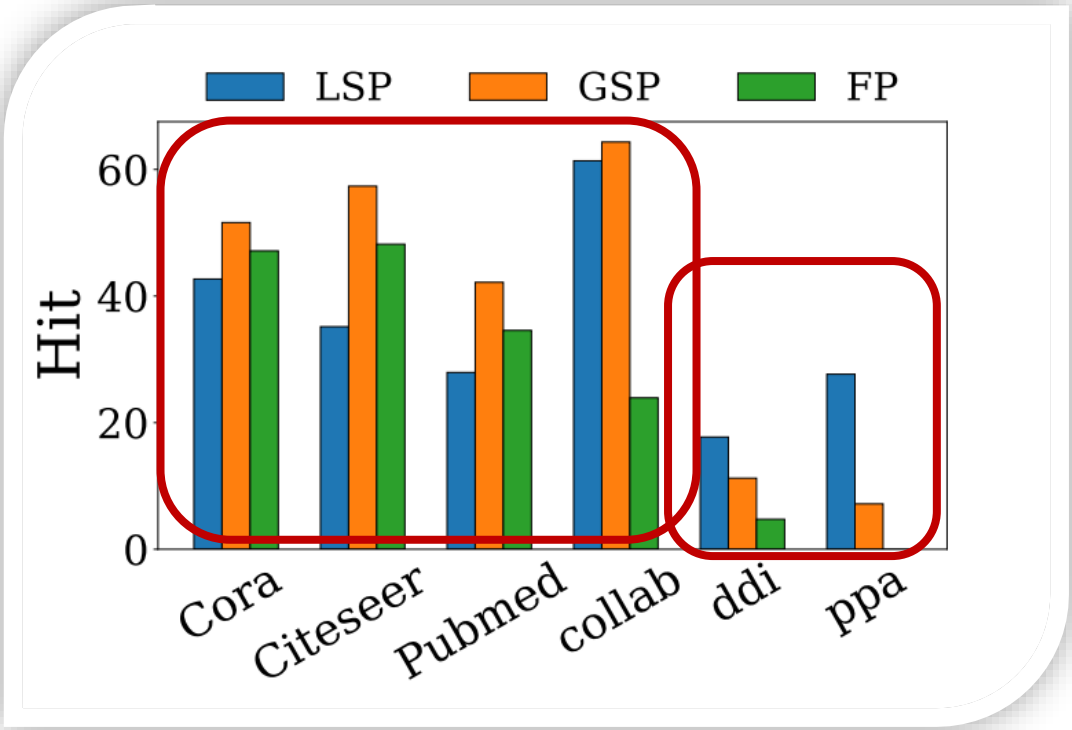
Does each data factor provide unique information?

Research Question

Does each data factor indeed play a key role?

Does each data factor provide unique information?

Effectiveness of factors



Performance of different heuristics on different factors

- No consistent winning solution
- The GSP works better on academic networks
- The LSP works better on chemistry and biology networks

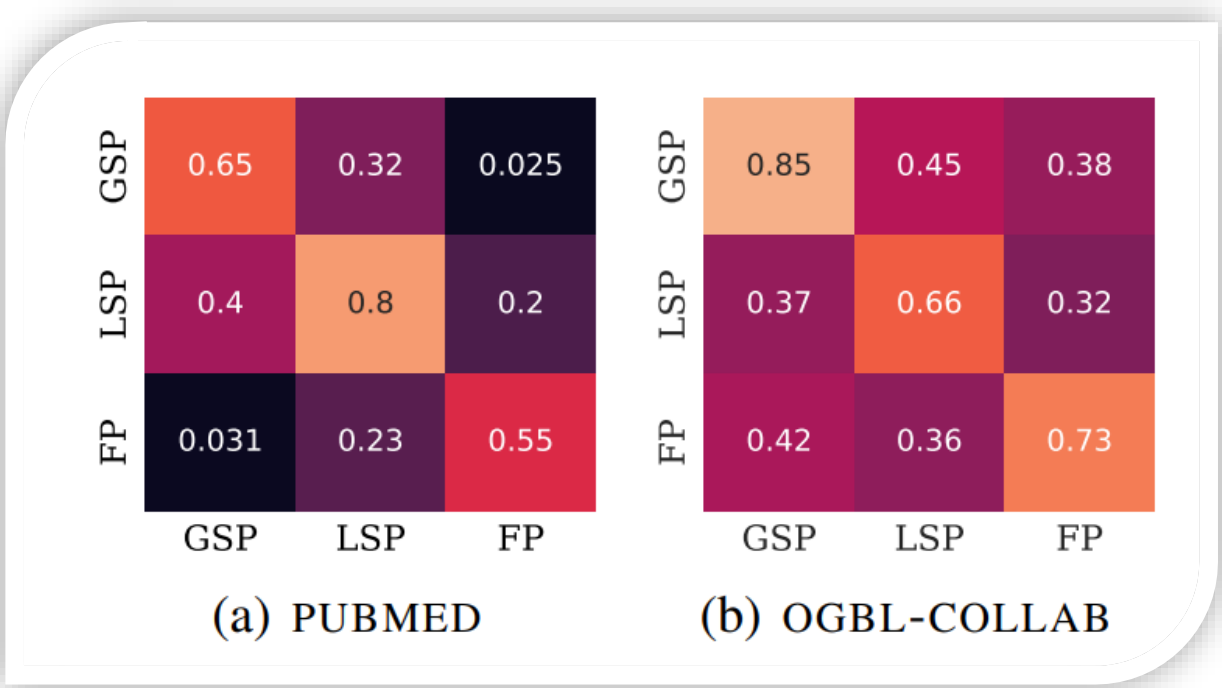
A high heuristic performance indicate the importance of the corresponding data factor

Research Question

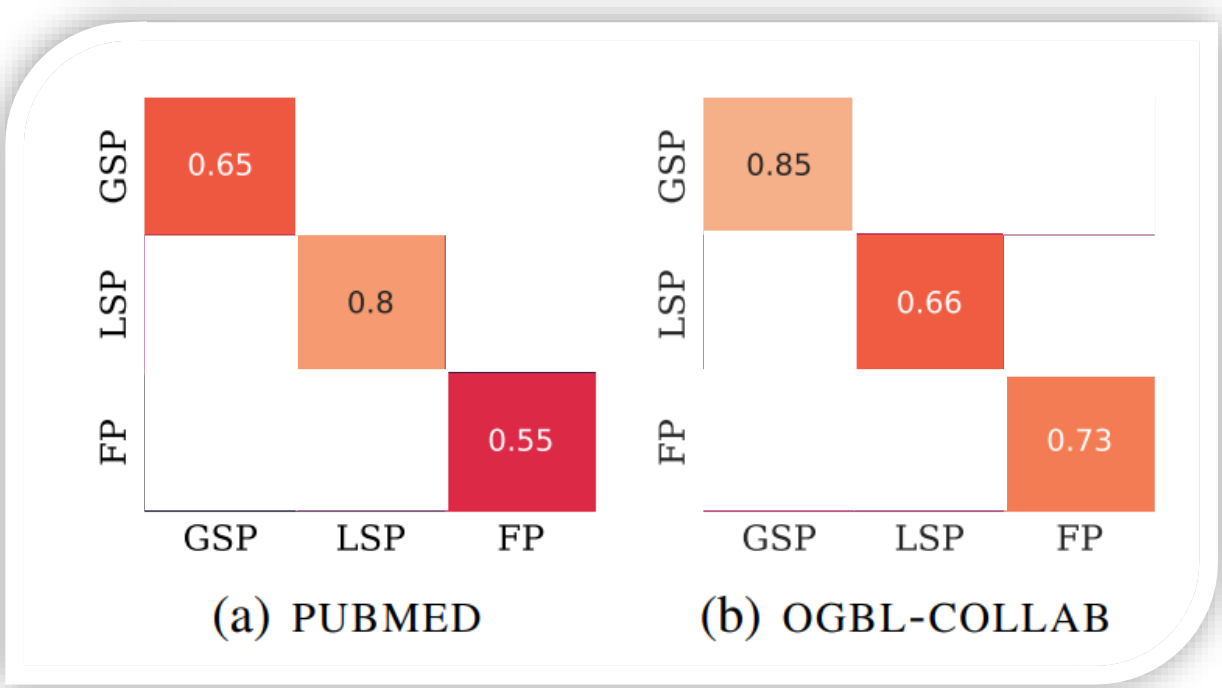
Does each data factor indeed play a key role?

