

When do Graph Neural Networks work and when not?

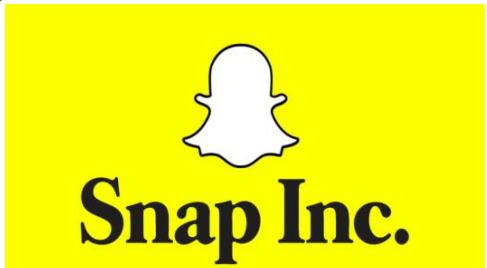
Haitao Mao

Department of Computer Science and Engineering

Michigan State University

haitaoma@msu.edu

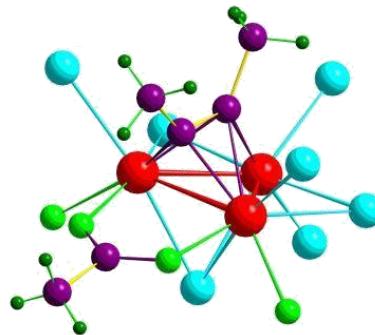
June, 2023



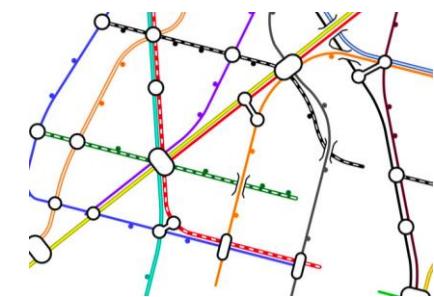
Graph can represent arbitrary data



Social Network



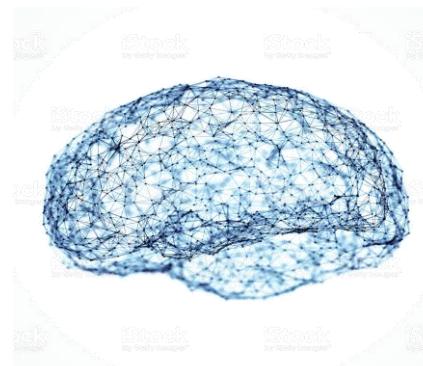
Molecules



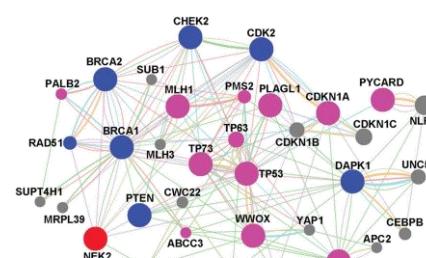
Transportation Networks



Web

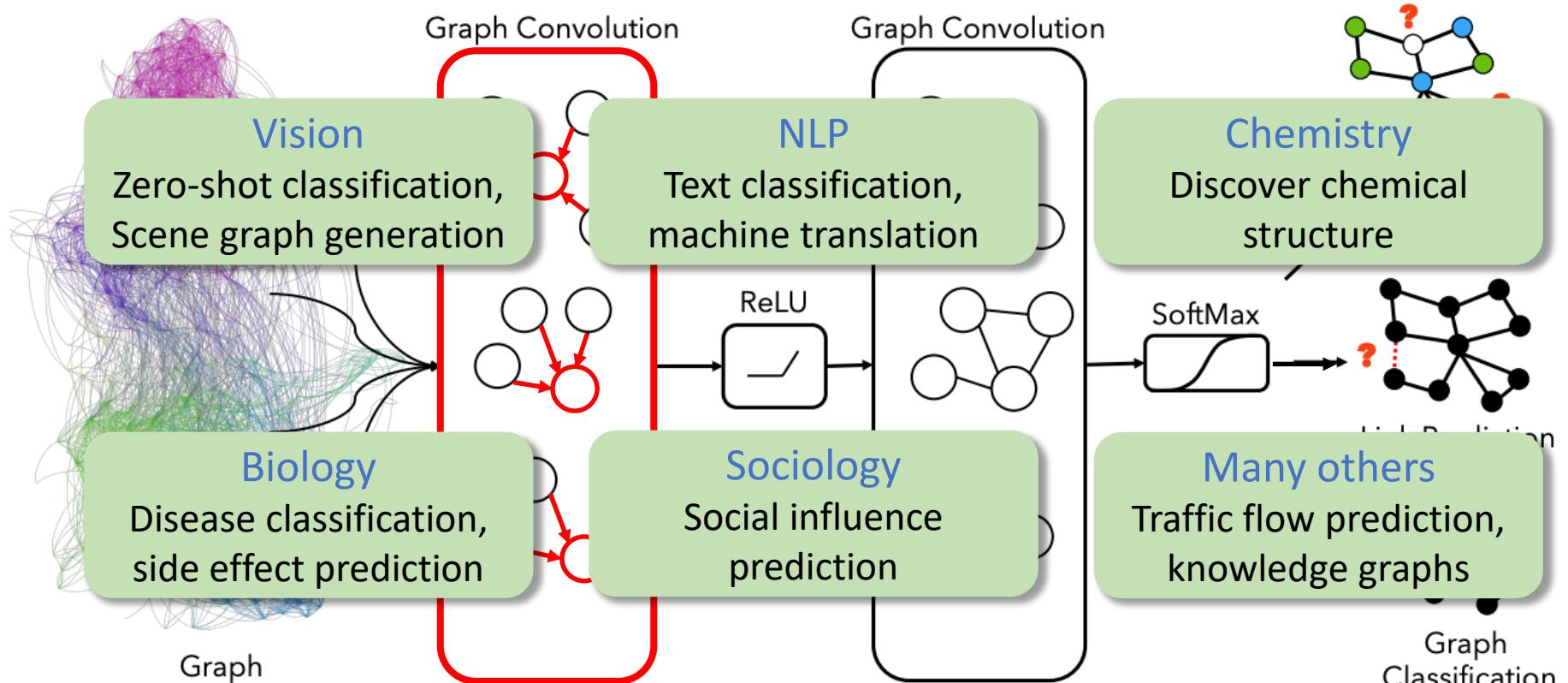


Brain Networks



Gene Interactions

Graph Neural Networks (GNNs)



For a given problem on graphs, can GNNs work?

Image credit: Swapnil Gandhi

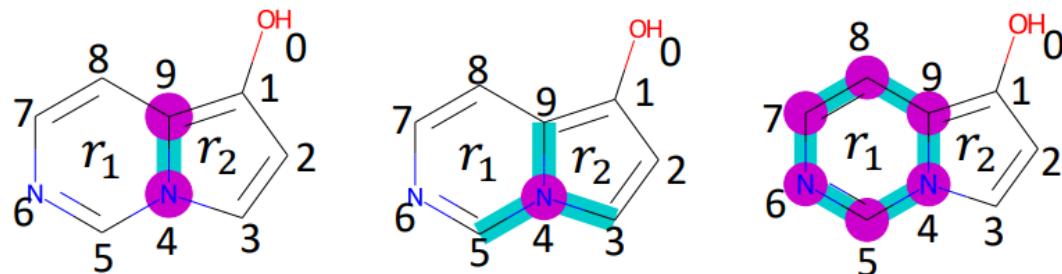
Can one GNN work for all graphs?

No, there is no consistent winning solution!

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

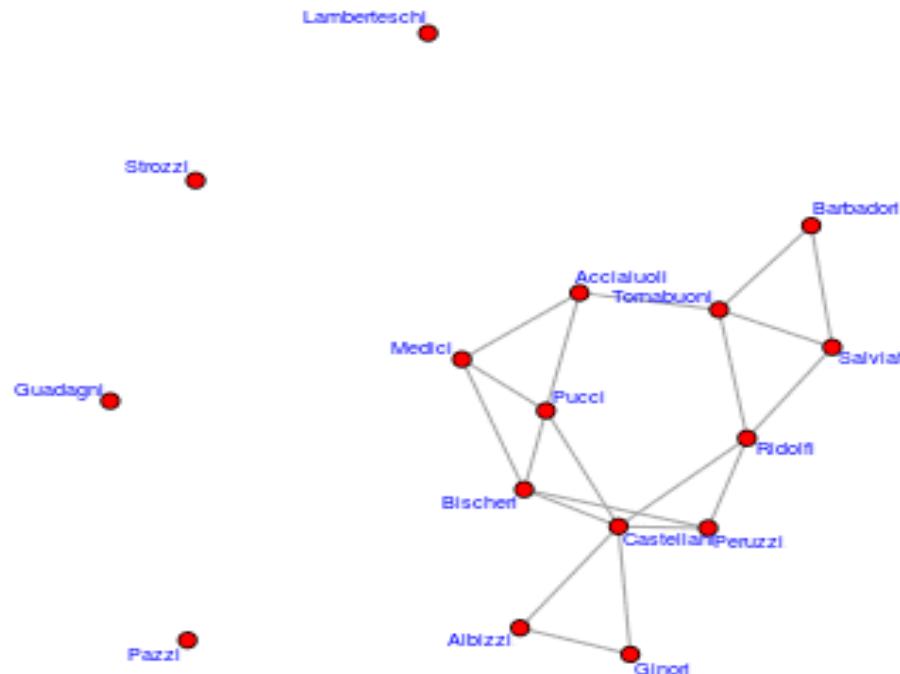


Can one GNN work for all graphs?

Data
mechanism

How is a link
formed in a graph?

Different graphs may have different
data formulation principles



Can one GNN work for all graphs?

Model
mechanism

⊗ Match!

Data
mechanism

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 perform well and when not?

Two tasks:

- Link Prediction
- Node Classification

Two mechanisms:

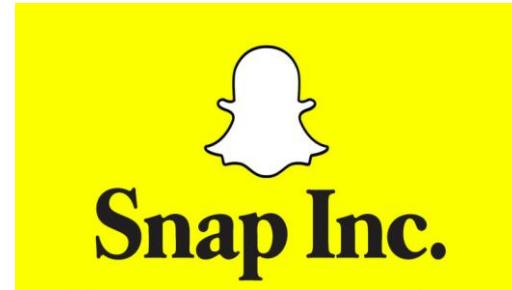
- Data mechanism
- Model mechanism

Multiple Applications:

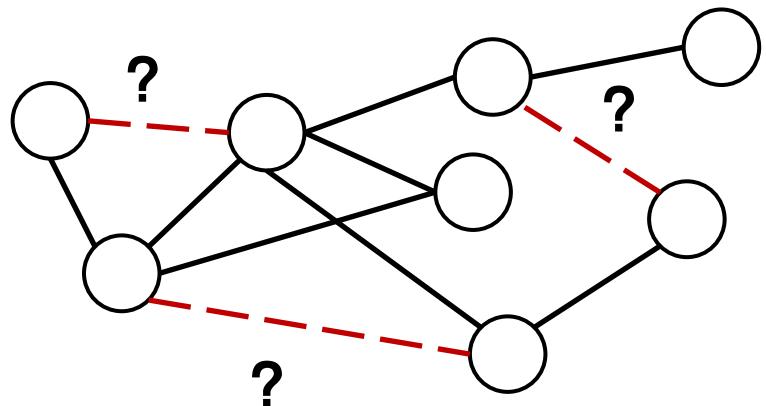
- New challenging scenarios
- Instruction for model design
-

Revisiting Link Prediction: a Data Perspective

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

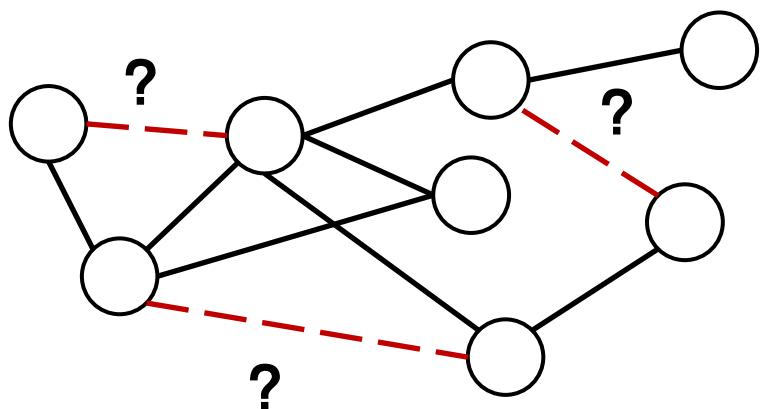


Link Prediction task

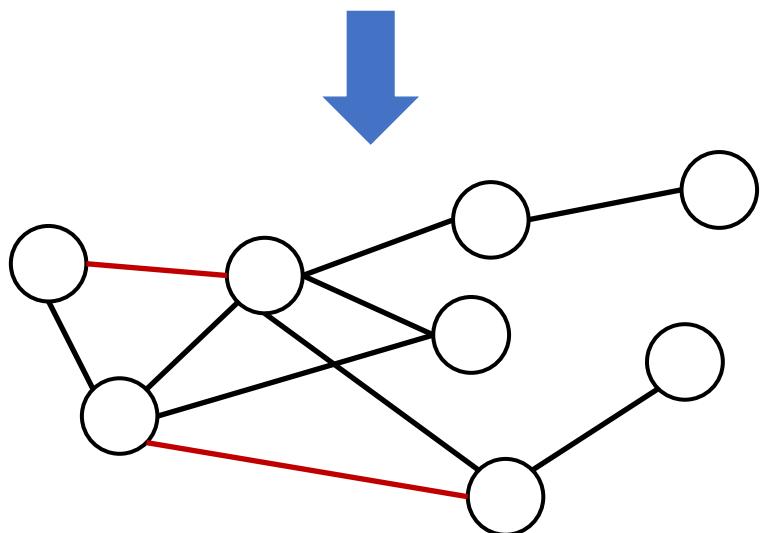


Predict the likelihood of
whether two nodes connected

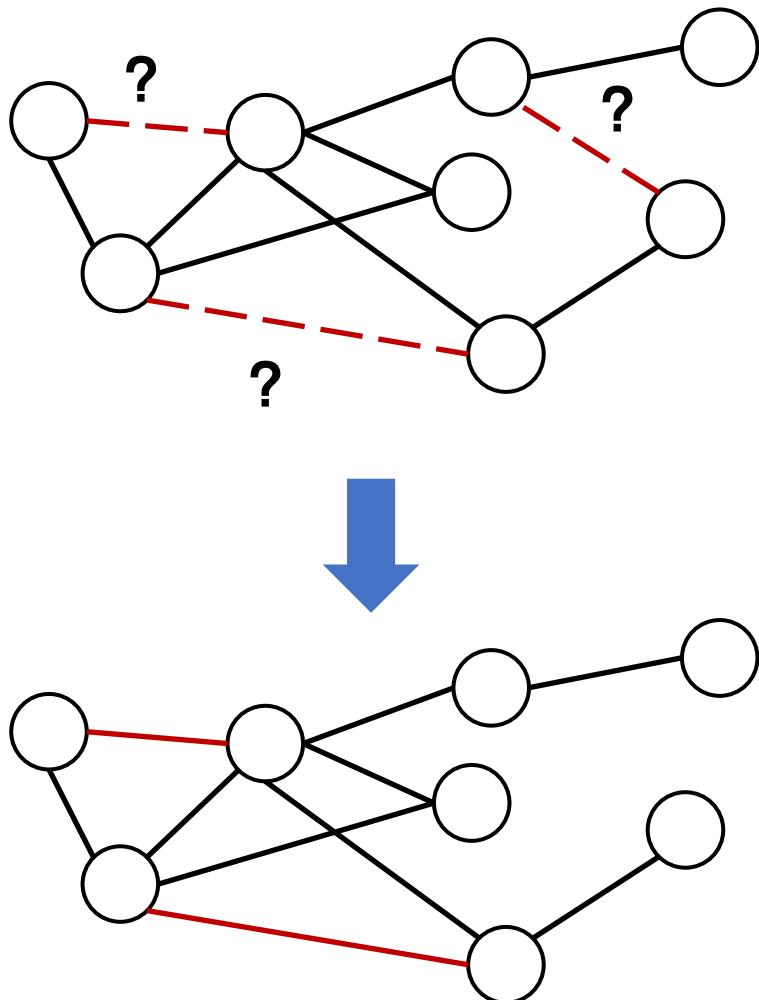
Link Prediction task



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

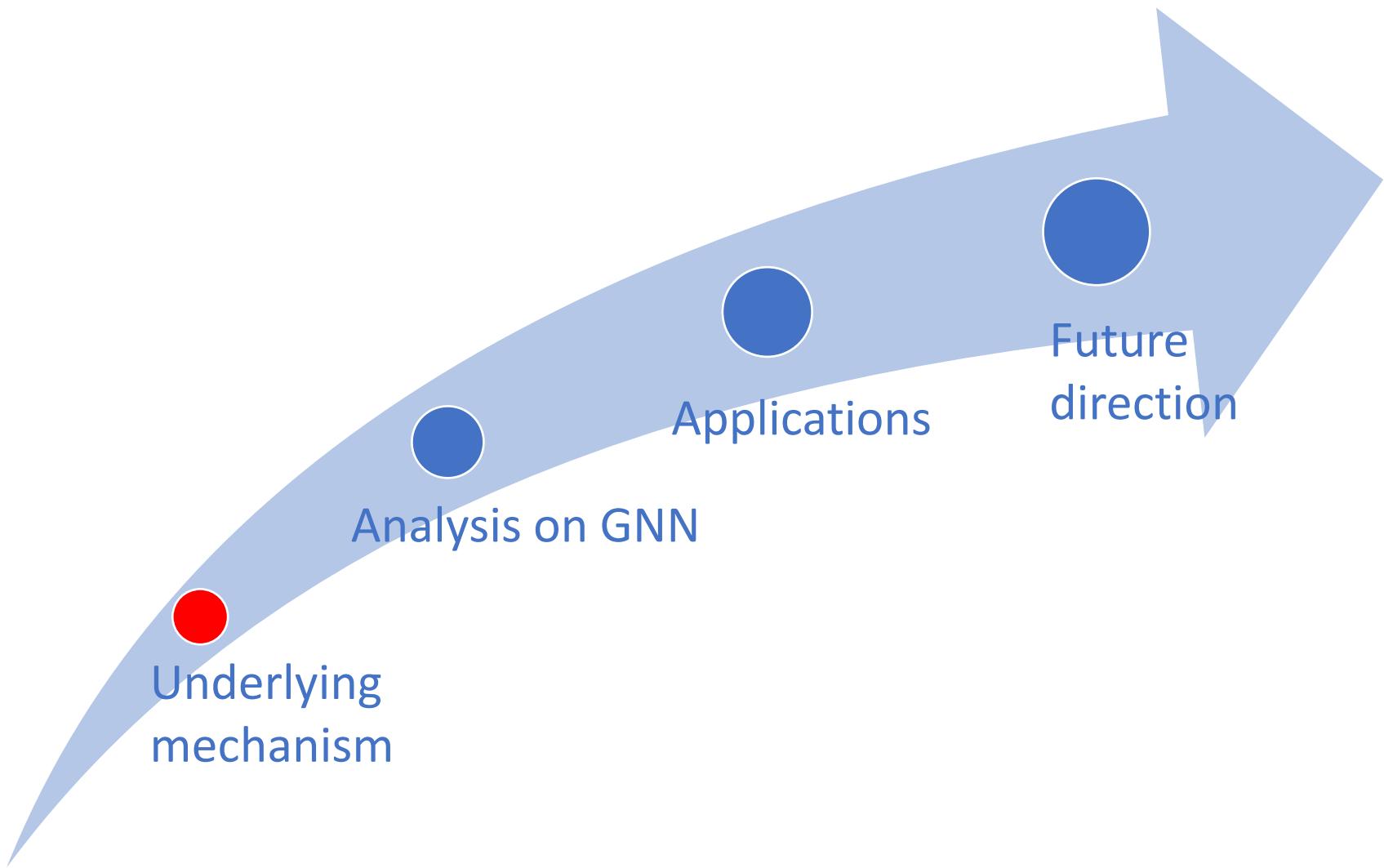
Table 1: Results on Cora, Citeseer, and Pubmed(%) under the existing evaluation setting. Highlighted are the results ranked **first**, **second**, and **third**.

	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, no consistent winning solution

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

Outline



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

- ❑ Heuristic algorithms



- ❑ Graph Neural Networks for Link Prediction (GNN4LP)

Different GNNs for Link Prediction incorporate different pairwise heuristics as the inductive bias.

Heuristic algorithms

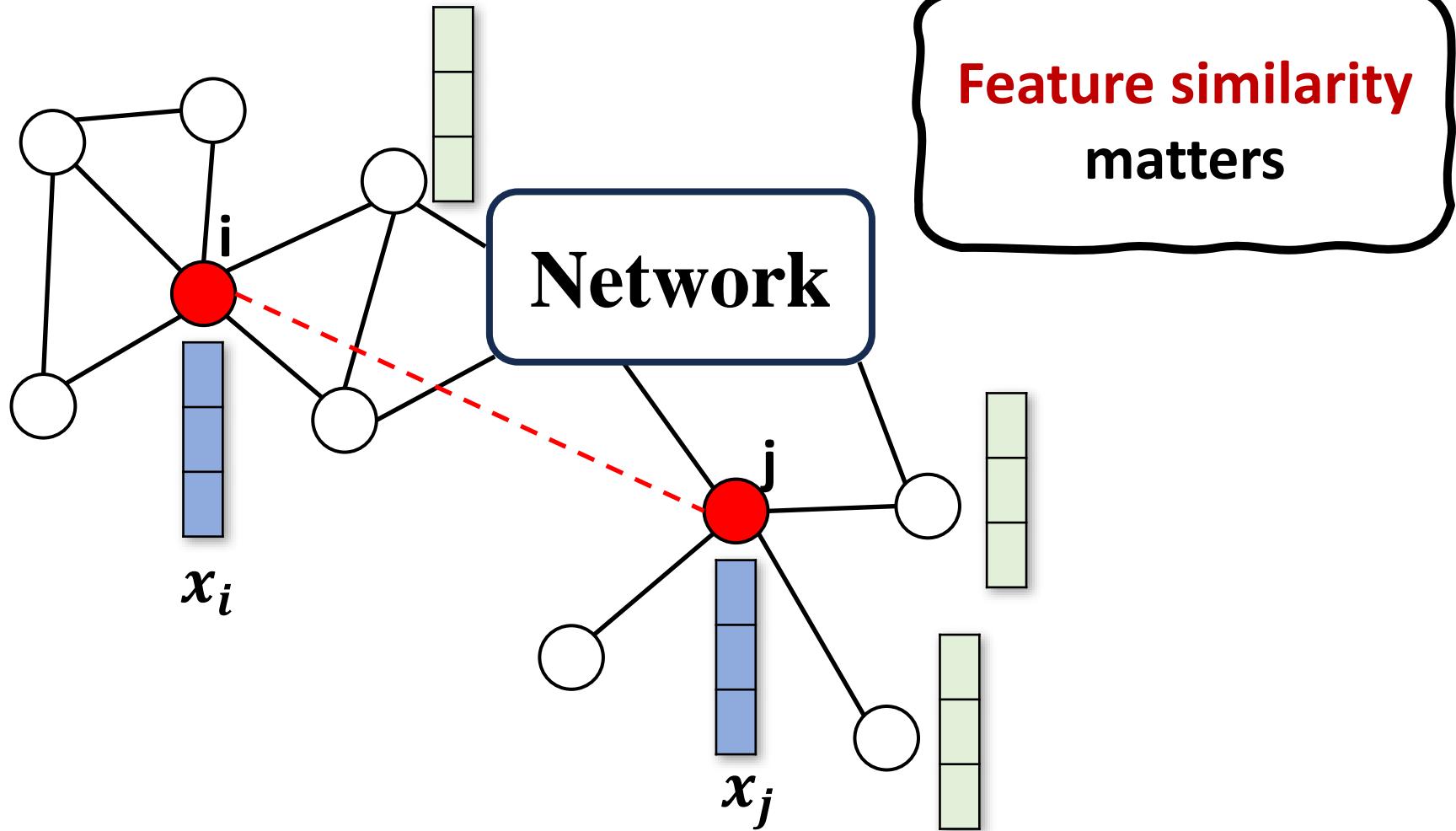
Assign a score to each node pair based
on specific heuristic function

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

- Interpretable
- Simple
- Derived from expert knowledge

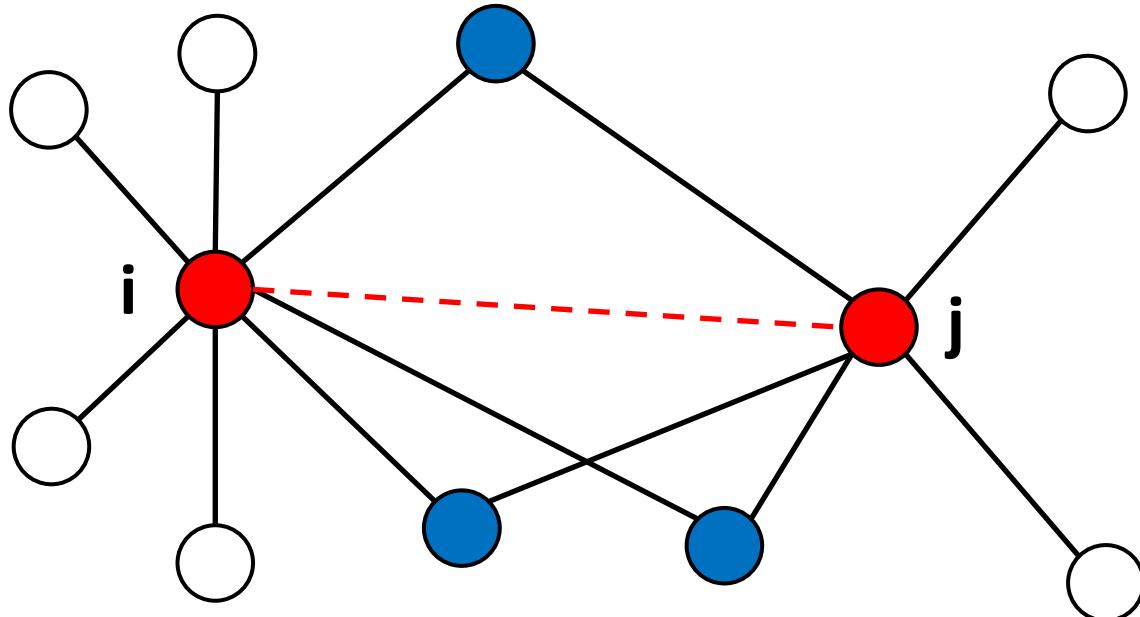
Heuristic algorithms

Feature homophily: $dis(x_i, x_j)$



Heuristic algorithms

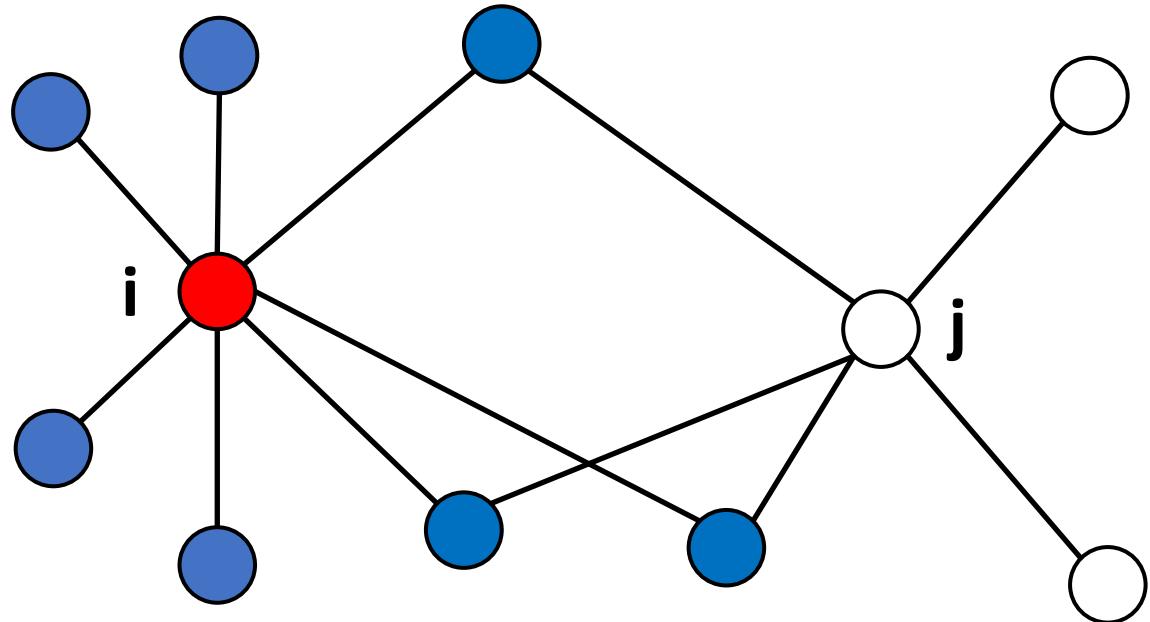
Common Neighbors (CN): $|\Gamma(i) \cap \Gamma(j)|$



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

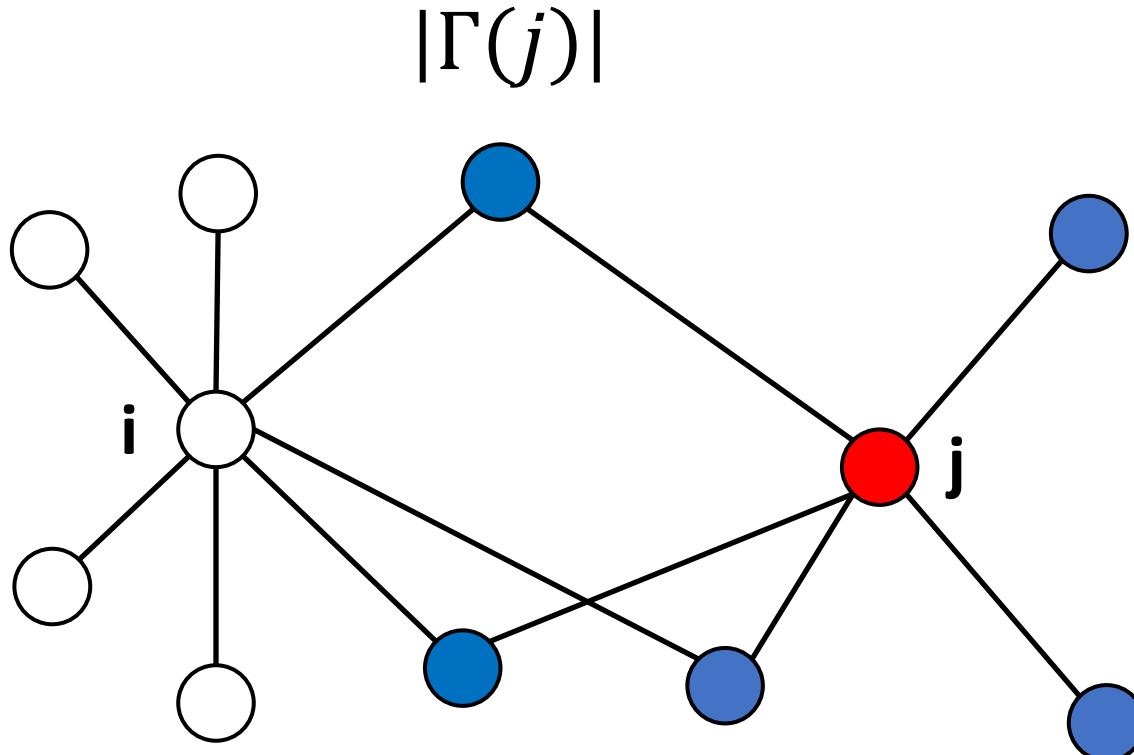
Heuristic algorithms

$$|\Gamma(i)|$$



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

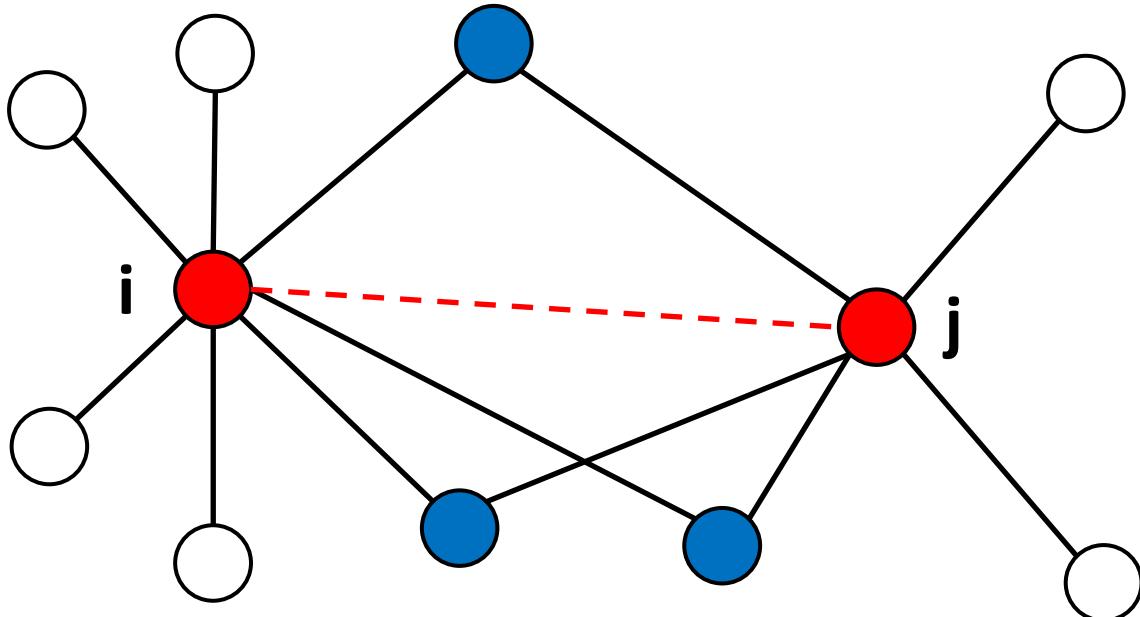
Heuristic algorithms



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

Heuristic algorithms

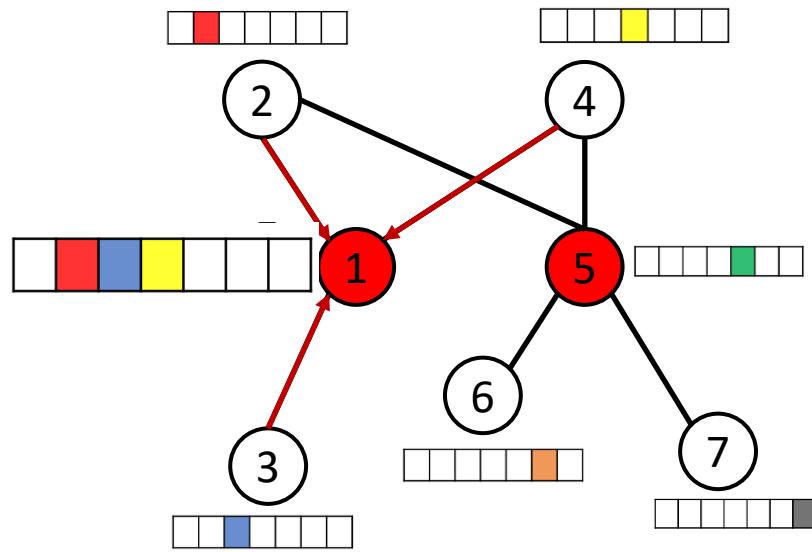
Common Neighbors (CN): $|\Gamma(i) \cap \Gamma(j)|$



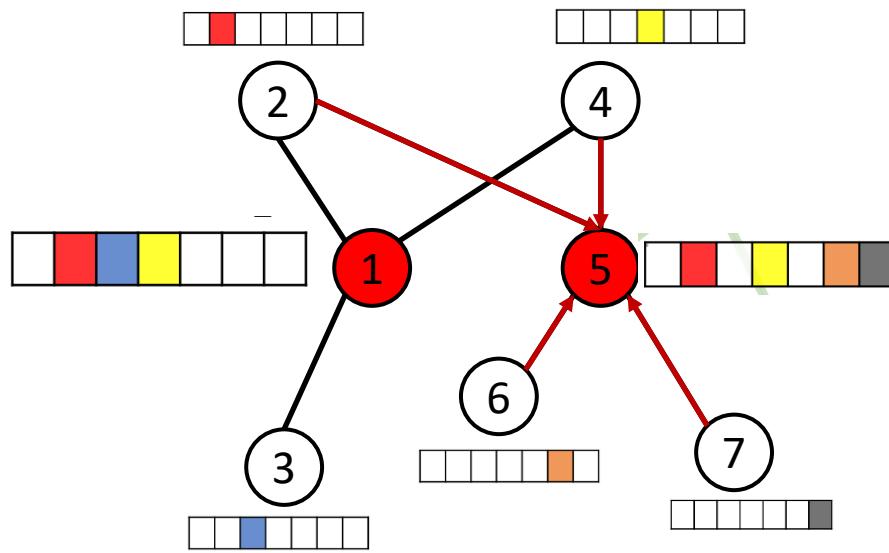
Only **first-order**
neighbors

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

Graph Neural Network -- Neighborhood Overlap GNN

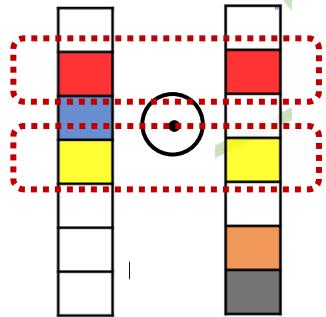


Graph Neural Network -- Neighborhood Overlap GNN



Graph Neural Network -- Neighborhood Overlap GNN

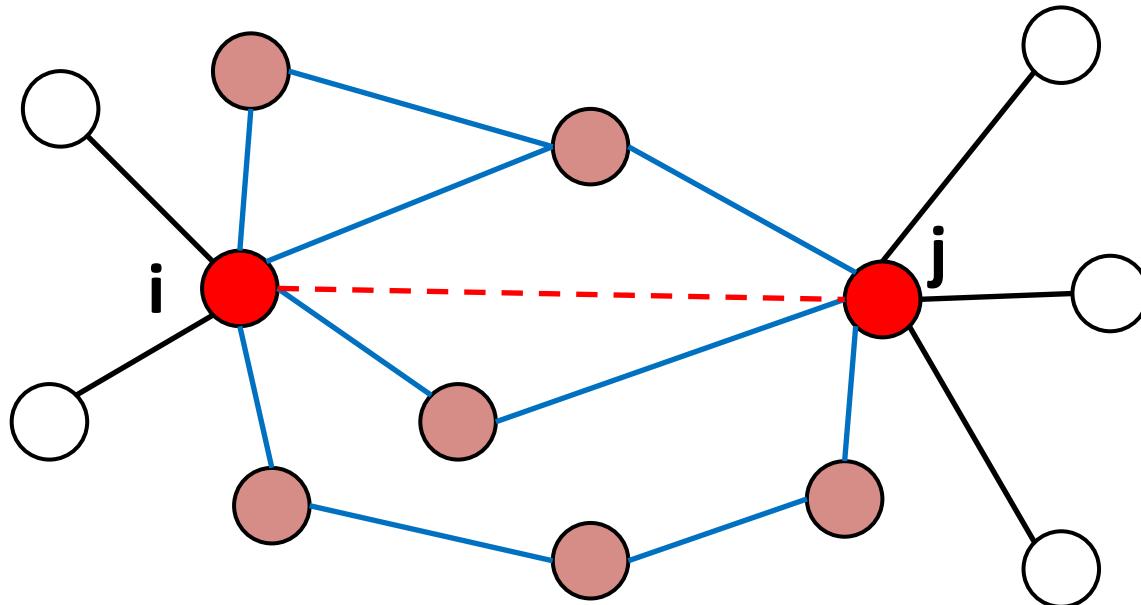
Two common
neighbors



Identify the important Neighborhood similarity information

Heuristic algorithms

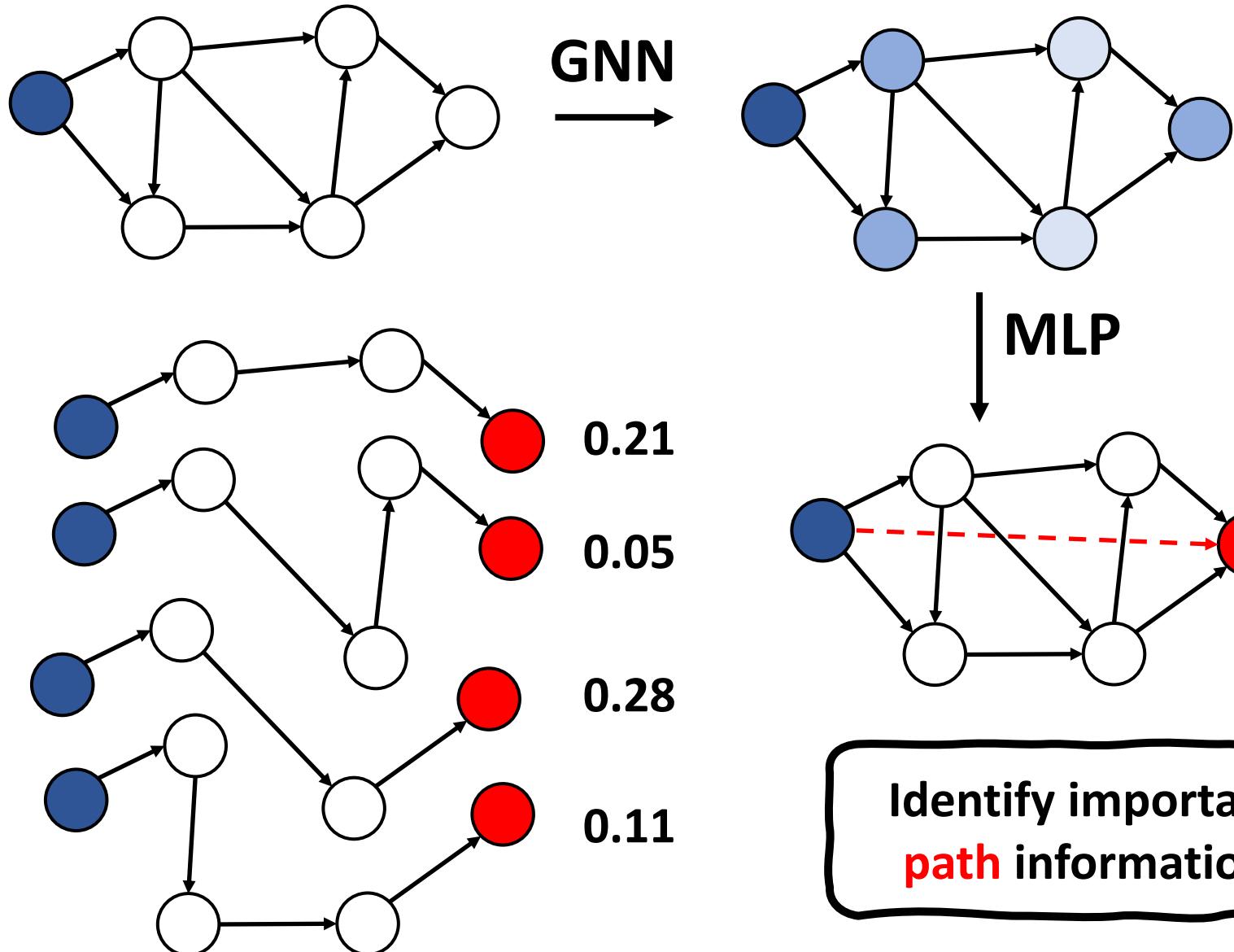
Katz Index: $\sum_{l=1}^{\infty} \beta^l |walks(x, y) = l|$