Does each data factor provide unique information?

Complementary effects

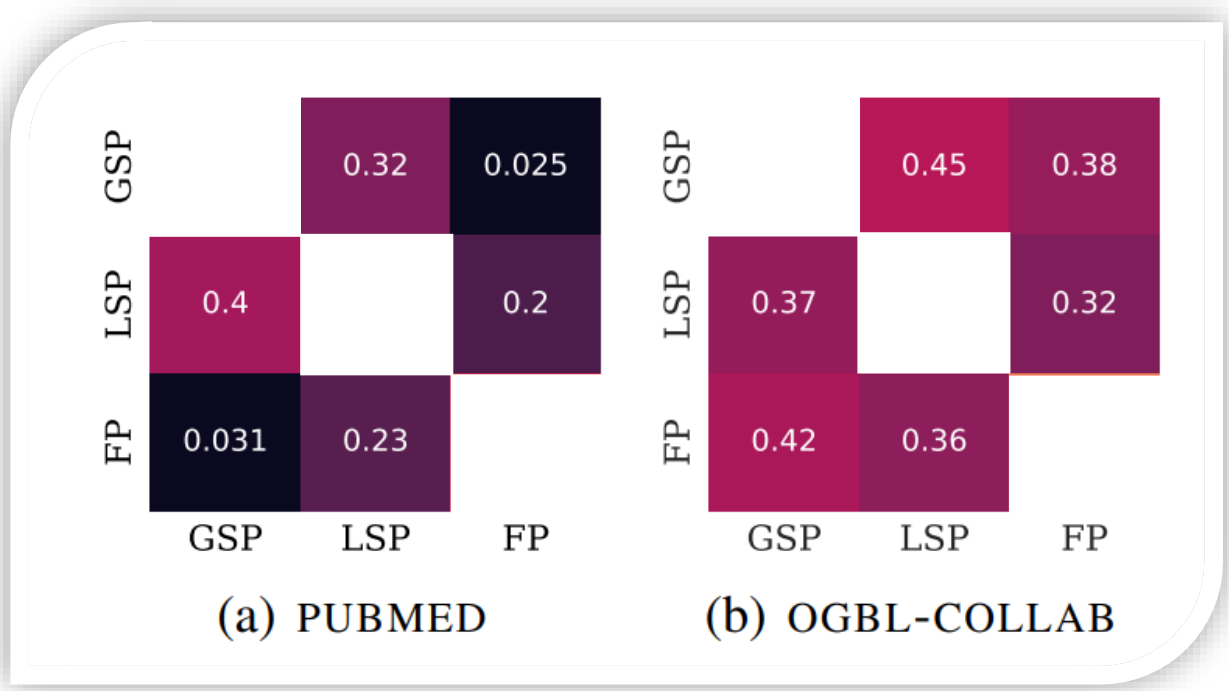


Complementary effects



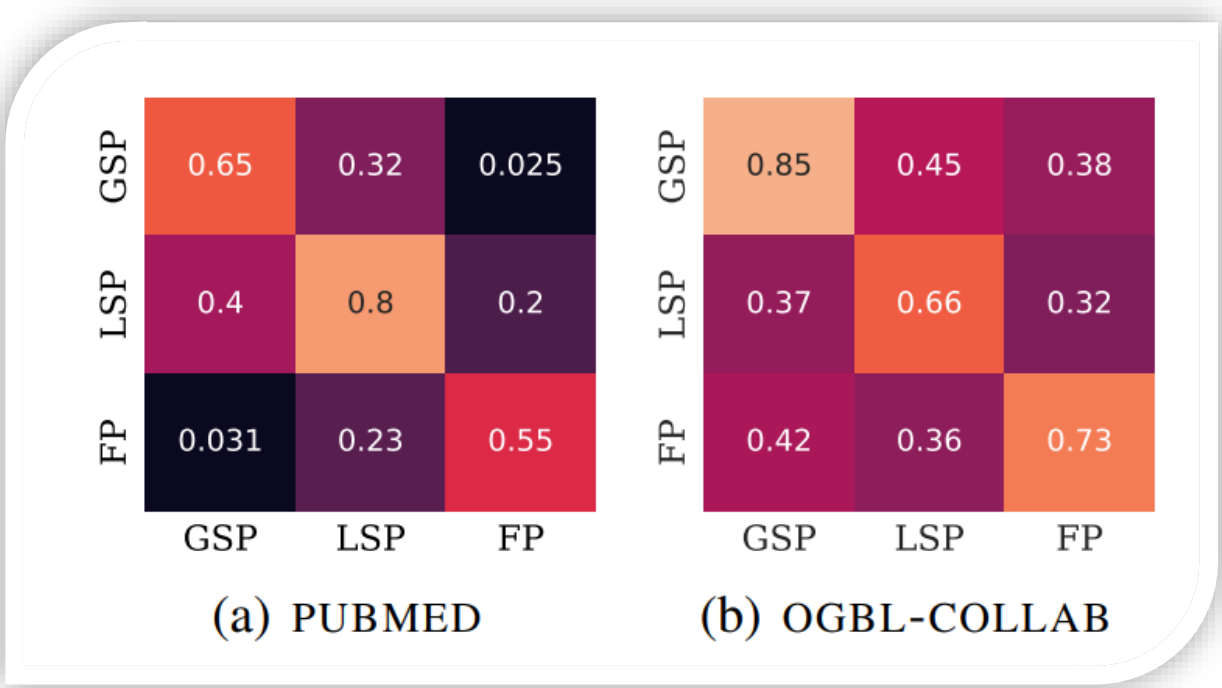
The prediction overlapping between heuristics from the same factor is large

Complementary effects



The prediction overlapping between heuristics from different data factors

Complementary effects



The prediction overlapping between heuristics from different factors is small

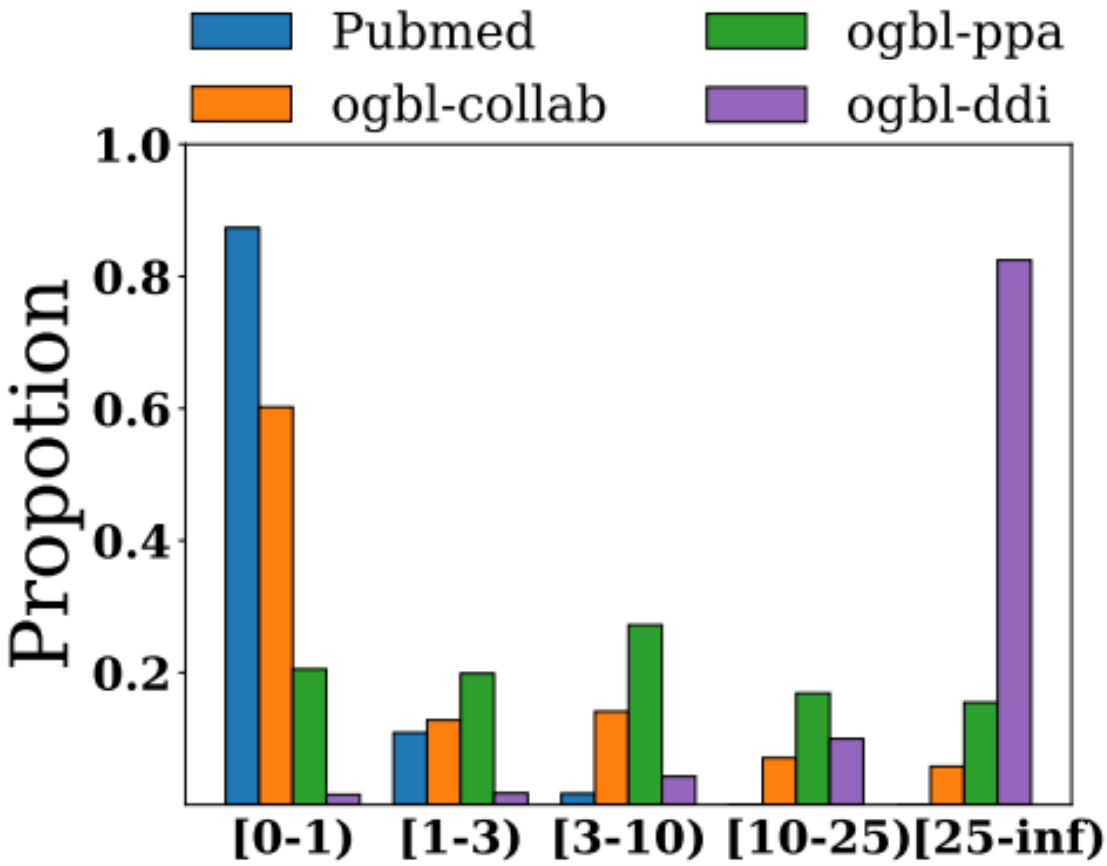
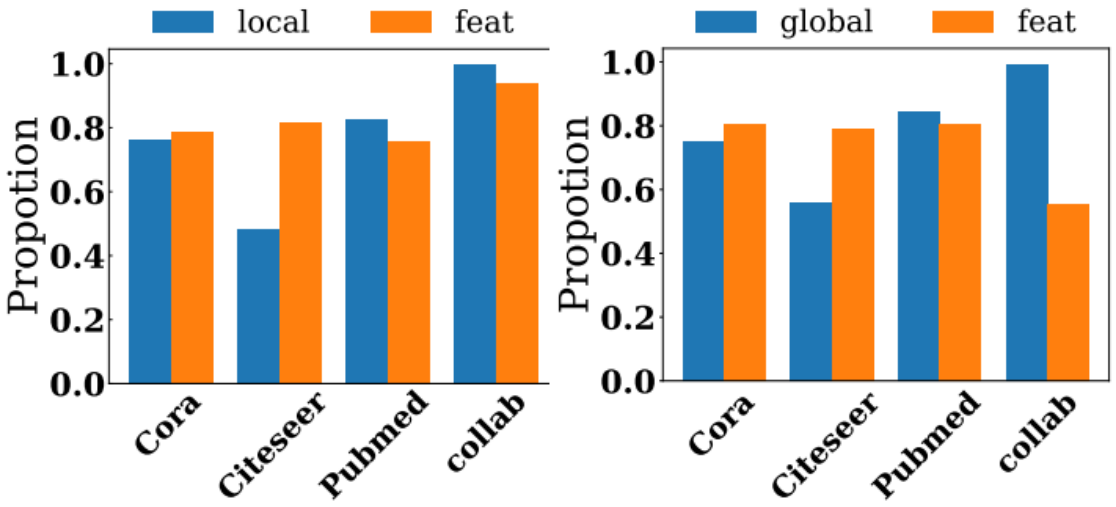


Figure 1: Distribution disparity of Common Neighbors across datasets.

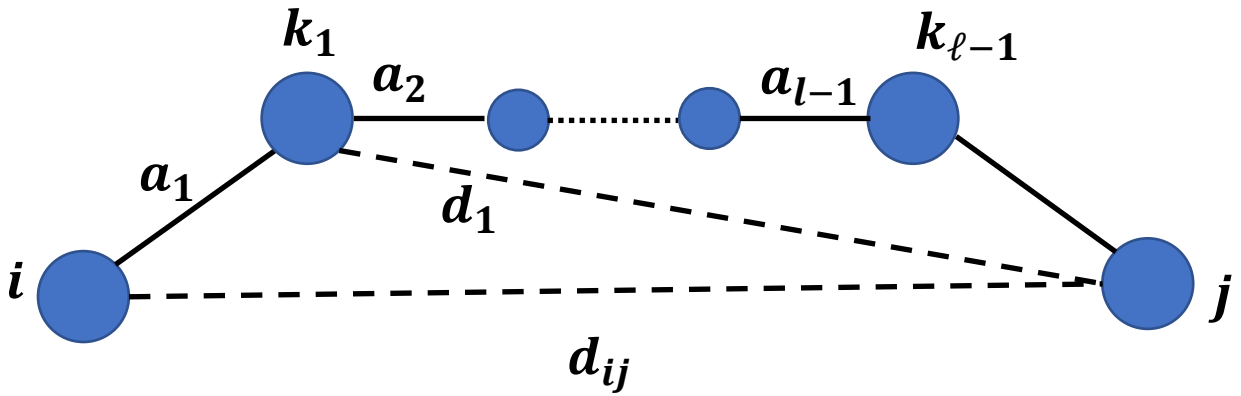
(c) OGBL-DDI

(d) OGBL-PPA



(a) Local (b) Global

Figure 4: Distinctions between FP heuristic and structural heuristics. Each bar depicts the proportion of node pairs accurately predicted by the heuristic on one factor but not on the other one.



$$
\begin{aligned}
\mathrm{P}_\ell(i,j) &= P\left(i \sim k_1 \sim \ldots \sim k_{\ell-1} \sim j \mid d_{ij}\right) \\
&= P\left(a_1 \leq r_1' \cap \ldots \cap a_\ell \leq r_{\ell-1}' \mid d_{ij}\right) \\
&= \int_{d_1,\ldots,d_{\ell-2}} P\left(a_1 \leq r_1', \ldots, a_{\ell-1} \leq r_{\ell-1}', d_1, \ldots, d_{\ell-2} \mid d_{ij}\right) \\
&= \int_{d_{\ell-2}=\left(d_{ij}-\sum_{n=0}^{\ell-3} r_n\right)_+}^{r_{\ell-1}+r_\ell} \cdots \int_{d_1=(d_{ij}-r_0)_+}^{\sum_{m=2}^{\ell} r_m} P\left(a_1 \leq r_1', d_1 \mid d_{ij}\right) \ldots P\left(a_{\ell-1} \leq r_{\ell-1}', a_\ell \leq r_\ell' \mid d_{\ell-2}\right) \\
&\leq \quad A\left(r_1', \sum_{m=2}^{\ell} r_m, d_{ij}\right) \times A\left(r_2', \sum_{m=3}^{\ell} r_m, (d_{ij}-r_0)_+\right) \times \ldots \times A\left(r_{\ell-1}', r_\ell, (d_{ij}-\sum_{n=0}^{\ell-3} r_n)_+\right) \\
&\leq \prod_{p=1}^{\ell-1} A\left(r_p', \sum_{m=p+1}^{\ell} r_m, \left(d_{ij}-\sum_{n=0}^{p-2} r_n\right)_+\right)
\end{aligned}
\tag{9}
$$

Table 3: Selected Datasets Statistics.

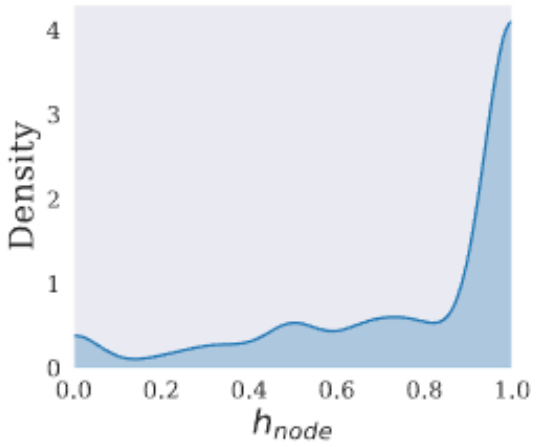
	Power	Reddit	Amazon Photo	Flicker
#Nodes	4,941	232,965	7,650	334,863
#Edges	6,594	114,615,892	238,162	899,756
#Feature	NA	602	745	500
Mean Degree	2.67	98.38	6.21	5.69
Split Ratio	80/10/10	80/10/10	80/10/10	80/10/10
Domains	Transport	Social	Web	Social

Table 4: Web Domain Datasets Statistics.