Higher-order
neighbors on
the entire graph

Notations: $\beta < 1$ is damping factor

Graph Neural Network -- NBFNet



Explore underlying mechanism

Model
mechanism

Data
mechanism

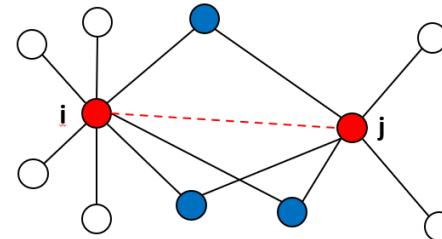
Estimate the proximity
between node pairs

How a link is
formed in a graph

Important data factors

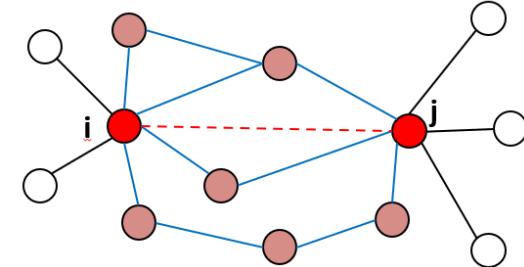
Local Structural Proximity (LSP)

The overlapping between immediate neighbors
CN, RA, AA,



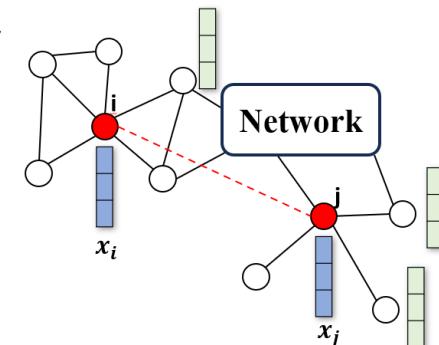
Global Structural Proximity (GSP)

The number of paths with different length
Katz, SimRank, PPR,



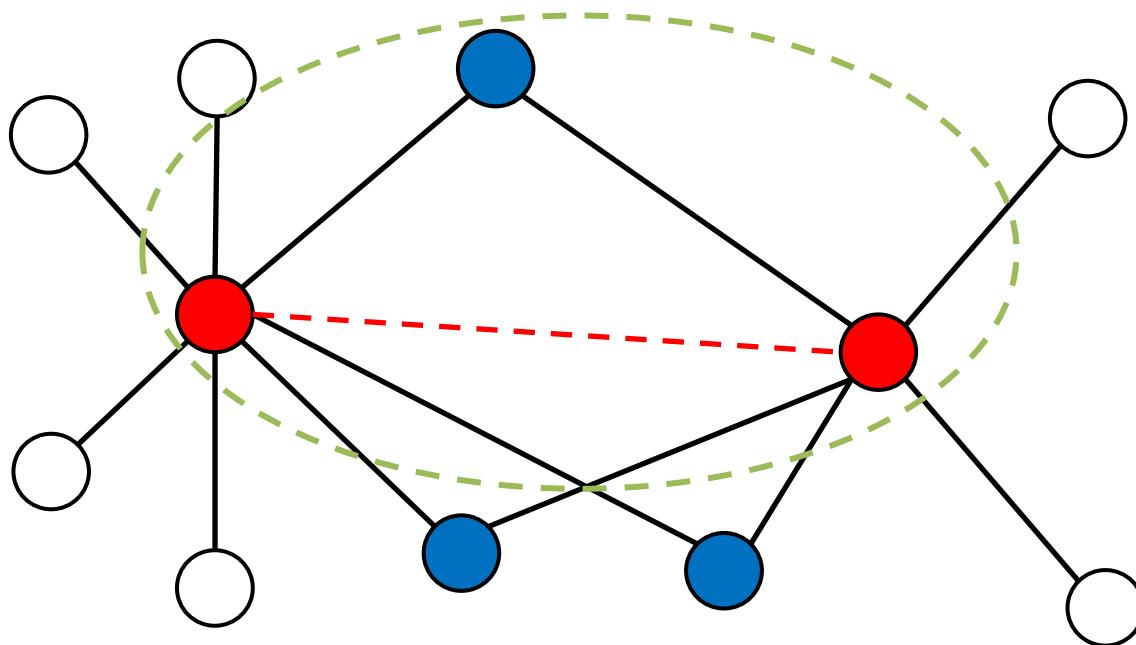
Feature Proximity (FP)

The node feature similarity



Data Factor -- LSP

Common Neighbors (CN), RA, AA,

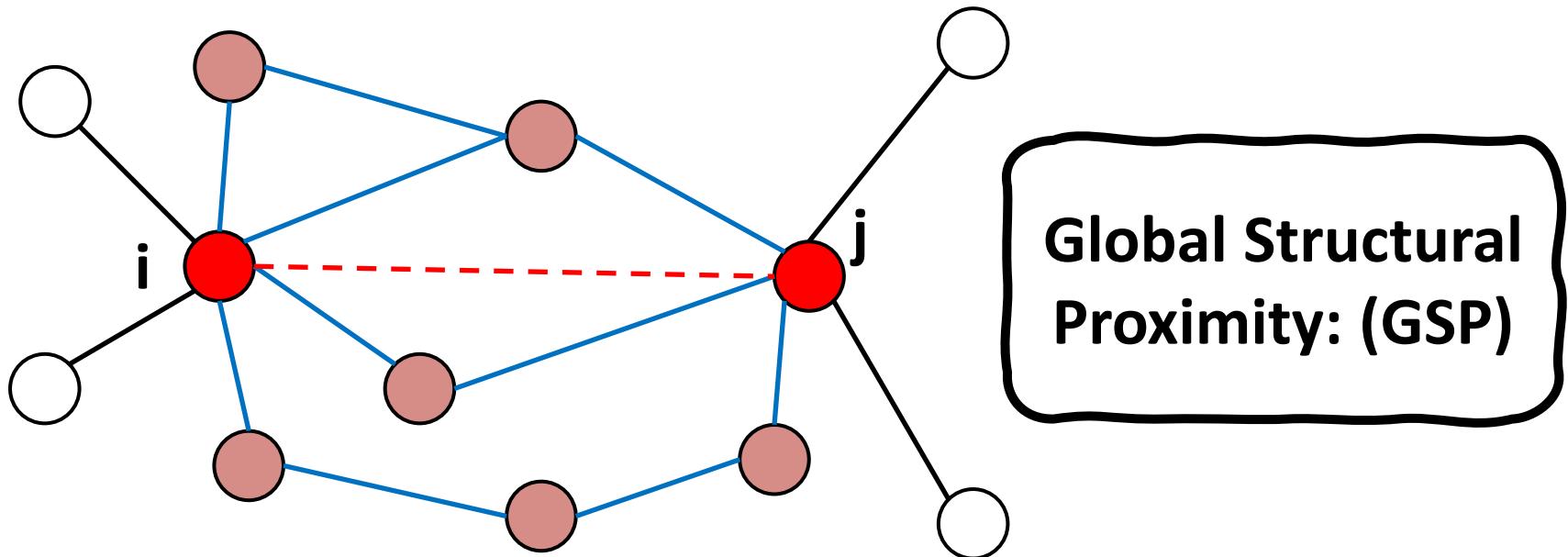


**Local Structural
Proximity: (LSP)**

Rooted in the triadic closure principle

Data Factor -- GSP

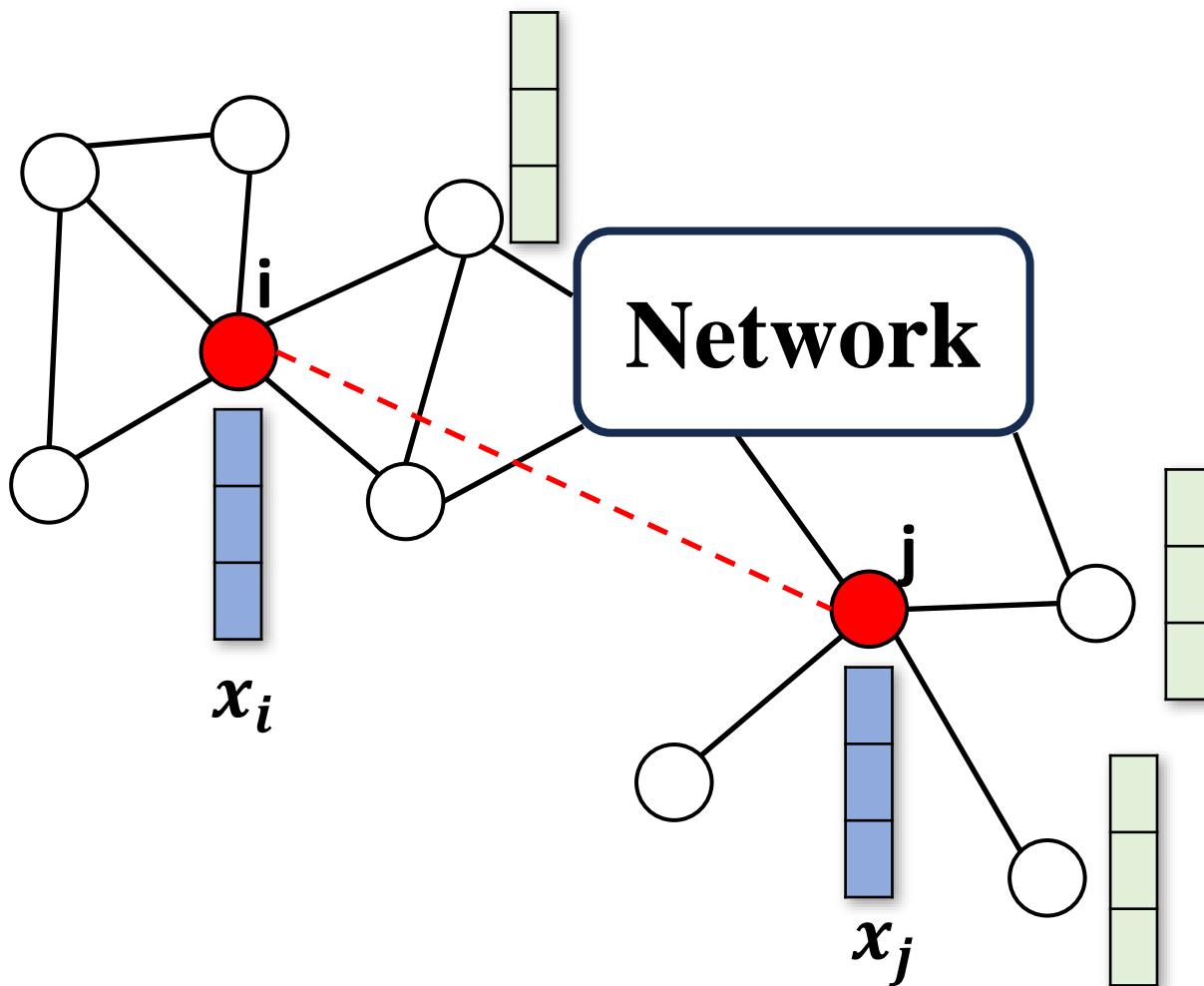
Katz Index, SimRank, PPR,



Rooted in the attenuation principle

Data Factor -- FP

Feature homophily (FH)



Feature
Proximity (FP)

Rooted in the
homophily
principle

Research Question

Does each data factor indeed play a key role?

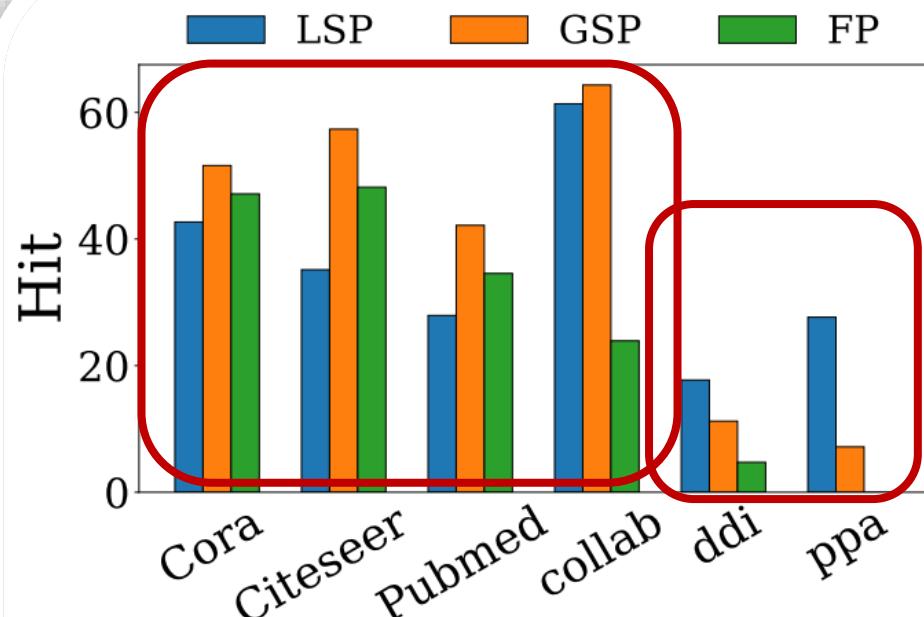
Does each data factor provide unique information?

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

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

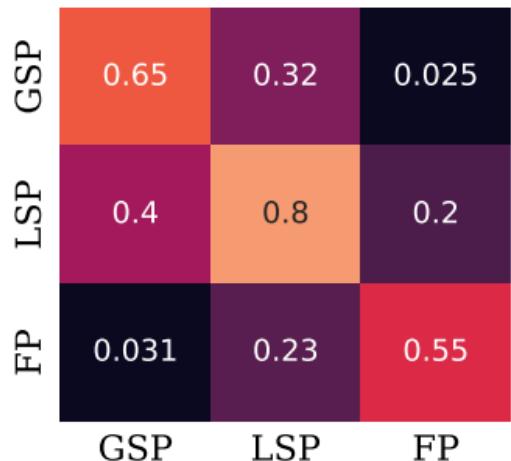
- No consistent winning solution
- The GSP works better on academic networks
- The LSP works better on chemistry and biology networks

Research Question

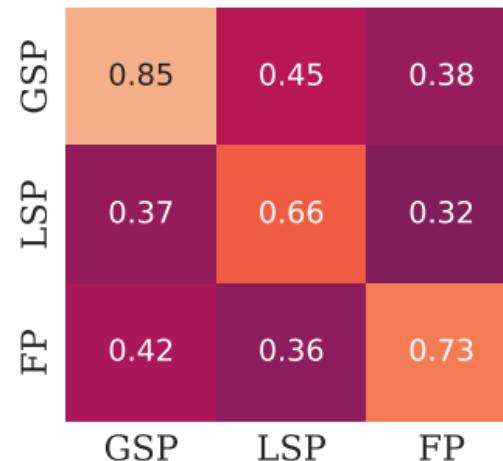
Does each data factor indeed play a key role?

Does each data factor provide unique information?