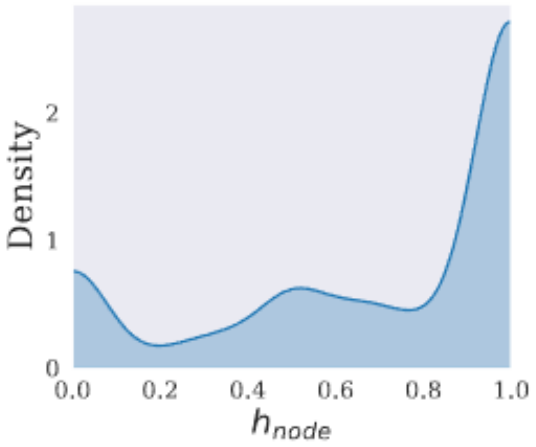
	PB	Email-Enron	Amazon Photo	Amazon	Google
#Nodes	1,222	36,692	7,650	334,863	875,713
#Edges	16,714	183,831	238,162	899,756	5,105,039
Mean Degree	27.36	10.02	6.21	5.69	116.49
Split Ratio	80/10/10	80/10/10	80/10/10	80/10/10	80/10/10
Domains	Web	Web	Web	Web	Web

Appendix for Node classification

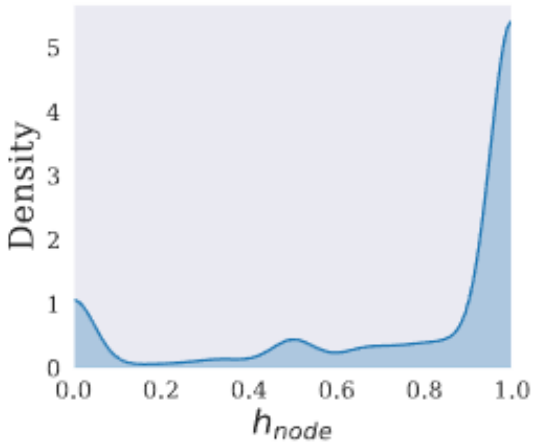
Homophily ratio distribution



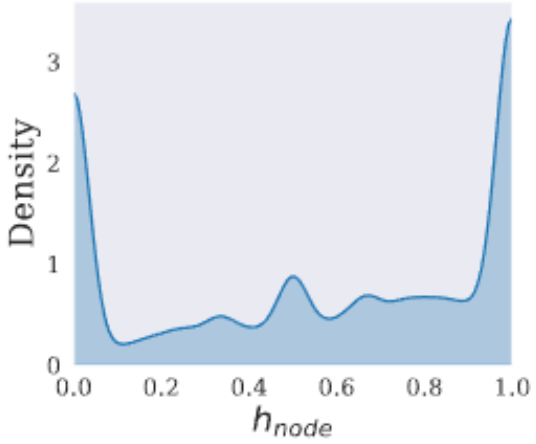
(a) Cora (h=0.81)



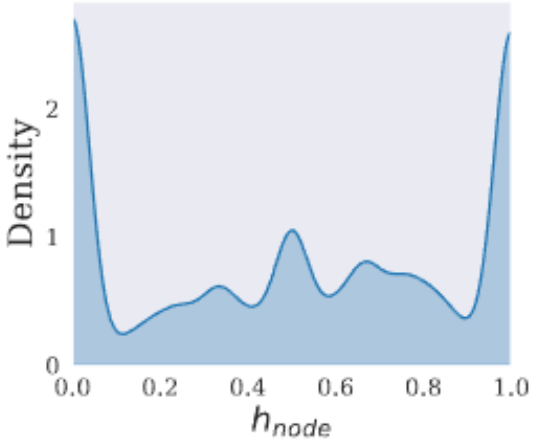
(b) CiteSeer (h=0.71)



(c) PubMed (h=0.79)

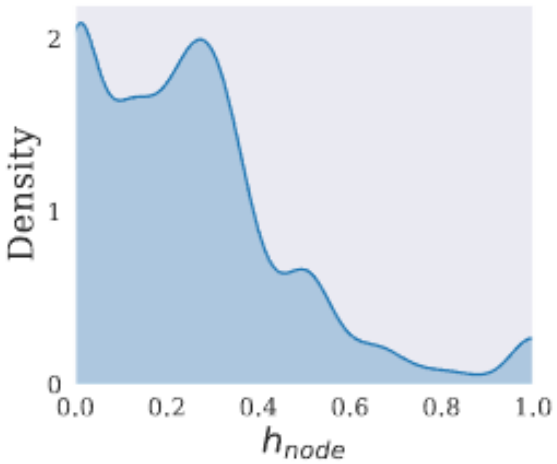


(d) Ogbn-Arxiv (h=0.63)

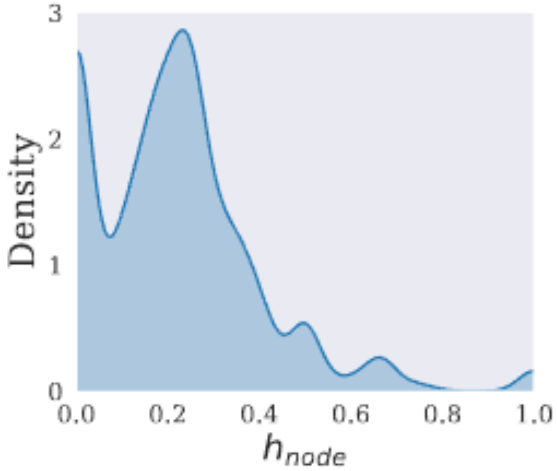


(e) IGB-tiny (h=0.57)

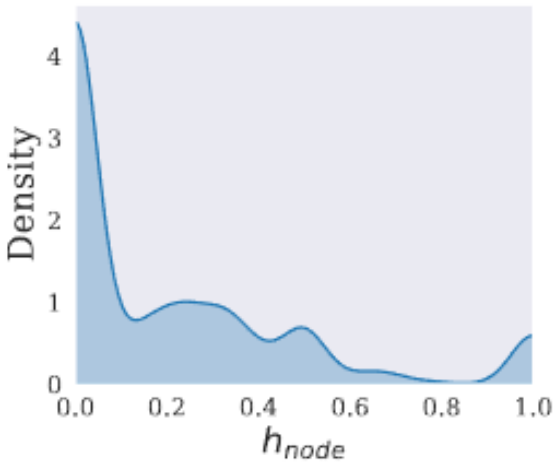
Homophily ratio distribution



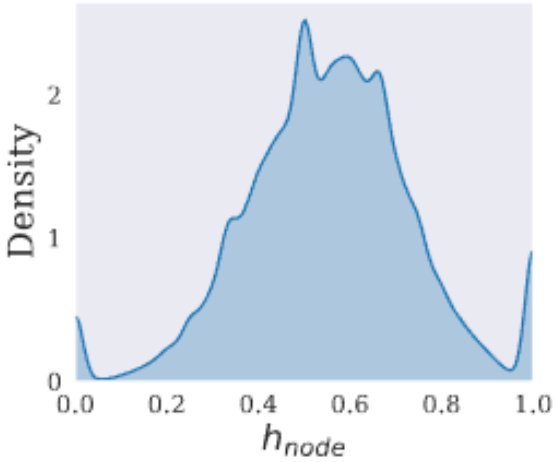
(f) Chameleon (h=0.25)



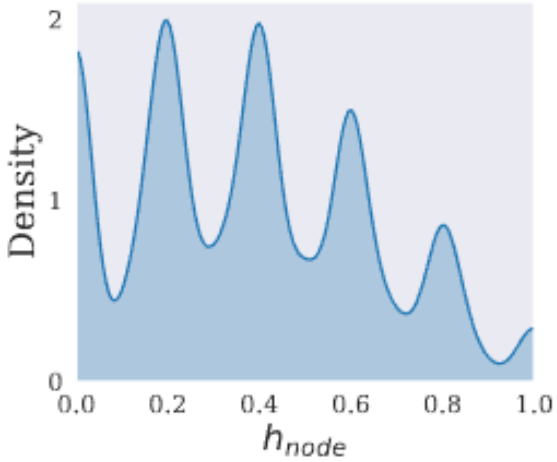
(g) Squirrel (h=0.22)



(h) Actor (h=0.22)



(i) Twitch-gamers (h=0.56)



(j) Amazon-ratings (h=0.38)

Model performance

Table 10: The accuracy of GNN and MLP models on homophilic graphs

Dataset	Cora	Citeseer	Pubmed	Arxiv	IGB-tiny
MLP	61.1±1.2	60.0±1.4	69.0±2.3	54.0±0.1	73.2±0.1
GLNN	81.3±1.5	73.0±2.7	78.2±2.6	71.7±0.1	73.2±0.1
GCN	81.5±1.4	73.7±1.6	77.9±2.0	71.4±0.1	70.7±0.1
SGC	81.7±1.4	72.7±2.2	77.0±2.7	68.0±0.1	71.0±0.1
GAT	82.2±1.1	73.6±1.6	77.3±1.5	71.0±0.1	70.8±0.2
APPNP	83.1±1.3	75.0±1.1	79.6±1.3	70.3±0.5	71.2±0.1
GCNII	82.8±1.1	73.8±1.7	79.0±2.5	71.7±0.5	73.5±0.1
GPRGNN	82.9±1.4	72.4±1.8	78.3±2.1	72.3±0.3	73.9±0.1

Table 11: The accuracy of GNN and MLP models on heterophilic graphs

Dataset	Chameleon	Squirrel	Twitch-gamers	Actor	Amazon-ratings
MLP	49.0±2.4	30.1±1.7	60.7±0.2	37.0±0.7	45.9±0.8
GLNN	39.2±2.7	52.3±1.4	61.1±0.1	37.3±1.0	54.0±0.7
GCN	68.0±2.0	54.7±1.4	62.2±0.2	30.7±0.9	49.0±0.6
SGC	69.1±1.8	53.0±1.1	62.0±2.0	30.0±1.5	46.5±0.6
GAT	67.0±1.9	53.2±1.7	59.9±0.3	30.7±1.0	48.0±0.5
APPNP	56.7±2.5	42.4±1.9	59.7±0.1	37.0±1.3	44.9±0.8
GCNII	64.7±1.8	44.0±1.5	64.5±0.3	36.0±1.2	50.0±0.5
GPRGNN	68.5±1.4	53.8±1.4	61.9±0.2	36.5±1.4	49.8±0.5

Data and model Assumption

Definition 1 (CSBM-S$(\mu_1, \mu_2, (p^{(1)}, q^{(1)}), (p^{(2)}, q^{(2)}), \Pr(\text{homo}))$). *The generated nodes consist of two disjoint sets $\mathcal{C}_1$ and $\mathcal{C}_2$. each node feature x is sampled from $N(\mu_i, I)$ with $i \in \{1, 2\}$. Each set $\mathcal{C}_i$ consists of two subgroups: $\mathcal{C}_i^{(1)}$ for nodes in homophilic pattern with intra-class and inter-class edge probability $p^{(1)} > q^{(1)}$ and $\mathcal{C}_i^{(2)}$ for nodes in heterophilic pattern with $p^{(2)} > q^{(2)}$. $\Pr(homo)$ denotes the probability that the node is in homophilic pattern. $\mathcal{C}_i^{(j)}$ denotes node in class i and subgroup j with $(p^{(j)}, q^{(j)})$. We assume nodes follow the same degree distribution with $p^{(1)} + q^{(1)} = p^{(2)} + q^{(2)}$.*

Definition 2 (Generalized CSBM-S model). *Each node subgroup V_m follows the CSBM distribution $V_m \sim CSBM(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, p^{(i)}, q^{(i)})$, where different subgroups share the same class mean but different intra-class and inter-class probabilities $p^{(i)}$ and $q^{(i)}$. Moreover, node subgroups also share the same degree distribution as $p^{(i)} + q^{(i)} = p^{(j)} + q^{(j)}$.*

Assumption 1 (GNN model). *We focus on SGC [15] with the following components: (1) a one-hop mean aggregation function g with $g(X, G)$ denoting the output. (2) MLP feature transformation $f(g_i(X, G); W_1, W_2, \cdots, W_L)$, where f is a ReLU-activated L-layer MLP with $W_1, \cdots, W_L$ as parameters for each layer. The largest width of all the hidden layers is denoted as b.*

Addition theoretical analysis on linear seperability

Lemma 2 (Linear separability on nodes with the same structural patterns). *Considering mean aggregated features are from the same structural pattern* $\mathbf{f}_i^{(j)}$, *for* $i \in \{1, 2\}$. *For any node* i, *the largest-margin linear classifier on* $\mathbf{f}_i^{(j)}$ *will have a lower probability to misclassify than* $\mathbf{x}_i$, *when* $d_i > \frac{(p^{(1)}+q^{(1)})^2}{(p^{(1)}-q^{(1)})^2}$

When $p^{(1)} = 0.9, q^{(1)} = 0.1$:

Improved linear separability can be found with $d_i > 1.75$

Lemma 3 (Linear separability on nodes with different structural patterns). *Consider features are from different structural patterns, where* $\mathbf{f}_i^{(1)}$ *for* $i \in \mathcal{C}_1$ *and* $\mathbf{f}_i^{(2)}$ *for* $i \in \mathcal{C}_2$. *For any node* i, *the largest-margin linear classifier will have a lower probability to misclassify* $\mathbf{f}_i^{(1)}$ *for* $i \in \mathcal{C}_1$ *and* $\mathbf{f}_i^{(2)}$ *for* $i \in \mathcal{C}_2$ *than* $\mathbf{x}_i$ *when* $d_i > \frac{(p^{(1)}+q^{(1)})^2}{(p^{(1)}-q^{(2)})^2}$

When $p^{(2)} = 0.2, q^{(2)} = 0.8$,

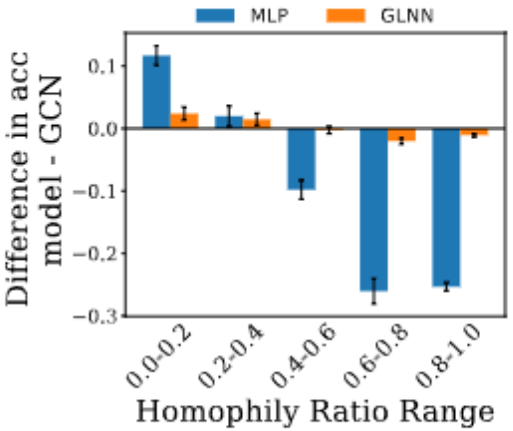
Improved linear separability can be found with $d_i > 100$

Empirical verfication

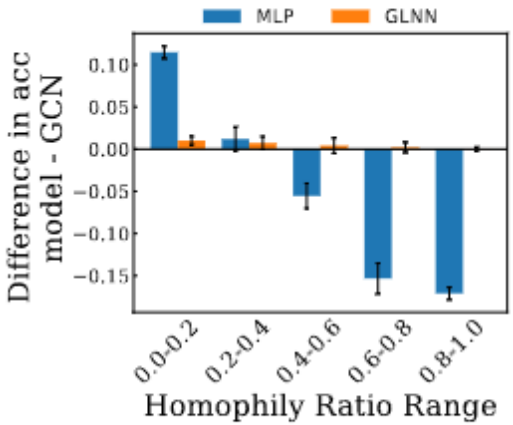
Table 6: The performance of logistic regression algorithm on homophilic nodes, heterophilic nodes, and a mixture of homophilic and heterophilic nodes. The results on the first row and first column correpond to the performance on homophilic nodes and heterophilic nodes, solely.

Hete\Homo	-	p=0.01, q=0.005	p=0.01, q=0.003	p=0.01, q=0.001
-	-	74.68±3.19	82.71±1.86	92.08±1.13
p=0.001, q=0.005	79.64±2.11	60.84±0.64	62.08±0.59	81.38±1.02
p=0.001, q=0.003	70.08±1.71	59.72±2.01	61.58±1.08	76.60±0.98
p=0.001, q=0.002	62.08±3.04	65.92±1.95	69.42±1.03	74.16±1.09

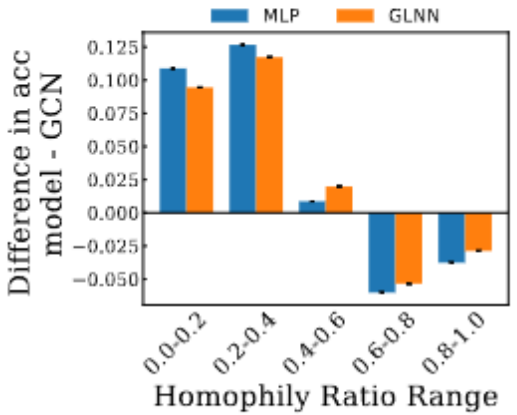
Additional comparison between GCN and MLP