Complementary effects

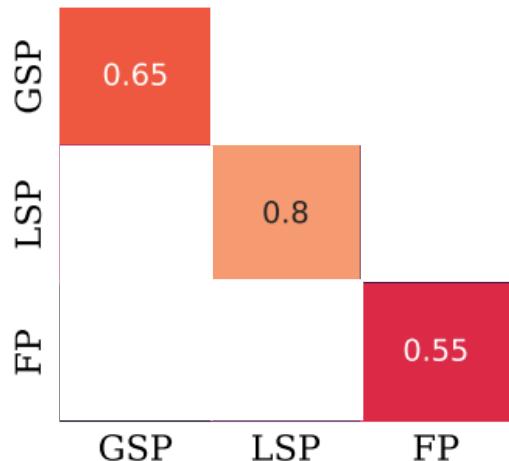


(a) PUBMED



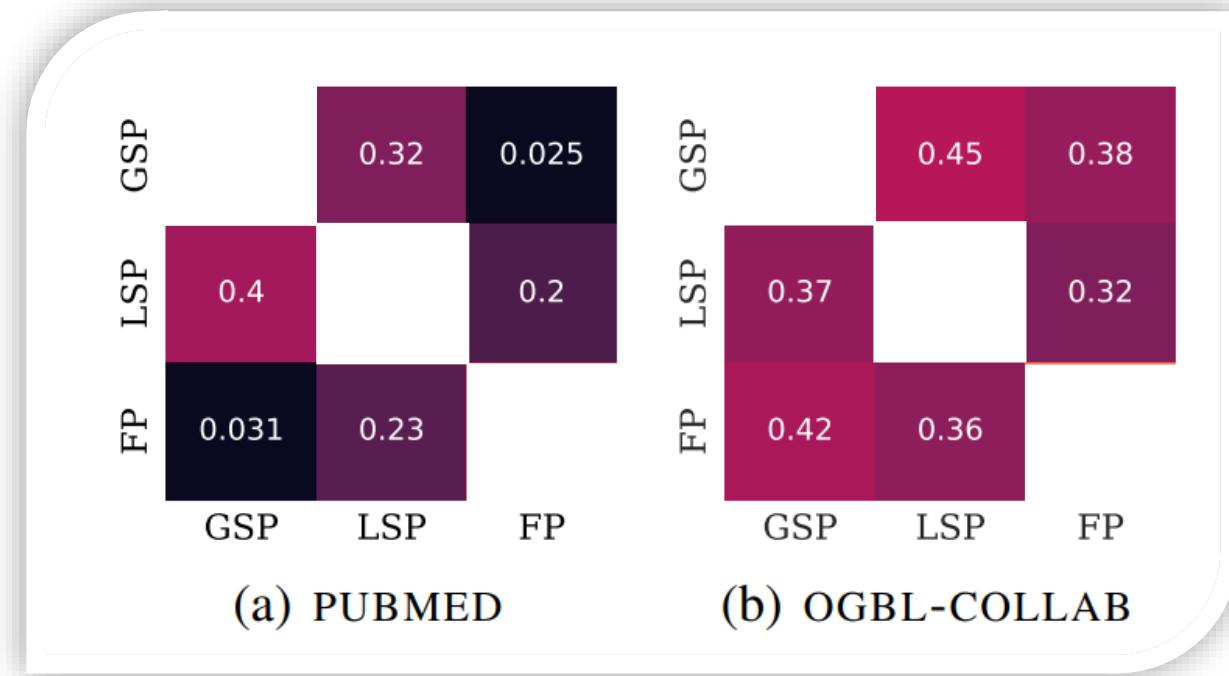
(b) OGBL-COLLAB

Complementary effects



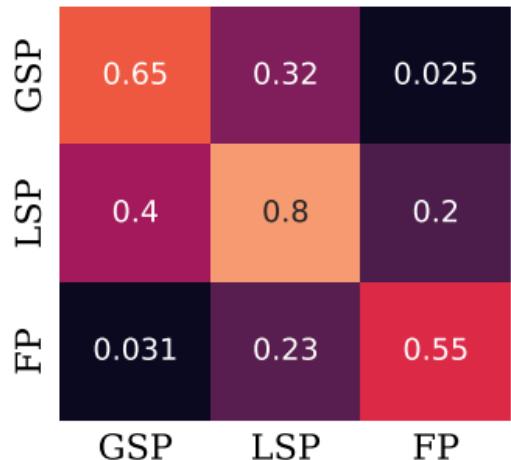
The **prediction overlapping** between heuristics from the same factor is large

Complementary effects

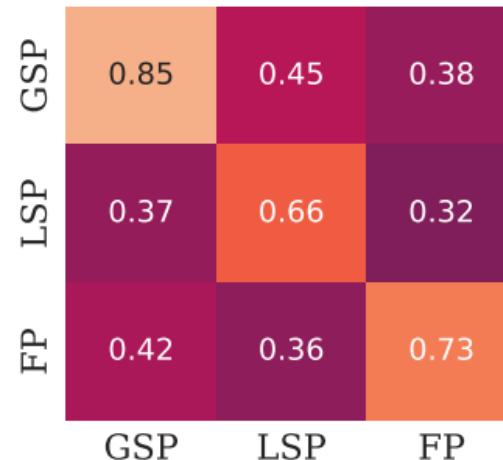


The prediction overlapping between heuristics from different data factors

Complementary effects



(a) PUBMED



(b) OGBL-COLLAB

The prediction overlapping between heuristics from different factors is small

A recap

Local Structural
Proximity (LSP)

The overlapping between immediate neighbors

Global Structural
Proximity (GSP)

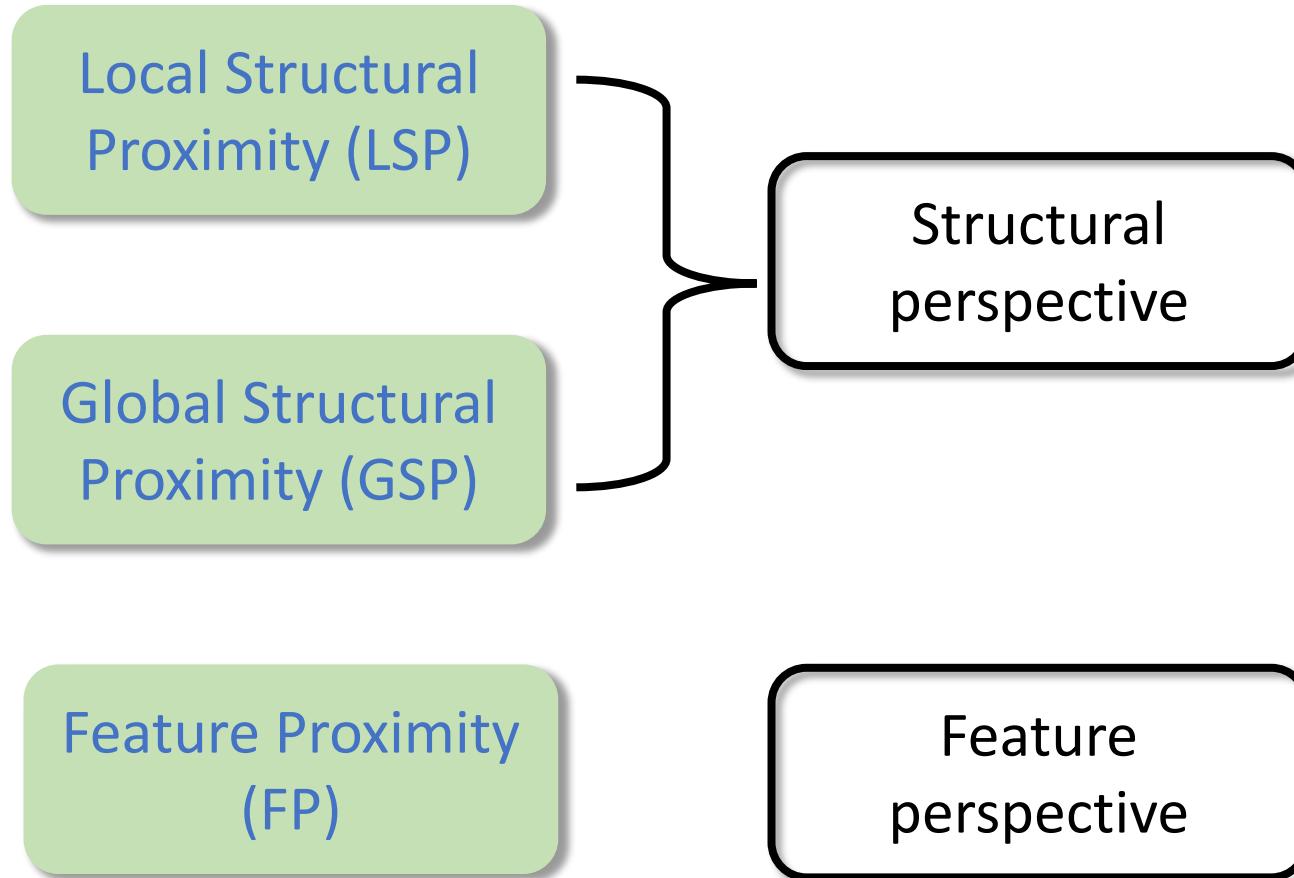
The number of paths with different length

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

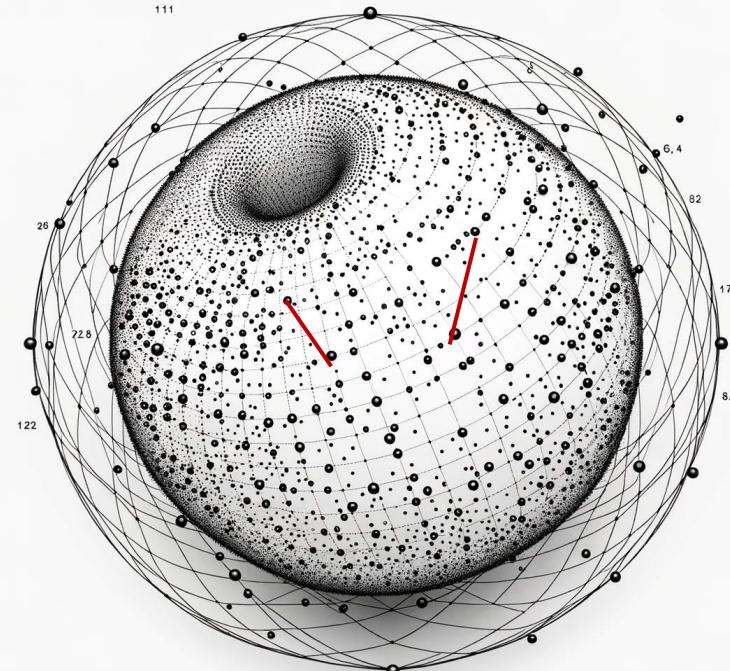
Instead of focusing on one node, similarity between **a pair of node** should be the focus.

Desired properties for the theoretical model

Latent space Model

Structural perspective

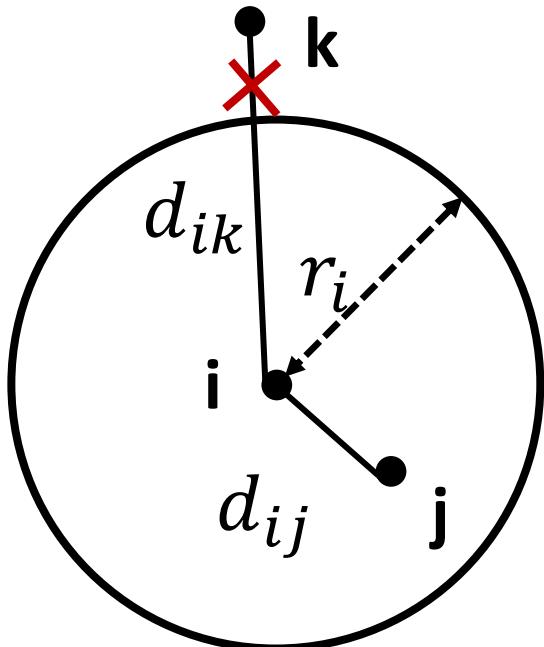
Feature perspective



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

Structural perspective

Structural
perspective



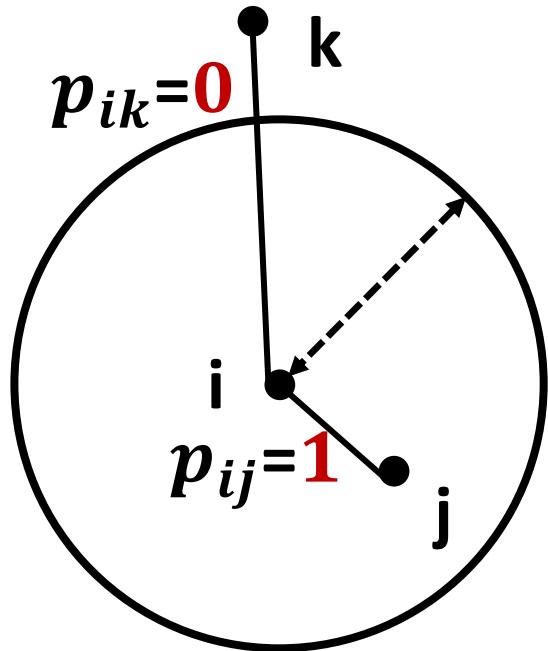
□ r_i is the accept threshold for node i
 d is the structural similarity distance

- 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

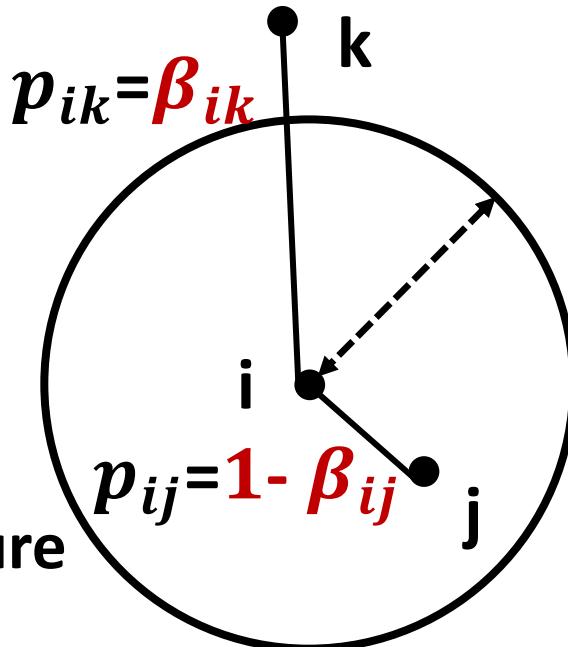
d_{ij} can be viewed as the social closeness

Feature perspective

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



Consider feature
similarity β_{ij}



Structural
Only

Structural
and Feature

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

Structural
distance

The number of CNs

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

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

The effectiveness of the GSP factor

GSP

**Structural
distance**

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

The number of paths with length ℓ

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

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

The feature similarity β_{ij}

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

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

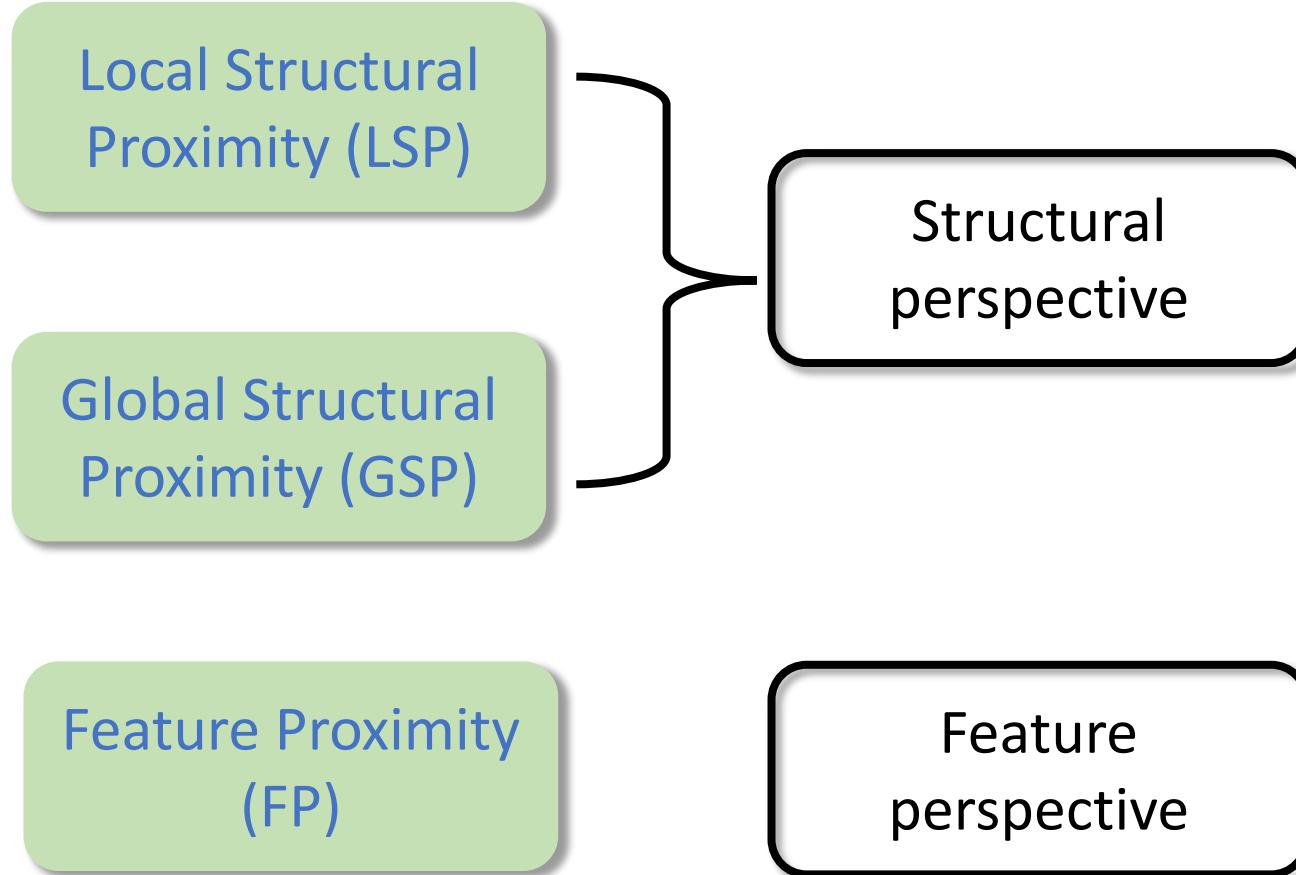
GSP

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

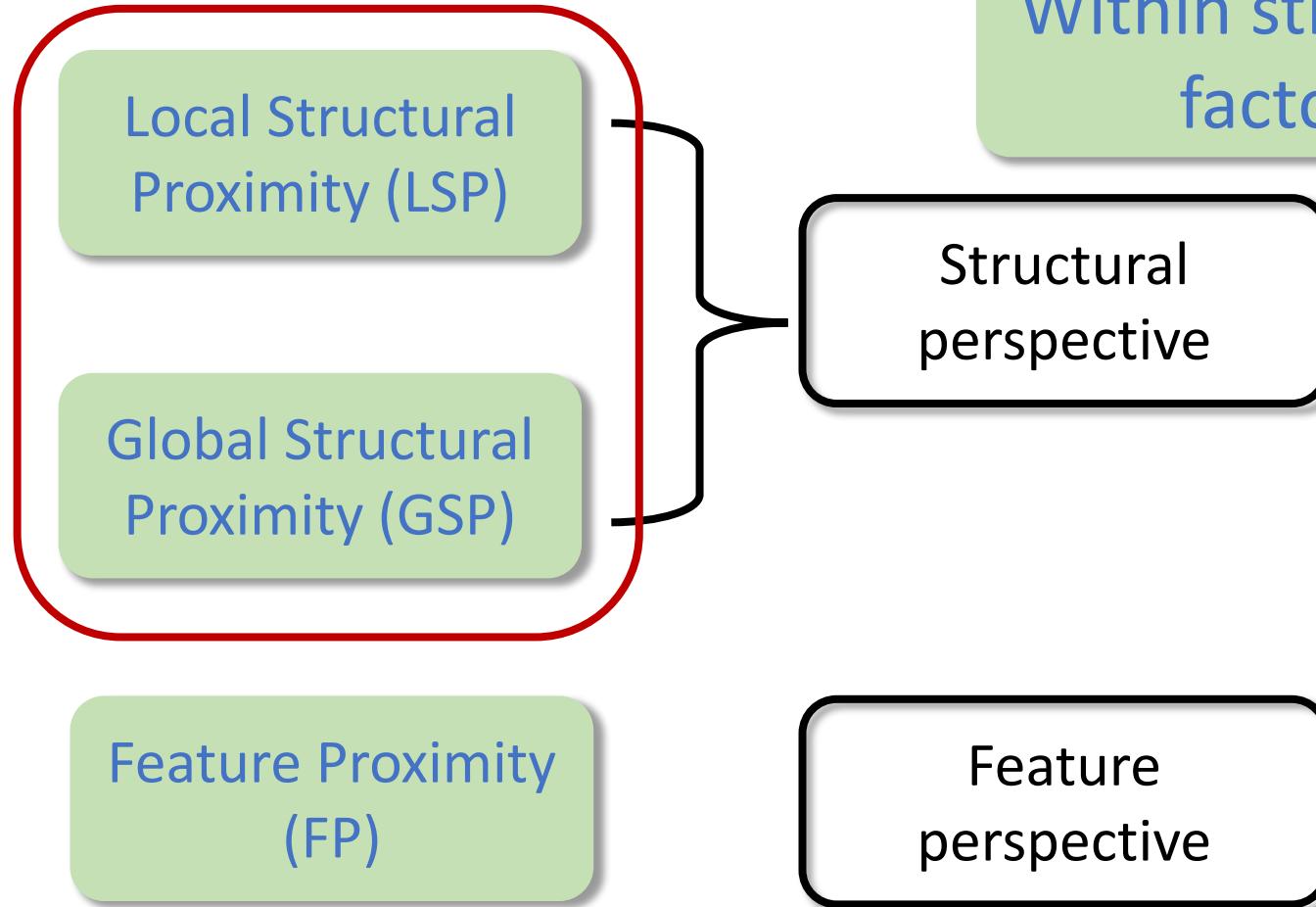
FP

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

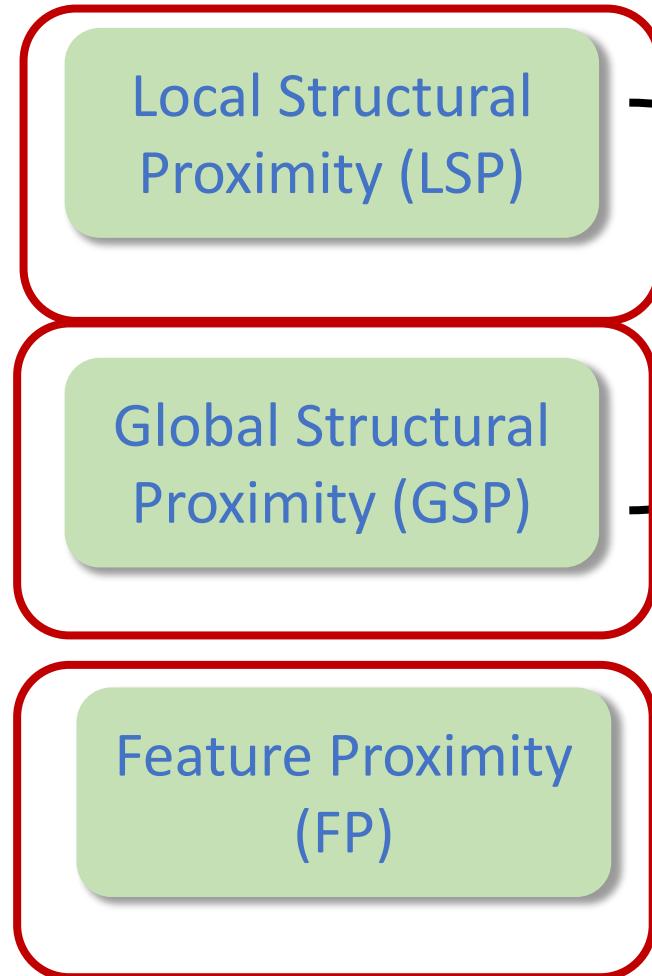
The relationship between different factors