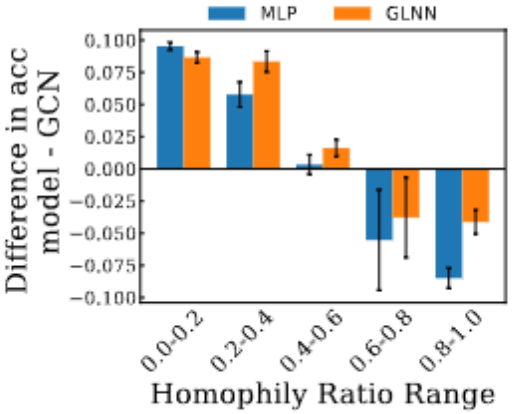
(a) Cora (h=0.81)



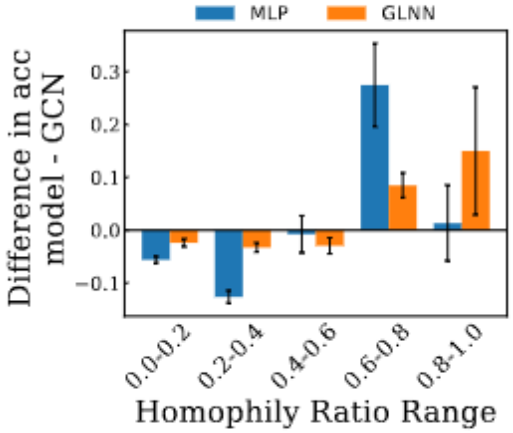
(b) CiteSeer (h=0.71)



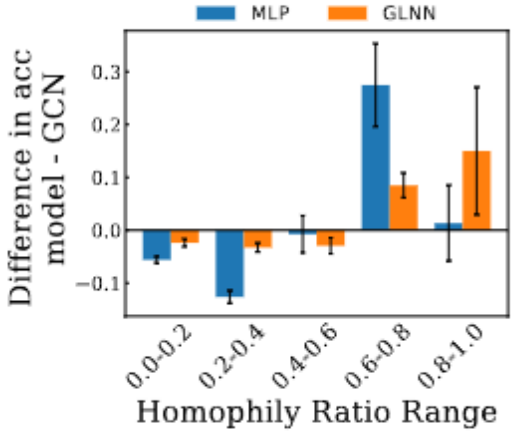
(c) IGB-tiny (h=0.58)



(d) Actor (h=0.22)



(e) twitch-gamers (h=0.56)



(f) Amazon-ratings (h=0.38)

Additional synthetic analysis

nodes with low homophily

Adding K intra-class edges

nodes with high homophily

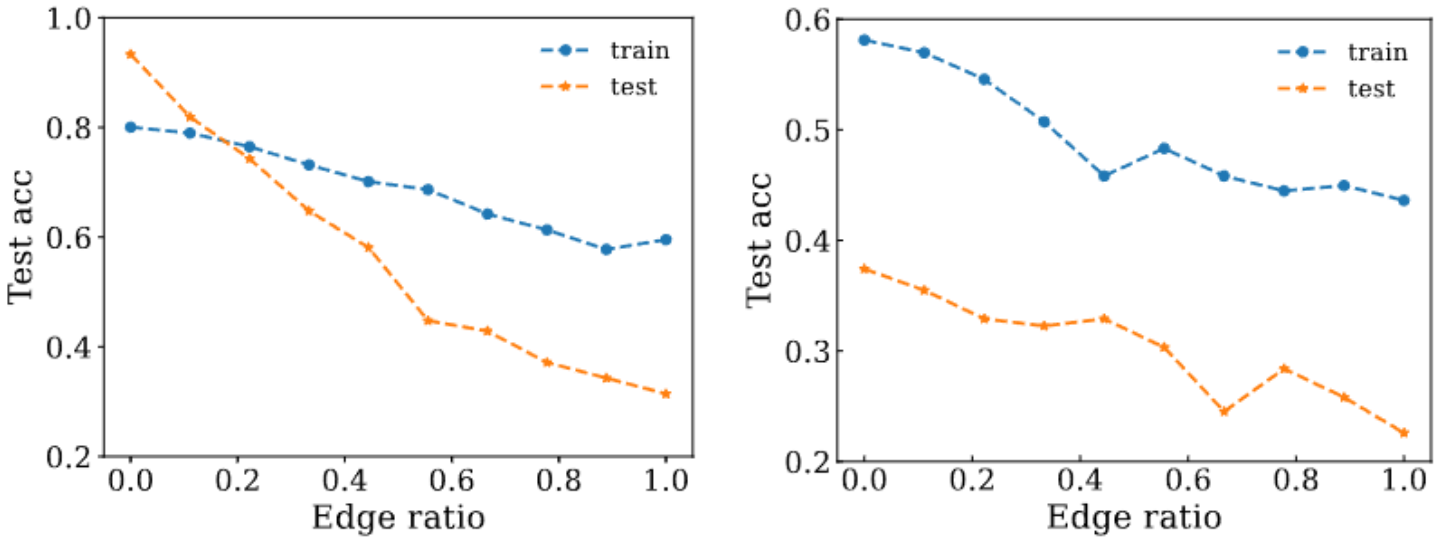
$$\mathcal{D}_0 : \text{Categorical}([0, 0.5, 0, 0, 0, 0.5]),$$
$$\mathcal{D}_1 : \text{Categorical}([0.5, 0, 0.5, 0, 0, 0]),$$
$$\mathcal{D}_2 : \text{Categorical}([0, 0.5, 0, 0.5, 0, 0]),$$
$$\mathcal{D}_3 : \text{Categorical}([0, 0, 0.5, 0, 0.5, 0]),$$
$$\mathcal{D}_4 : \text{Categorical}([0, 0, 0, 0.5, 0, 0.5]),$$
$$\mathcal{D}_5 : \text{Categorical}([0.5, 0, 0, 0, 0.5, 0]).$$

nodes with high homophily

Adding K across-class edges

nodes with low homophily

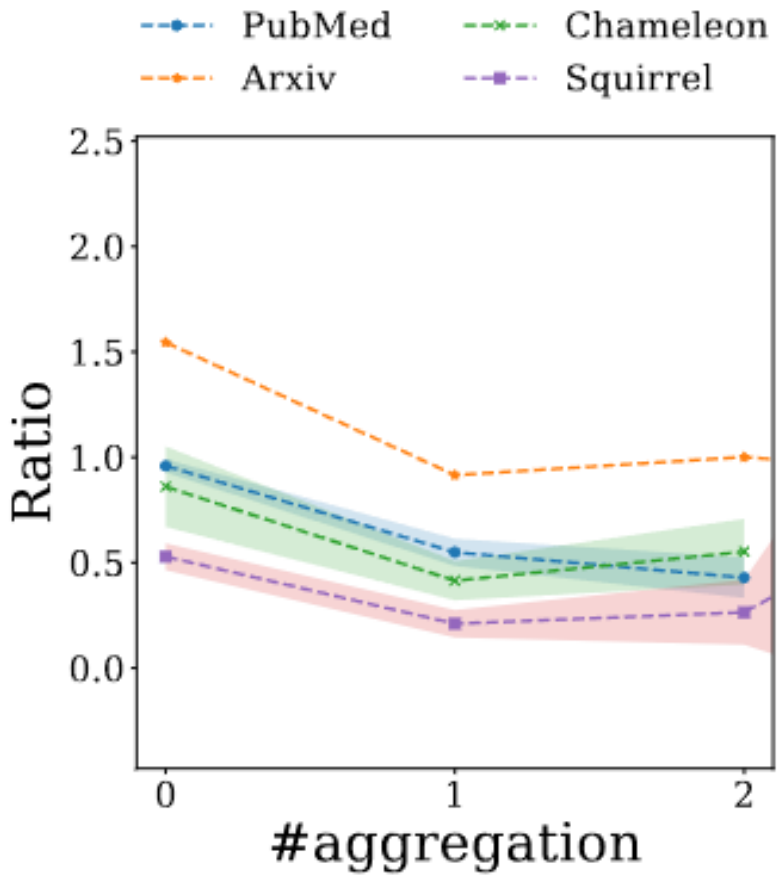
Additional synthetic analysis



(a) Synthetic graphs generated from Cora with the targeted heterophilic edge algorithm