The relationship between different factors



The relationship between different factors



Between structure and feature factors

Structural perspective

Feature perspective

The relationship between different factors

Within structural
factors

Between structure
and feature factors

GSP vs LSP

LSP vs FP

GSP vs FP

LSP vs GSP

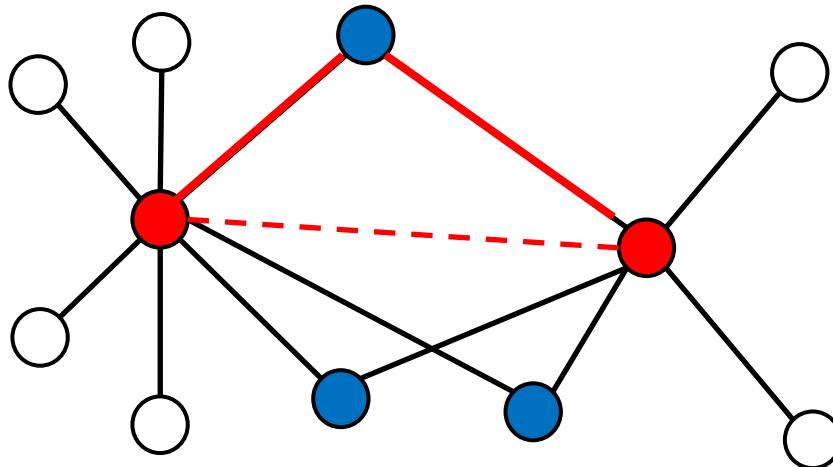
Within structural
factors

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

Between structure
and feature factors

GSP vs LSP

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:

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

$$c' < 0$$

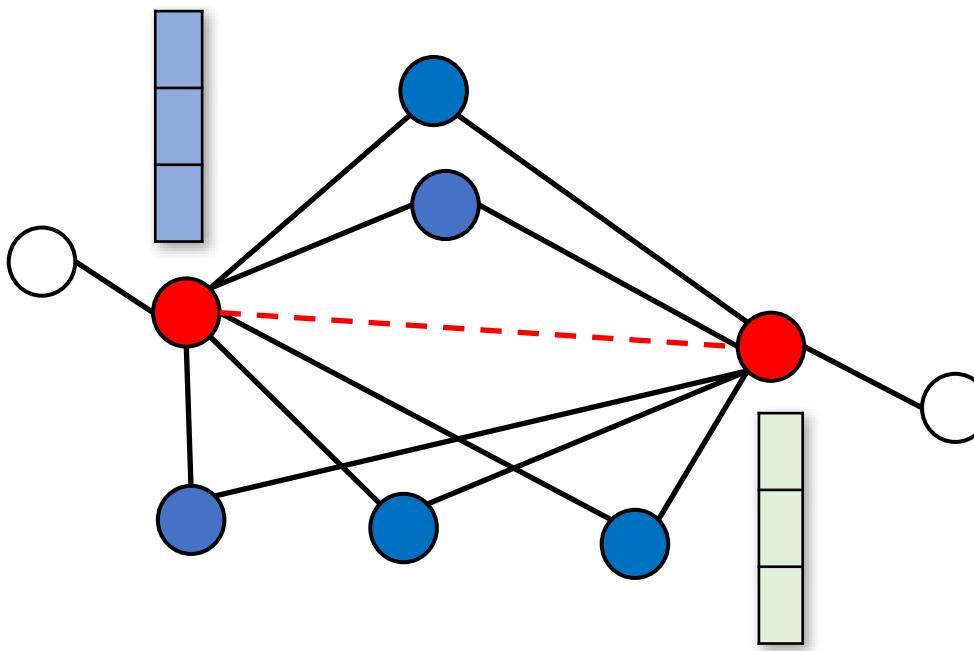
The number
of CNs

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}

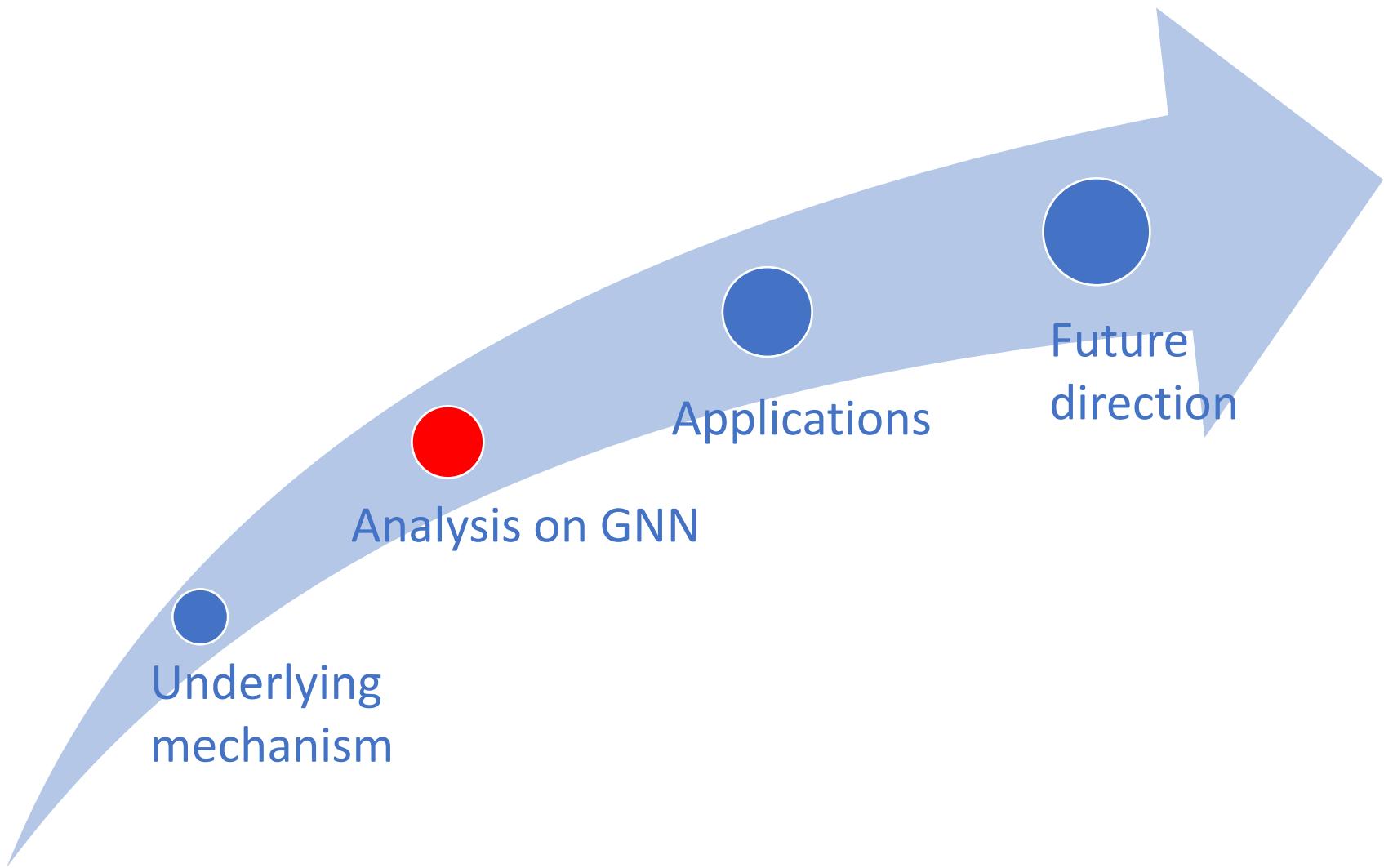
The incompatibility between LSP and FP

Example of incompatibility between LSP and FP



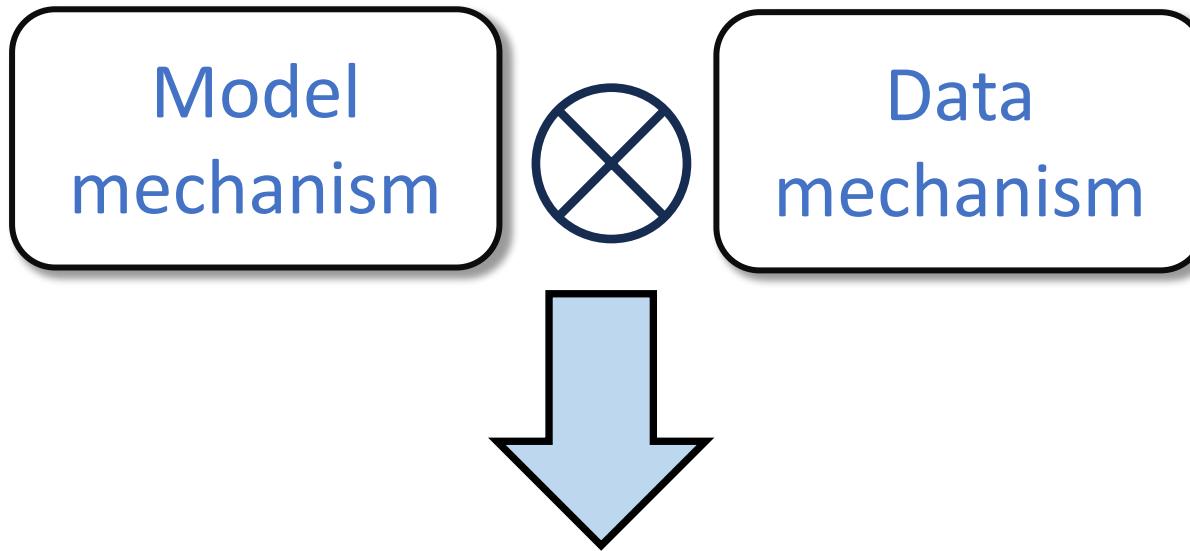
Two persons with a lot of common neighbors but still not connected is because of the low feature similarity

Outline



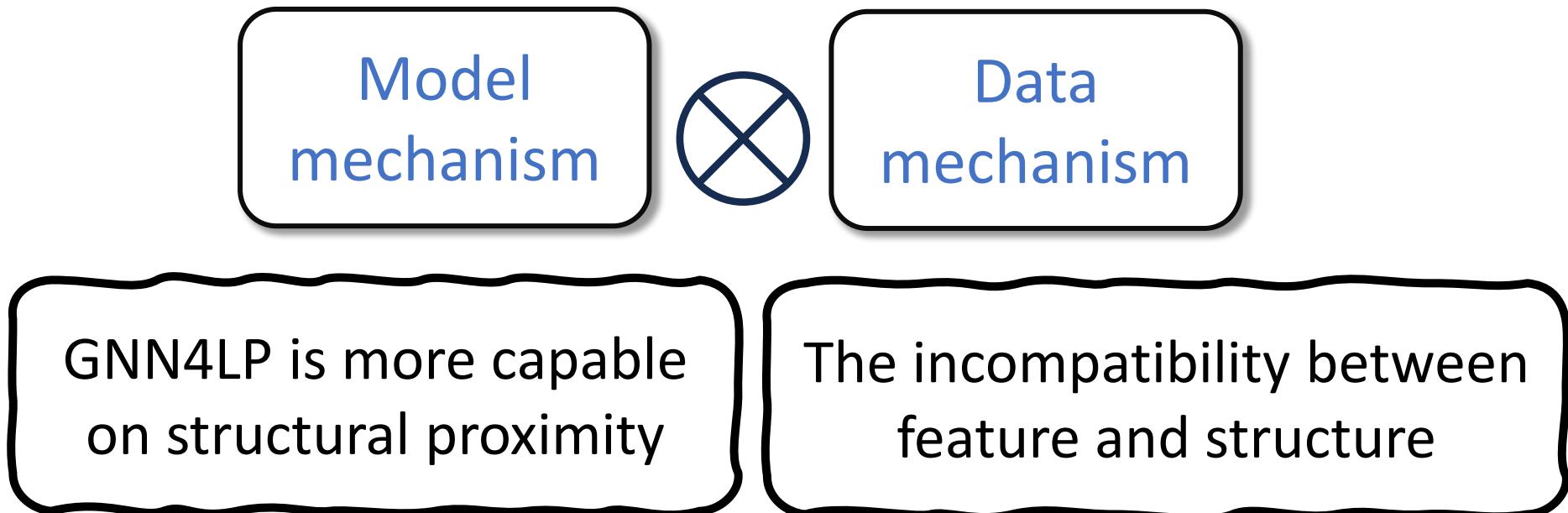
Analysis on GNNs

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

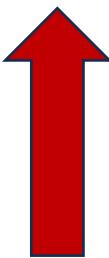


How models behave with such incompatibility?

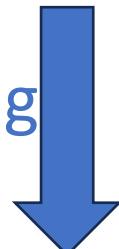
How models behave with such incompatibility?



There is no free lunch with incompatibility!

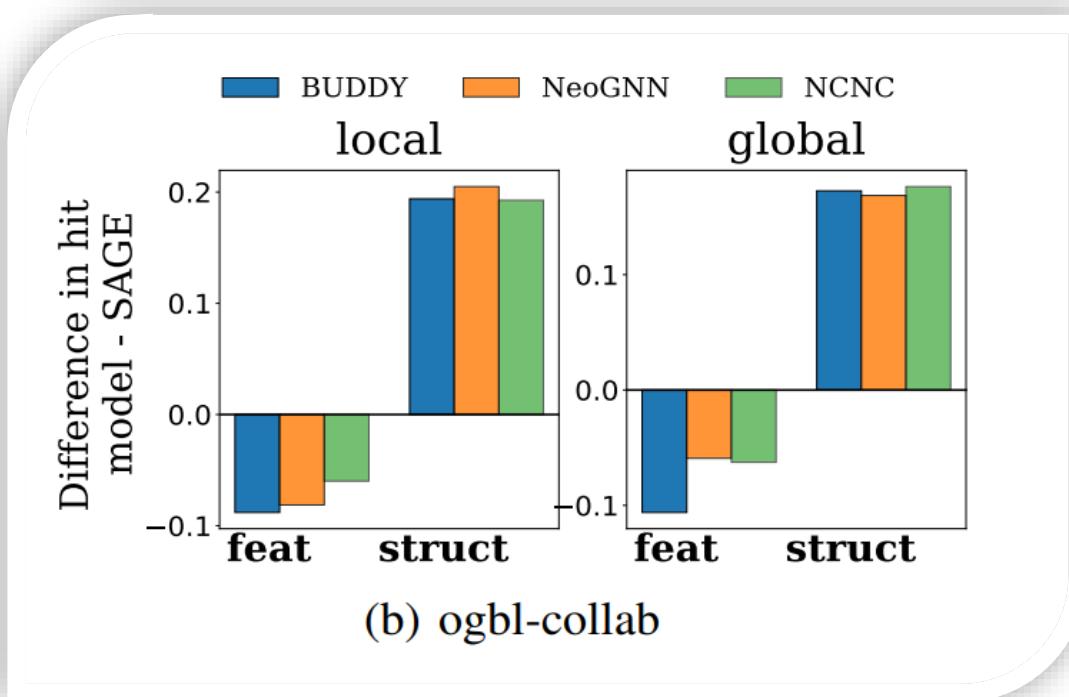


Better structure modeling Worse feature modeling



Empirical evidence

Performance comparison between GNN4LP models and GraphSAGE.

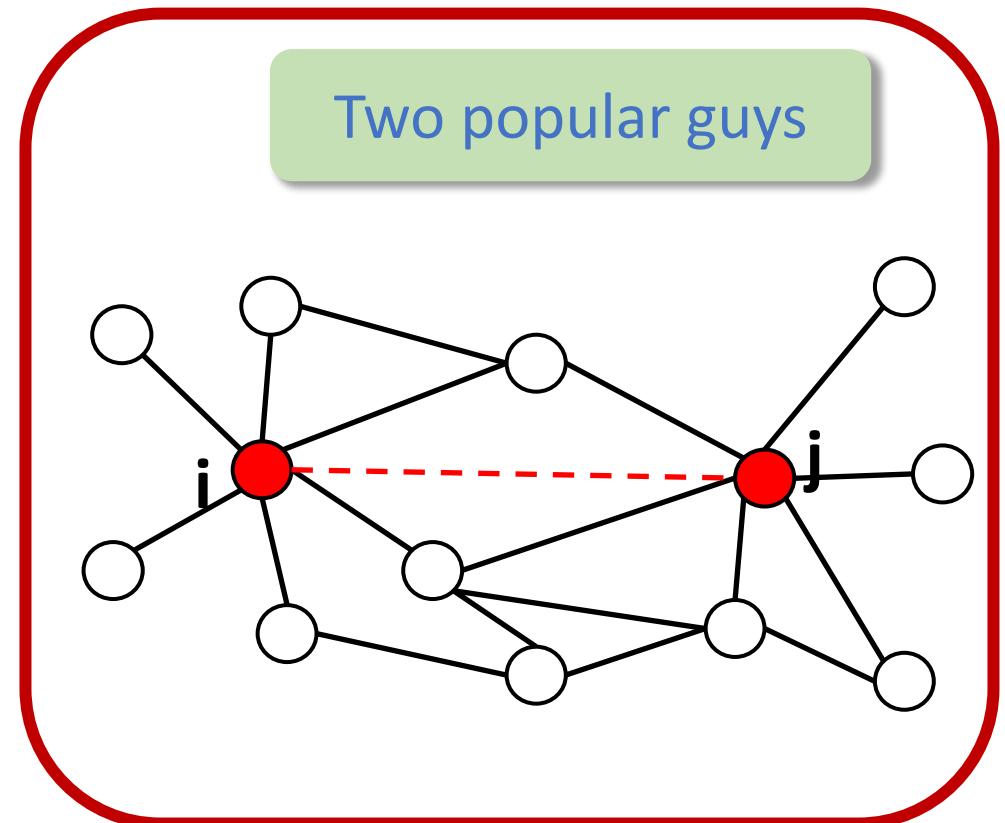
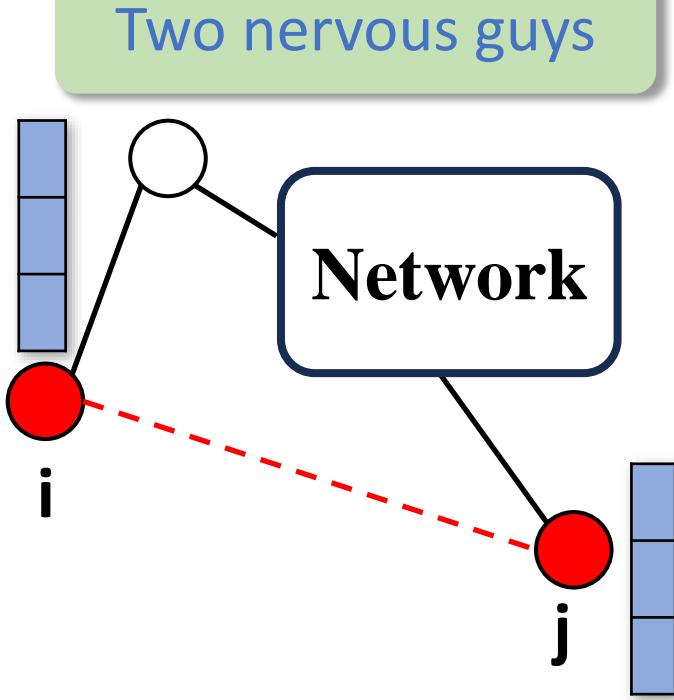


GNN4LP models outperform on node pairs with structure proximity, but fail on the feature ones

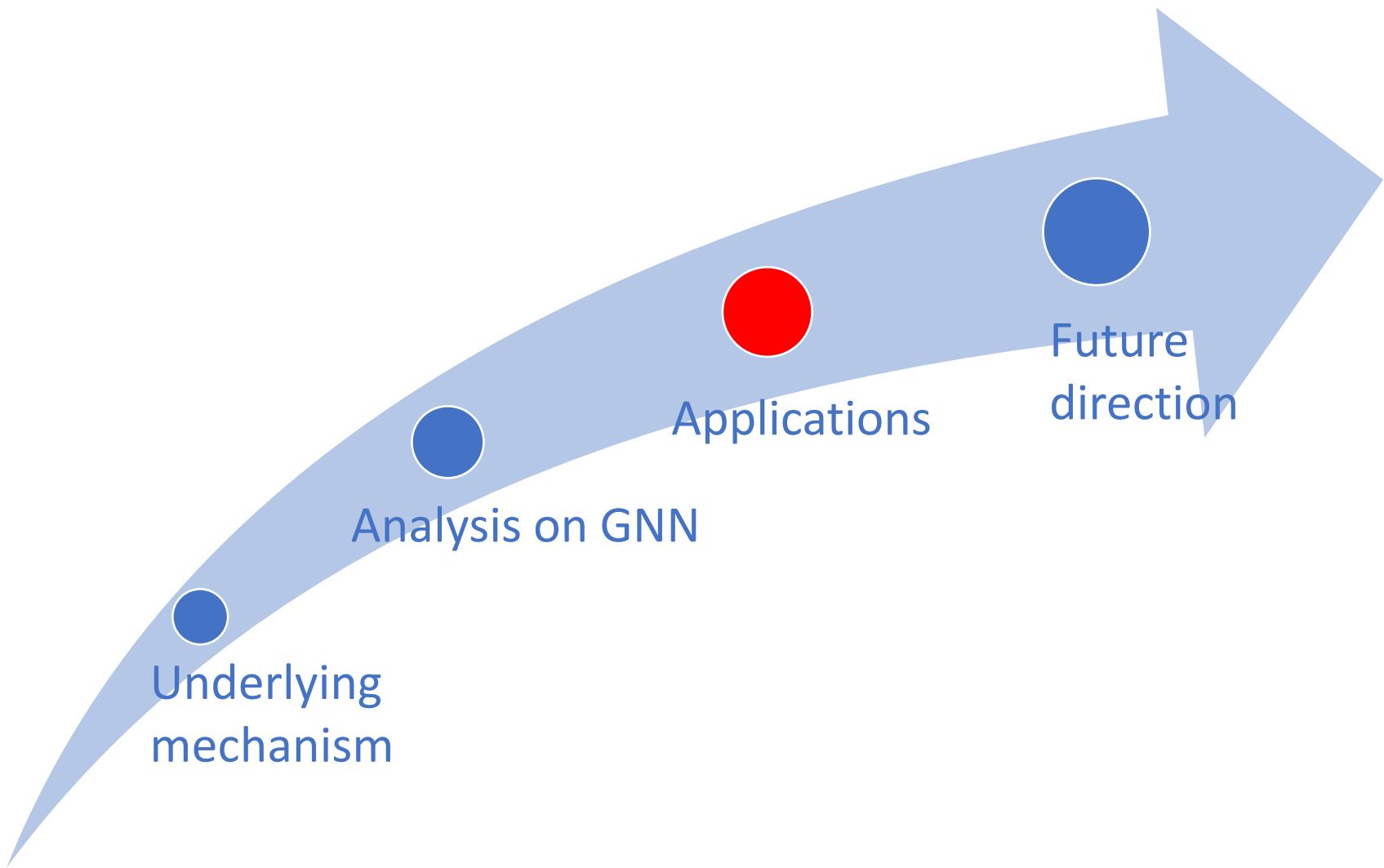
Broader impact

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

Leads to the essential fairness issue



Outline



Guidance for Practitioners

Model
mechanism

Data
mechanism

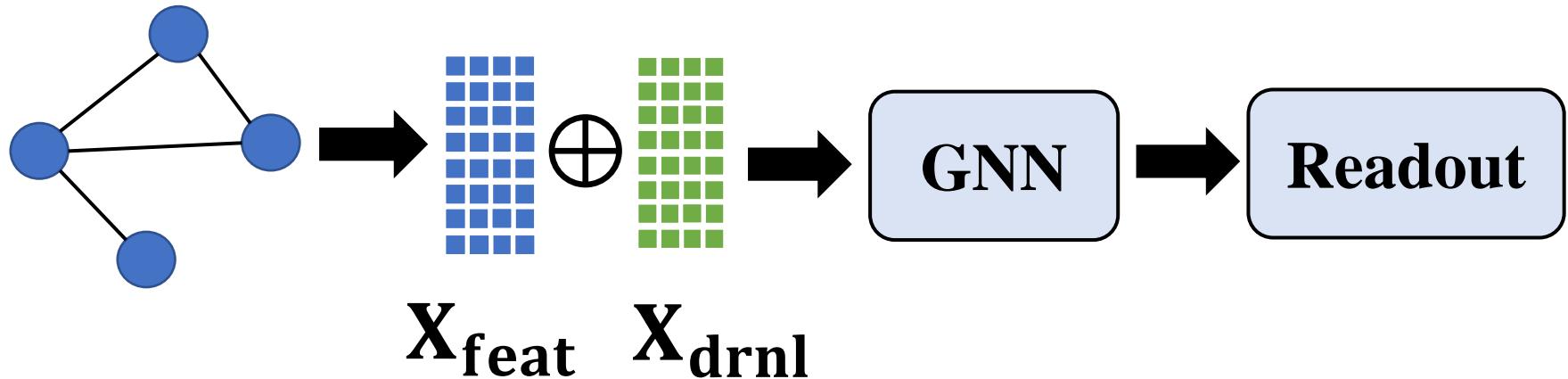
Guidance for Practitioners

Model
mechanism

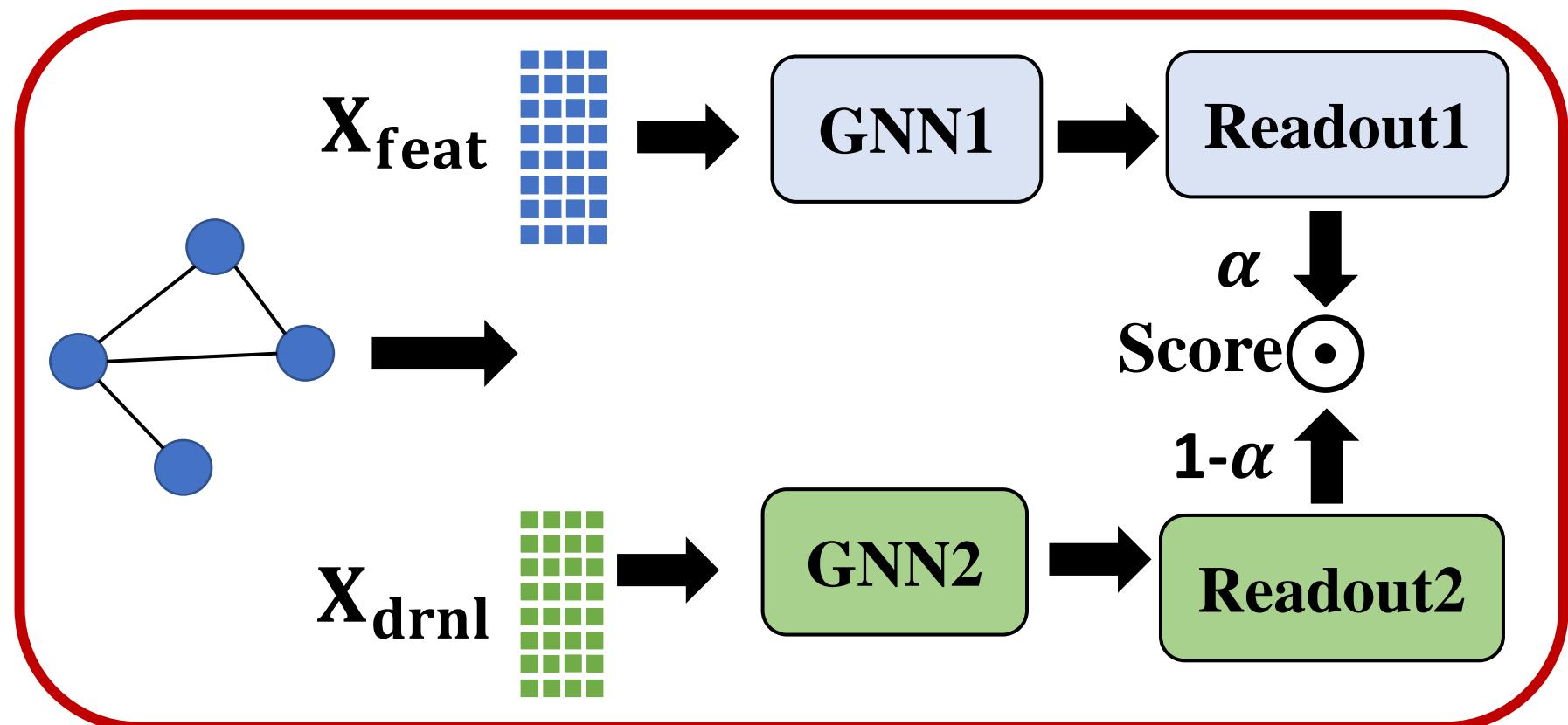
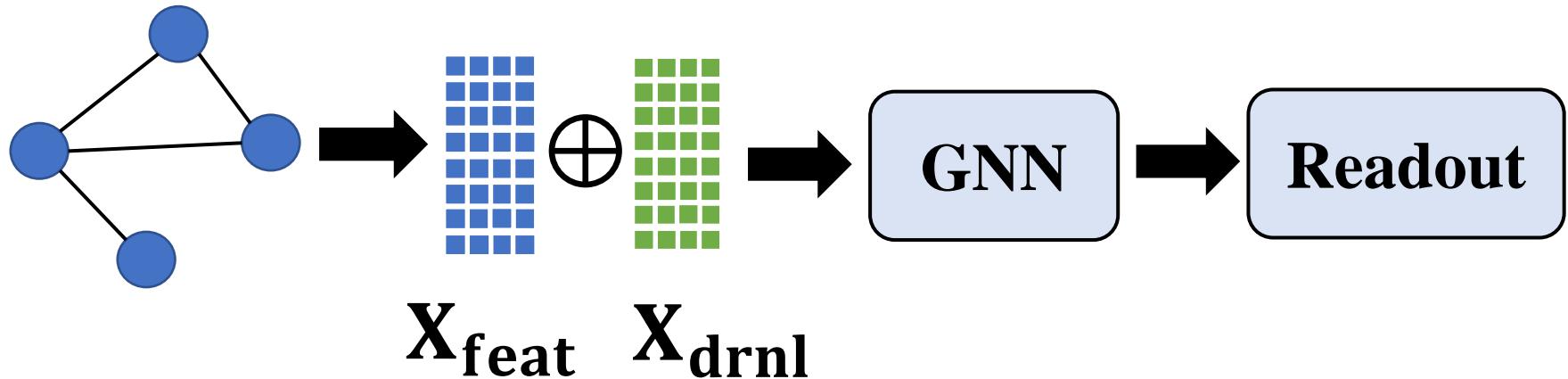
Data
mechanism

A new model design principle to
mitigate incompatibility

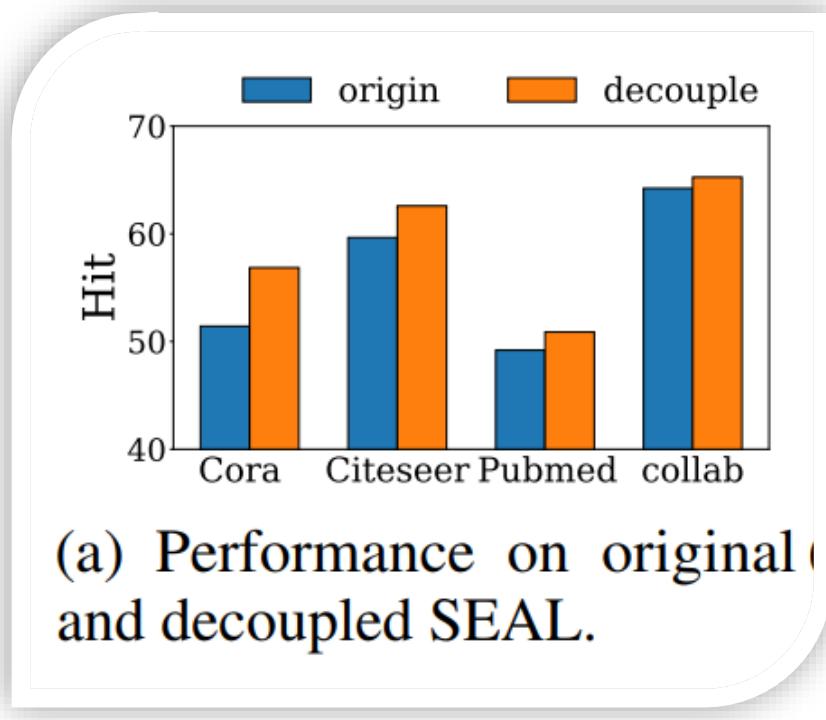
Do not model feature and structure together!



Combine feature and
structure into one GNN



Empirical evidence



Decoupling model leads to
the performance gain

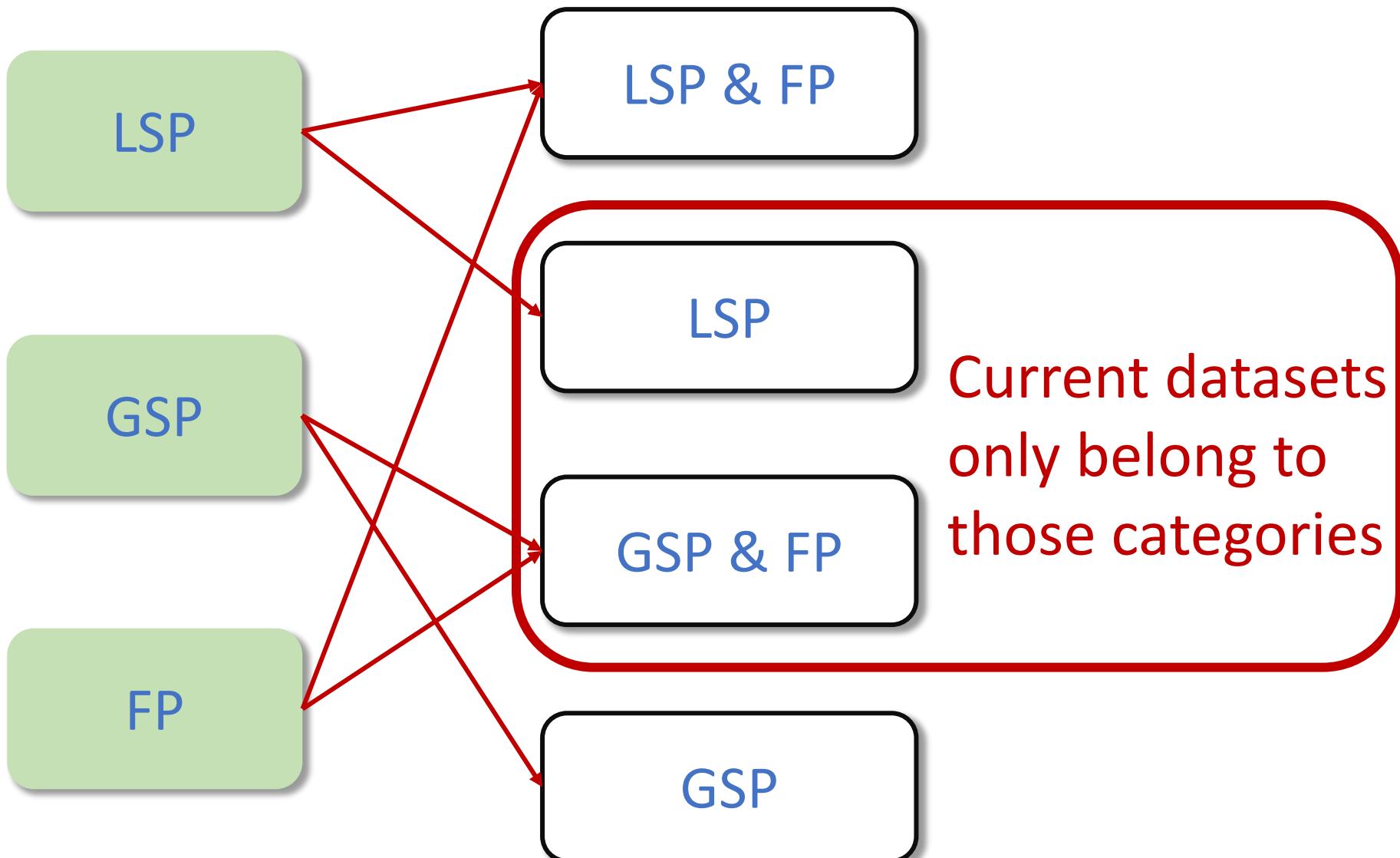
Guidance for Practitioners

Model
mechanism

Data
mechanism

How to achieve a comprehensive
evaluation fully covering all data factors?