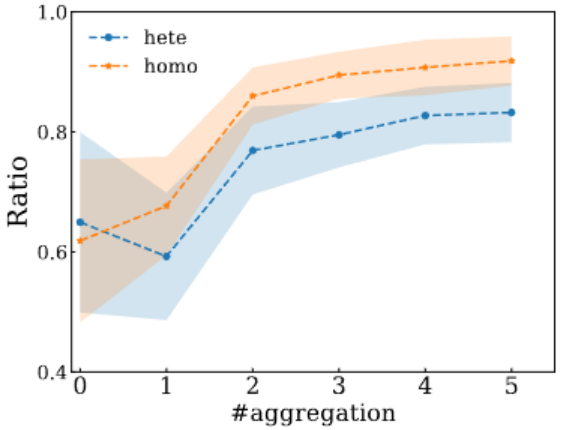
(b) Synthetic graphs generated from Squirrel with the targeted homophilious edge algorithm

Additional discriminative analysis on GCN

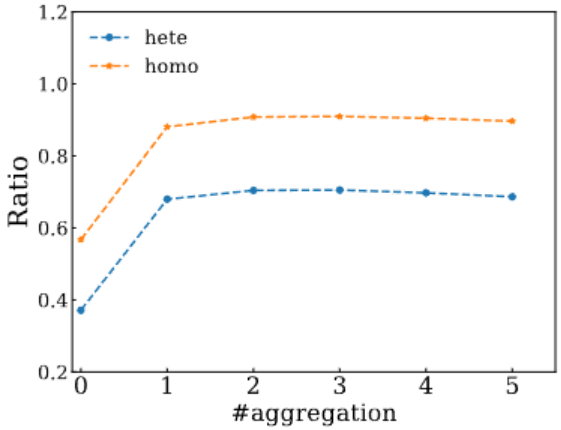


Additional local discriminative analysis

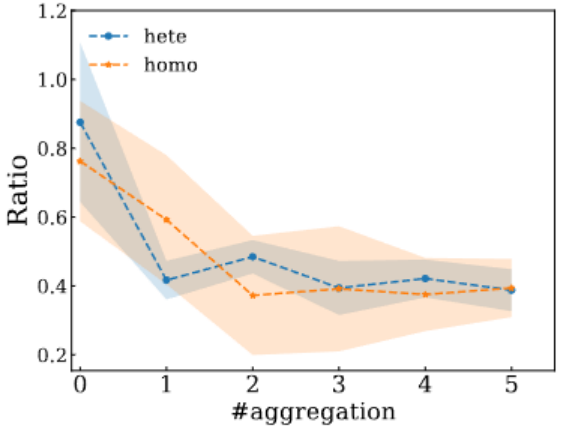
$$r = \frac{\sum_{v \in V_{\text{te}}} \mathbb{1}\left[\exists c \in \mathcal{C}, |\mathcal{M}_v^c| > \frac{k}{2}\right]}{|V_{\text{te}}|}$$



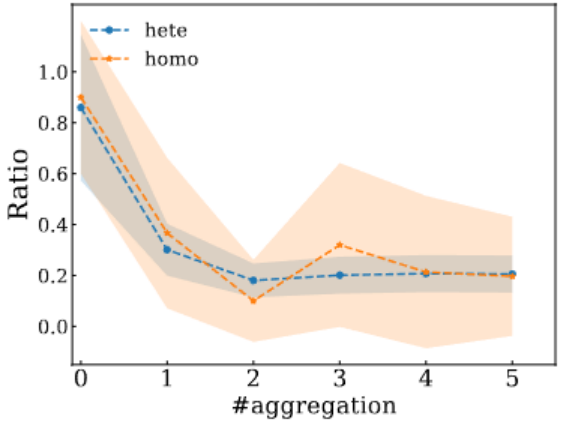
(c) PubMed (h=0.79)



(d) Ogbn-arxiv (h=0.63)



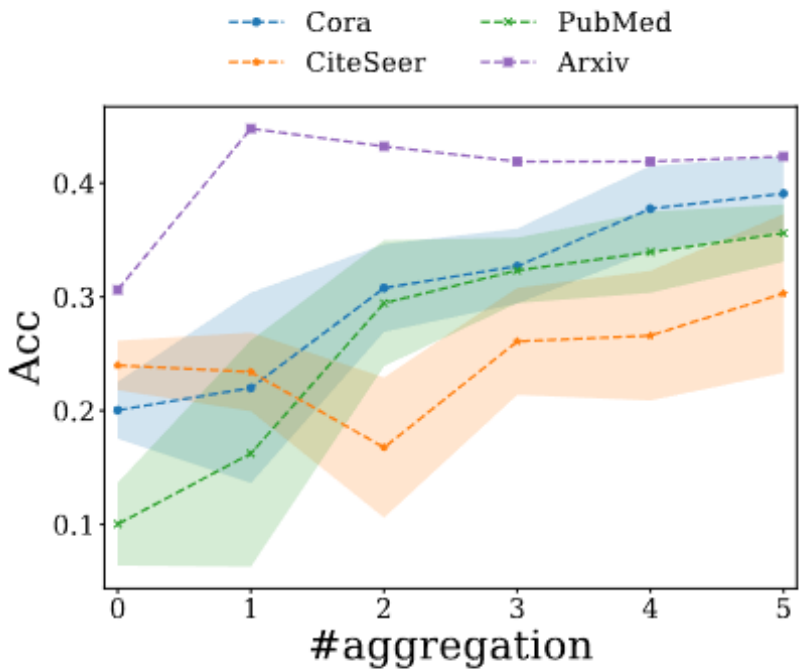
(e) Chameleon (h=0.25)



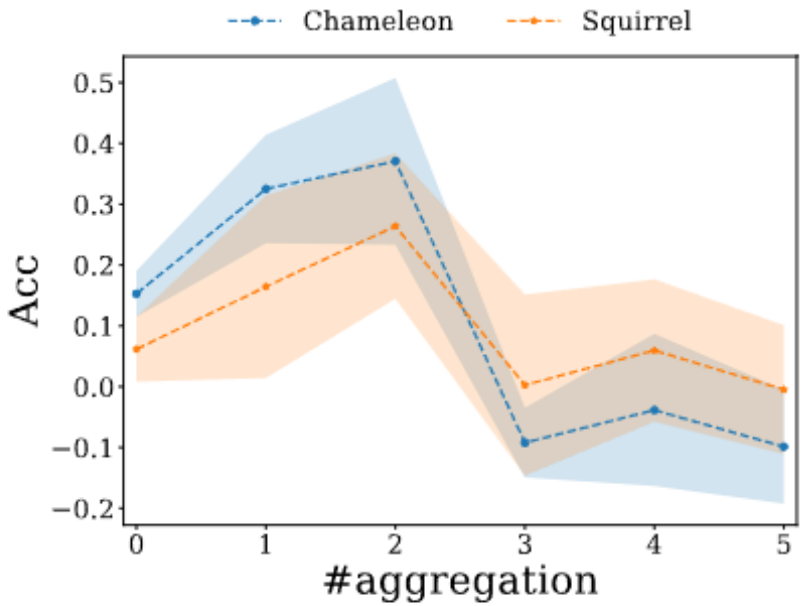
(f) Squirrel (h=0.22)

Additional local discriminative analysis

$$\text{Acc}_{\text{local}} = \frac{\sum_{v \in V_{\text{agree}}} \mathbb{1}\left[c_v = c_{\mathcal{N}_v}\right]}{|V_{\text{agree}}|}$$