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)

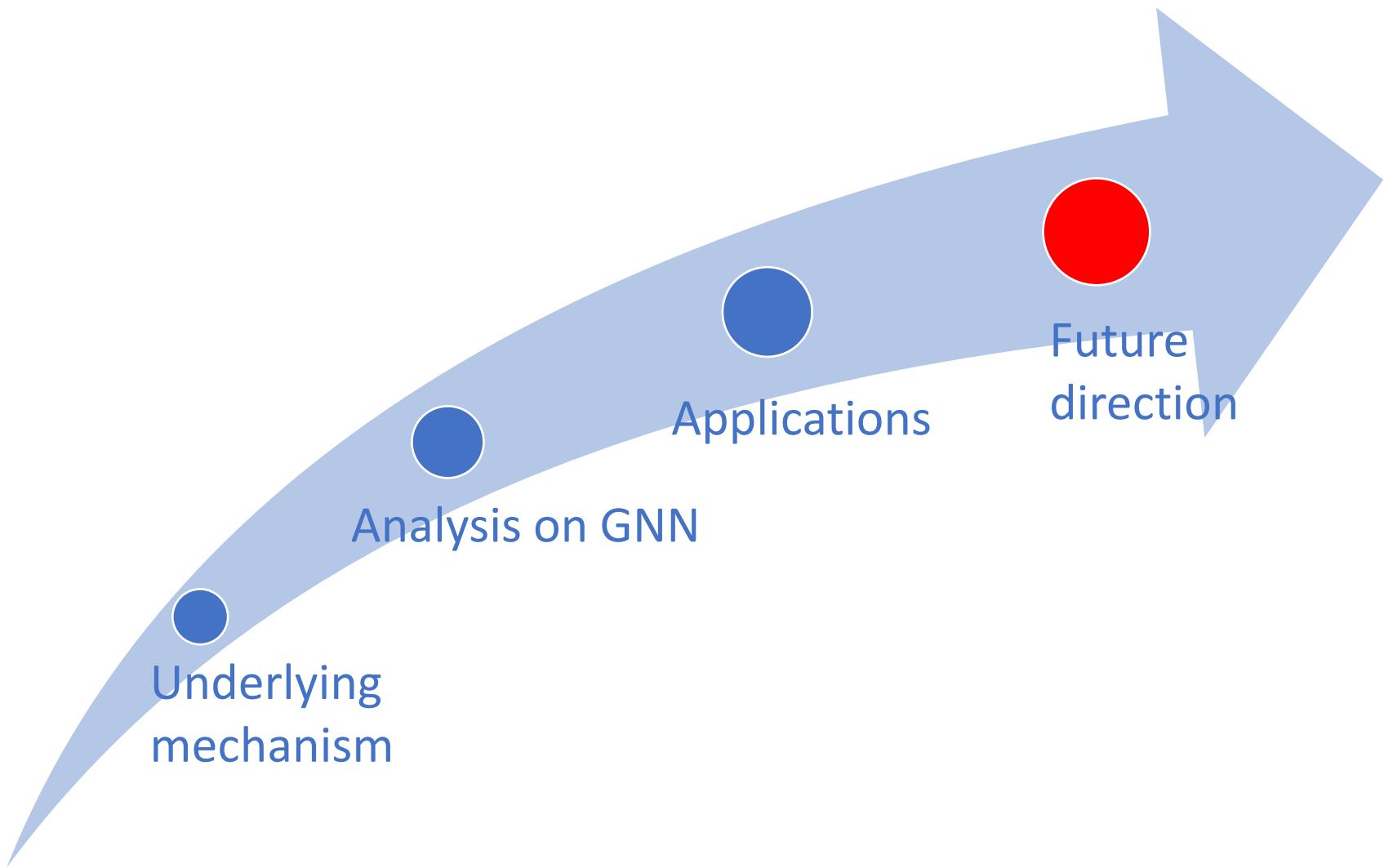
Guidance for a comprehensive evaluation

Select datasets in the same
category for the evaluation

Select datasets in the same
domain for the evaluation

- Biology
- Academia
- Transportation
- Social science
- Web

Outline



Main takeaways

There is no universal best algorithm in Link Prediction

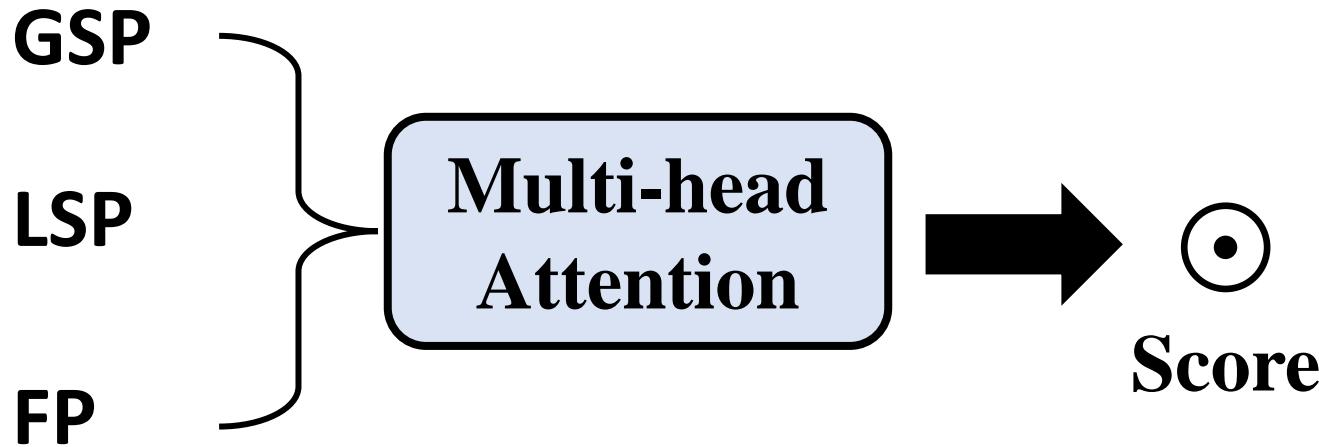
Take a closer look to data property before model design

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

Carefully handle incompatibility between FP and LSP

What are we doing now?

A more
fair model



Utilize transformer to capture different data factors together

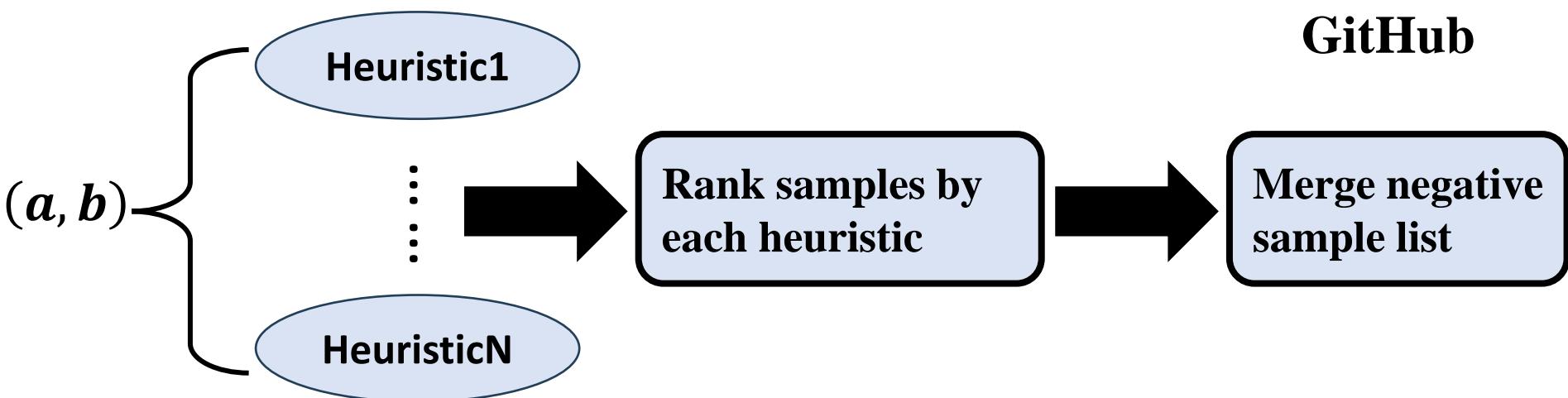
What are we doing now?

Link Prediction
benchmark

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



GitHub



Future work

New model design for graphs with new-found categories

AutoML for a link prediction fit in different problems

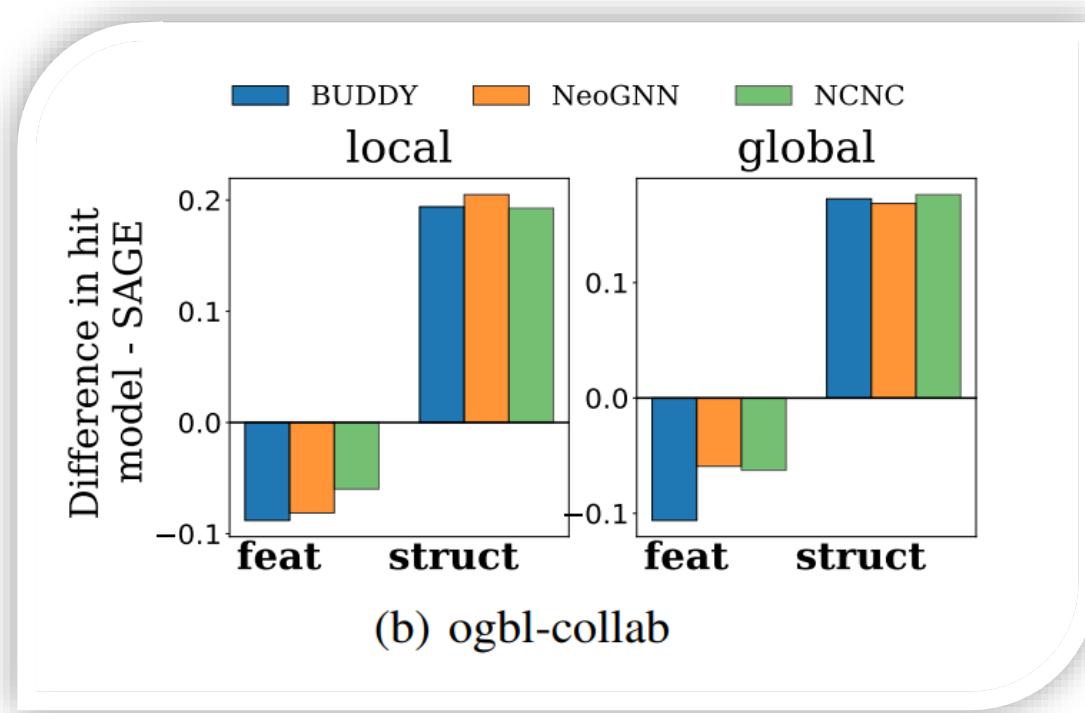
Formally Formulate the fairness problem in Link Prediction

Extend to recommendation with complex user behavior

A pause & QA!

Empirical evidence

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

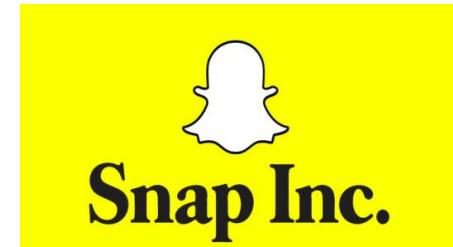


How about the node classification?

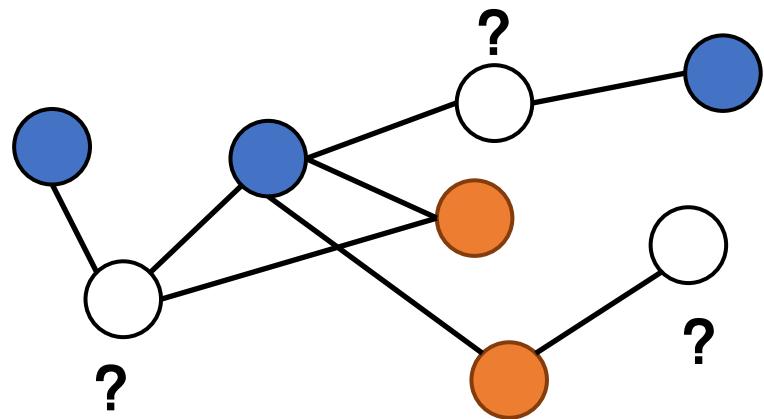
Demystifying Structural Disparity in Graph Neural Networks: Can one size Fit ALL?

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

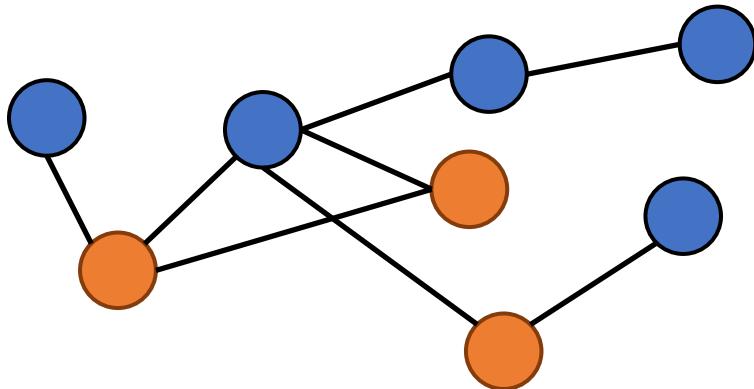
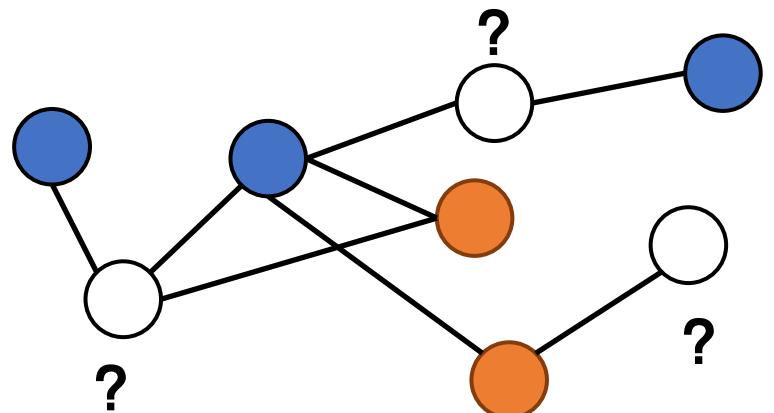
NeurIPS 2023



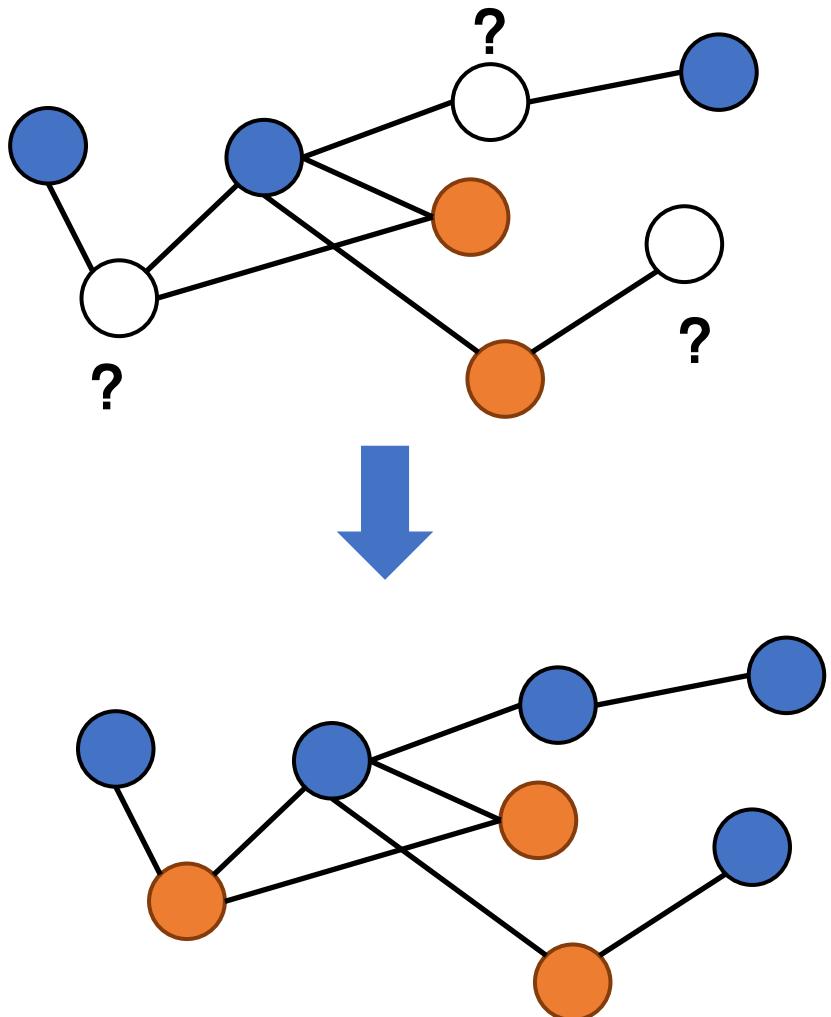
Node Classification task



Node Classification task

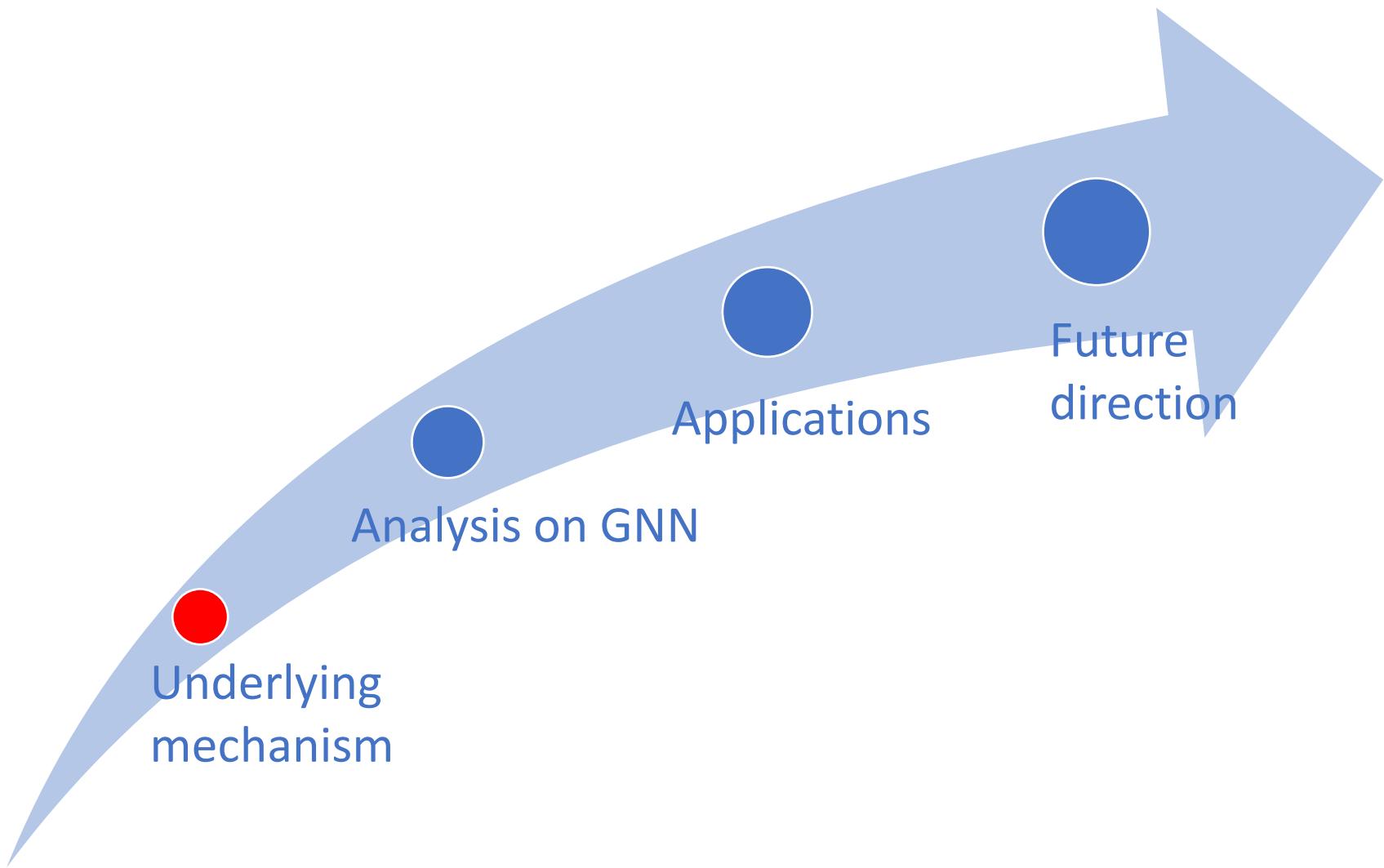


Node Classification task



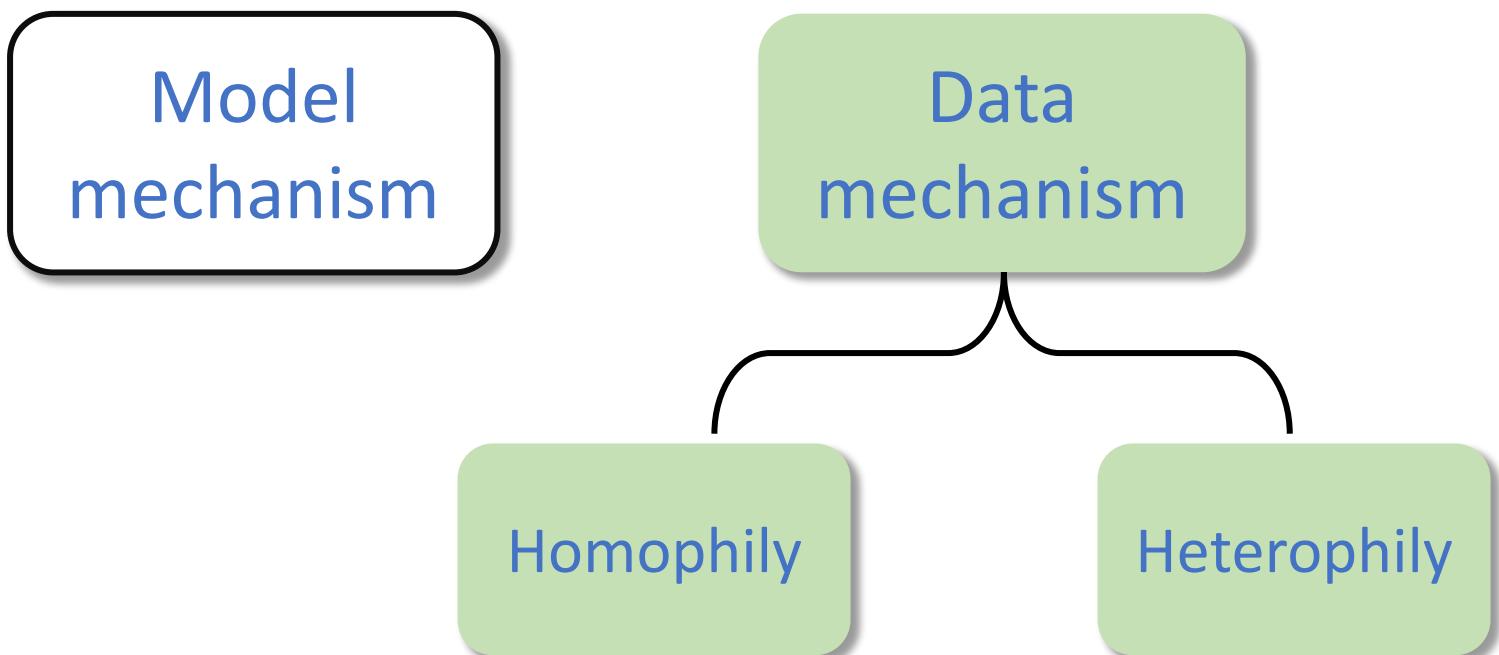
- Inferring Node Attributes
 - Social Influence Prediction
 - Traffic Prediction
 - Air Quality Prediction
- ⋮

Outline



Data mechanism

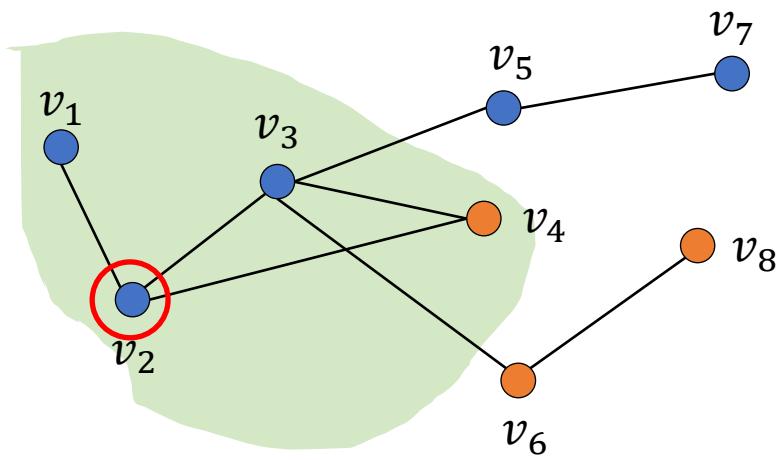
When do GNNs perform well on different nodes and when not?