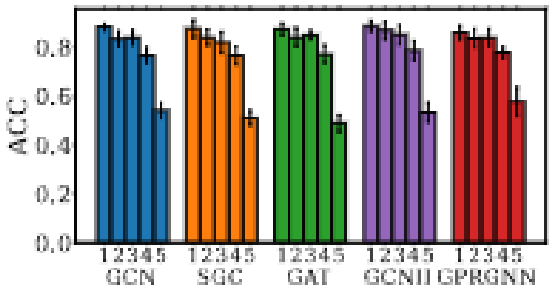
(a) Homophilic graphs



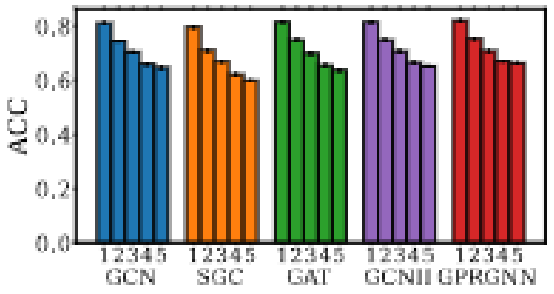
(b) Heterophilic graphs

Additional higher-order performance disparity

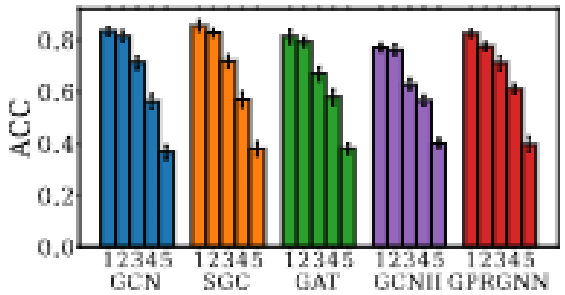
$$s = \epsilon_m + |h_{\text{tr}} - h_m|$$



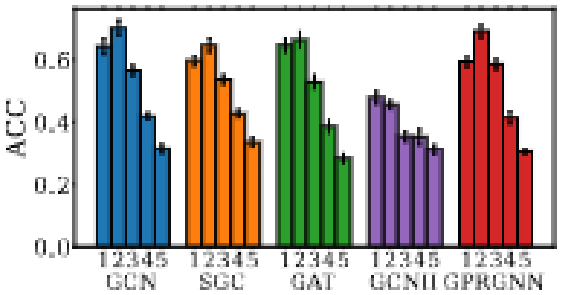
(a) PubMed



(b) Ogbn-arxiv



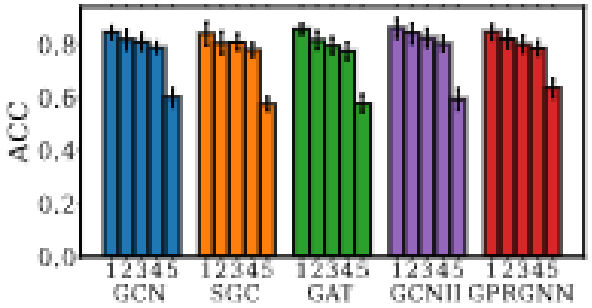
(c) Chameleon



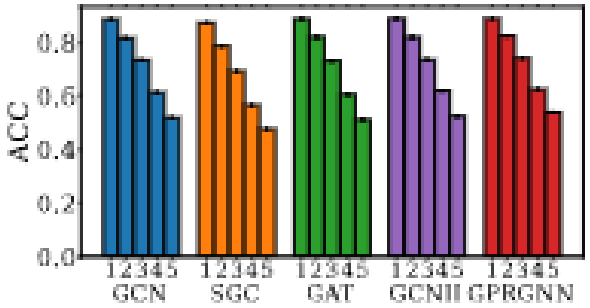
(d) Squirrel

Additional higher-order performance disparity

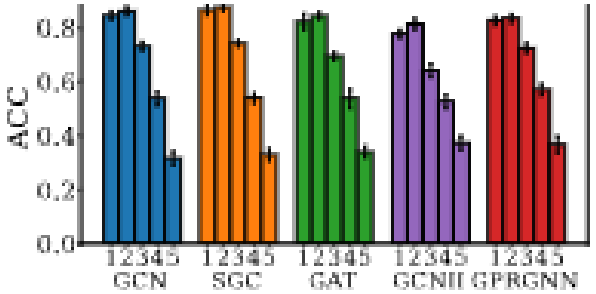
$$s = \epsilon_m$$



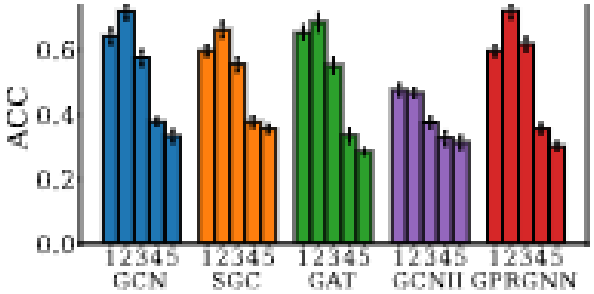
(a) PubMed



(b) Ogbn-arxiv



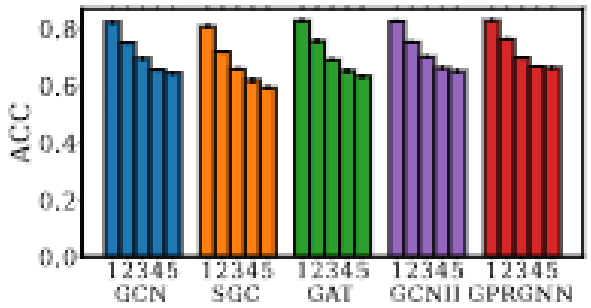
(c) Chameleon



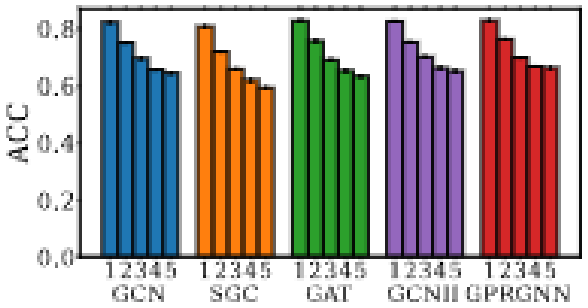
(d) Squirrel

Additional higher-order performance disparity

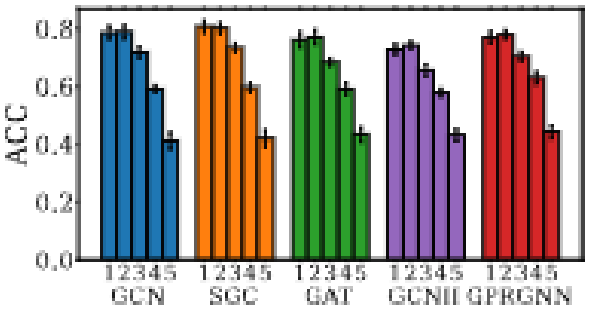
$$s = |h_{\text{tr}} - h_m|$$



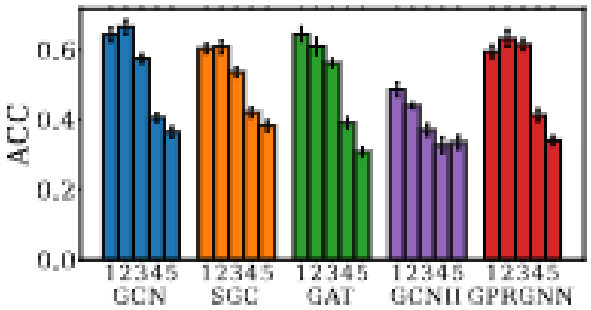
(a) PubMed



(b) Ogbn-arxiv



(c) Chameleon



(d) Squirrel

OOD statistics

Table 14: the numbers of train, validation, test nodes on OOD data split

Dataset	Cora	CiteSeer	PubMed	Arxiv	Squirrel	Chameleon
#train	1599	1160	12466	85788	3709	1642
#valid	400	290	3117	21447	928	441
#test	486	660	4134	62108	564	564

No observe covariance shift

Table 15: MMD distance between train and validation, test sets on both i.i.d. and ood settings.

Dataset	Cora	CiteSeer	PubMed	Arxiv	Chameleon	Squirrel
IID valid	0.565	0.345	0.082	0.149	0.951	1.04
IID test	0.610	0.600	0.050	0.276	0.882	0.92
OOD valid	0.564	0.233	0.127	0.211	0.977	1.192
OOD test	0.597	0.598	0.442	0.420	0.854	0.92

Table 17: Train and test homophily ratios on the OOD datasets in [61]

	Twitch-explicit	FaceBook-100	Ogb-arxiv	elliptic	Cora	Amazon-photo
Train Homo	0.53	0.18	0.38	0.12	0.69	0.90
Test Homo	0.53	0.54	0.42	0.57	0.69	0.90