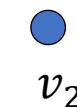
Homophily & Heterophily

Homophily: “nodes tend to connect with “similar” or “alike” others”

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

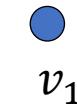


Center node:



v_2

Neighbor set:



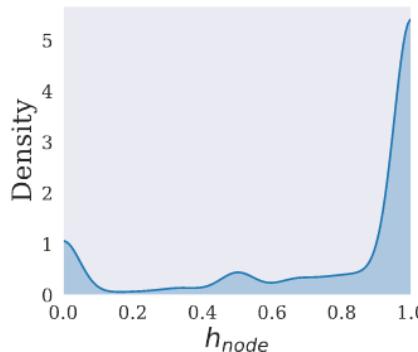
v_1 v_3 v_4

Node homophily ratio of v_2 :

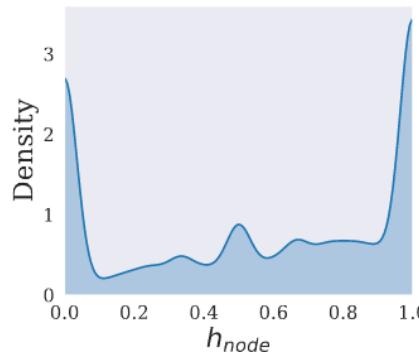
$$h_2 = \frac{2}{3}$$

Data analysis

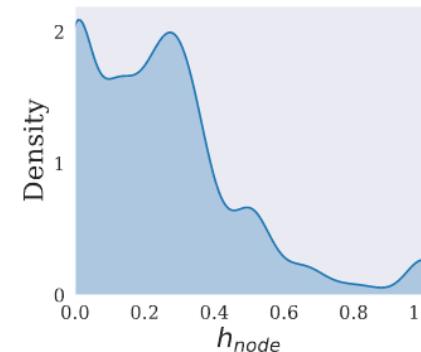
Node homophily ratio distribution



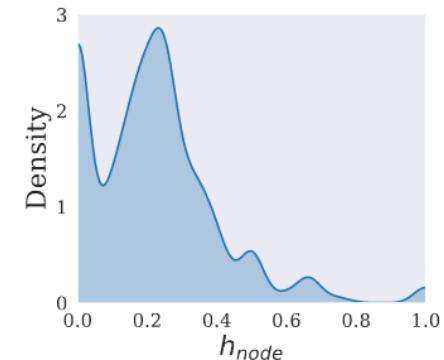
(a) PubMed ($h=0.79$)



(b) Ogbn-arxiv ($h=0.63$)



(c) Chameleon ($h=0.22$)



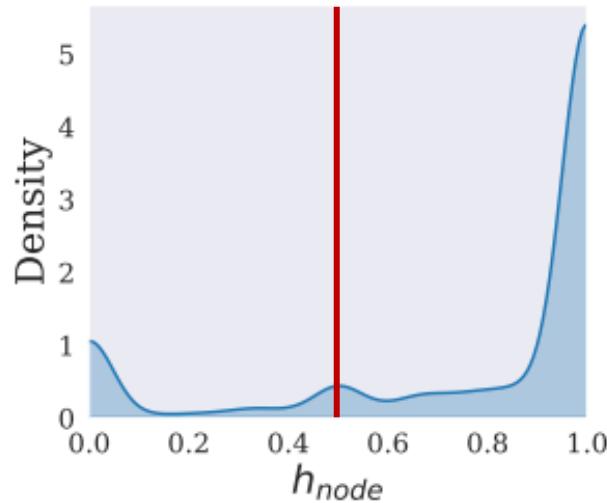
(d) Squirrel ($h=0.25$)

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

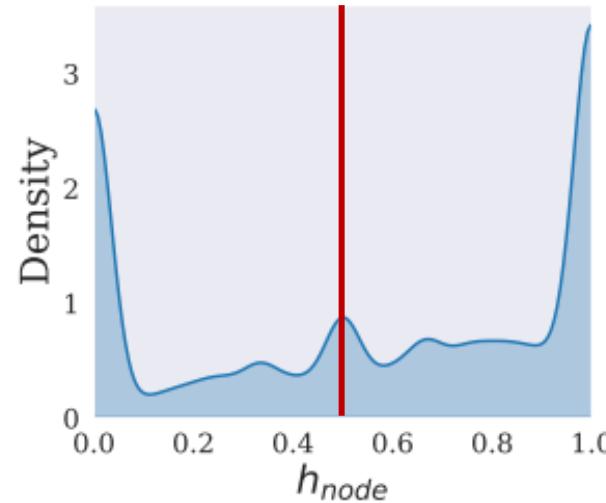
Structure disparity often happens in real world scenario!

Can one GNN fit all nodes?

Node homophily ratio distribution



(a) PubMed ($h=0.79$)



(b) Ogbn-arxiv ($h=0.63$)

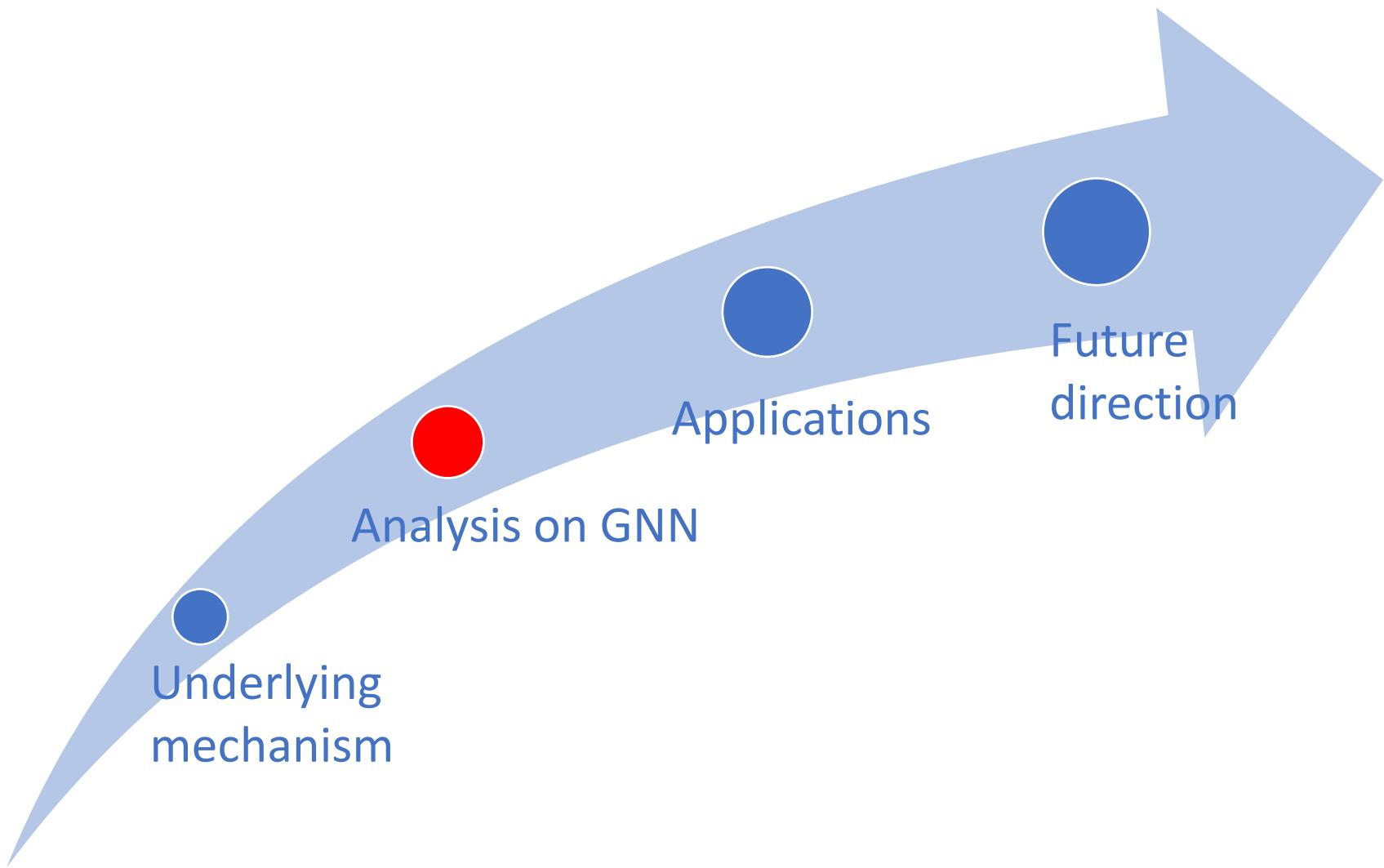
Majority pattern :

Homophily nodes in a homophily graph

Minority pattern :

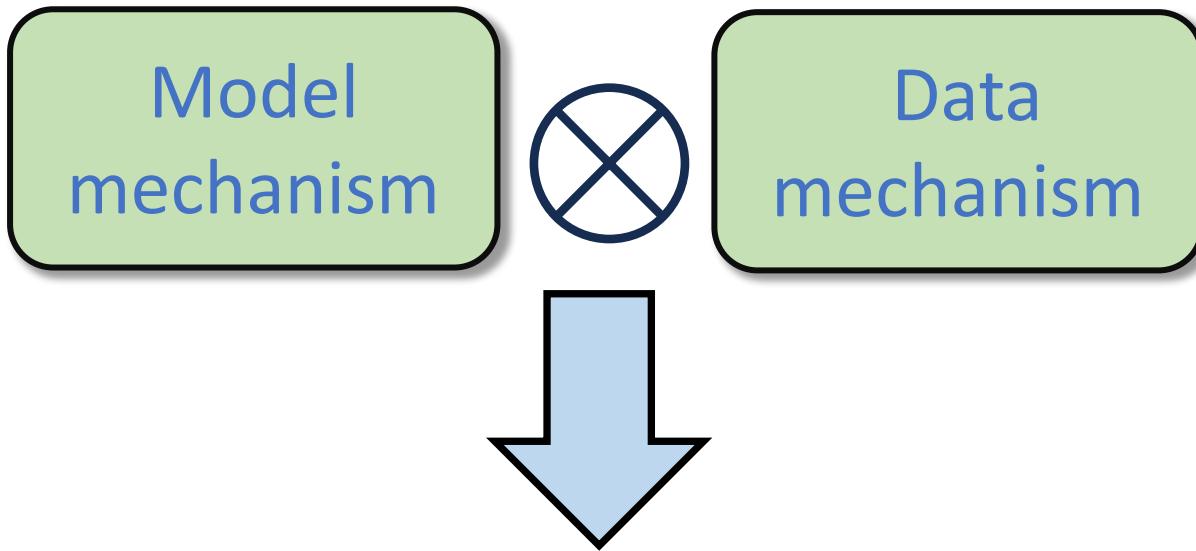
Heterophily nodes in a homophily graph

Outline



Analysis on GNNs

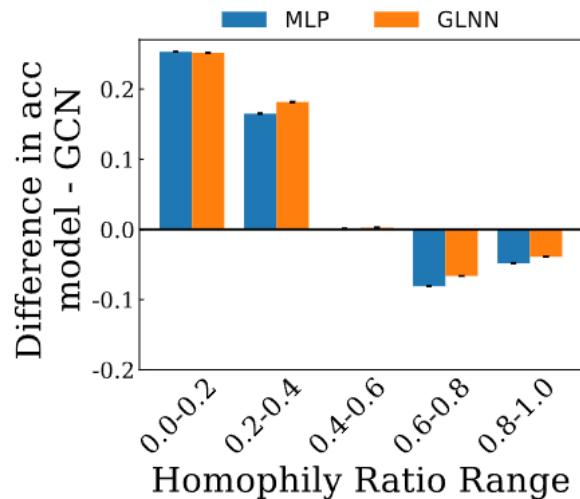
When do GNNs perform well on different nodes and when not?



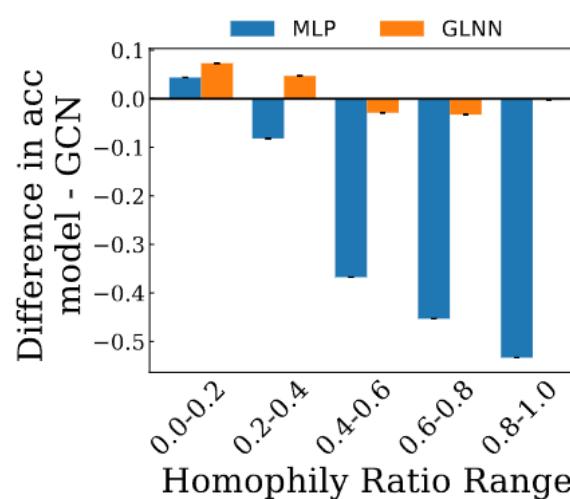
How is the performance?

Can one GNN fits all nodes?

Performance comparison between GCN and MLP-based models



(a) PubMed ($h=0.79$)

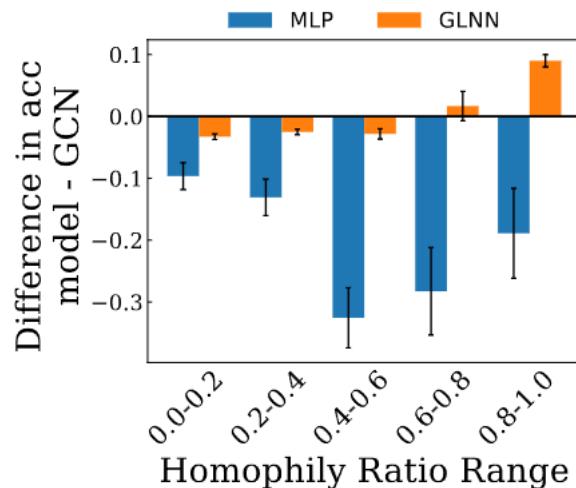


(b) Ogbn-arxiv ($h=0.63$)

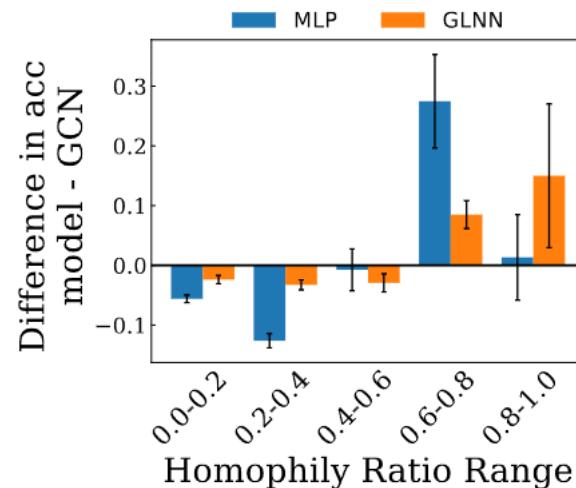
GCN outperforms on the majority pattern,
but fails in the minor pattern

Can one GNN 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 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 $\{\tilde{W}_l\}_{l=1}^L$, 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:*

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

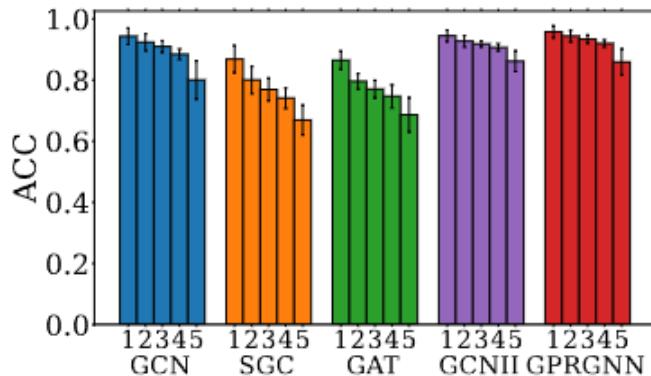
$\epsilon_m = \|f_i - f_j\|_F^2$ is the aggregated feature distance between train and test subgroup(s).

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

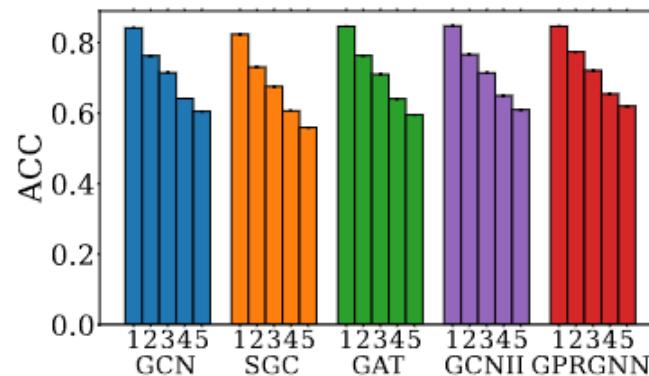
**Structure
disparity**

Empirical verification

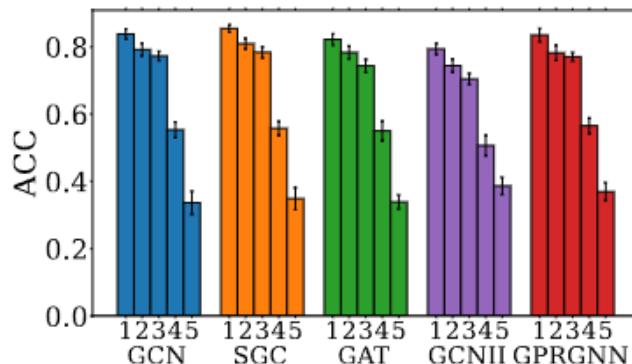
$$s = \epsilon_m + |h_{\text{tr}} - h_m|$$



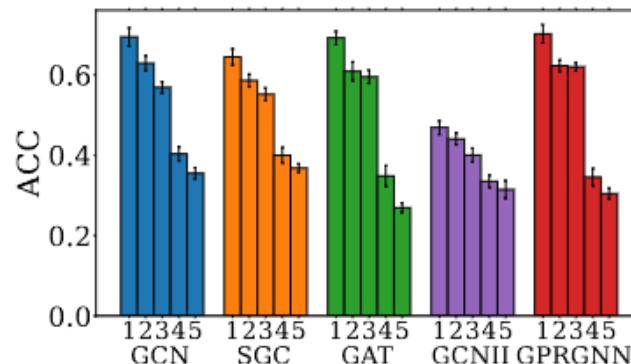
(a) PubMed ($h=0.79$)



(b) Ogbn-arxiv ($h=0.63$)



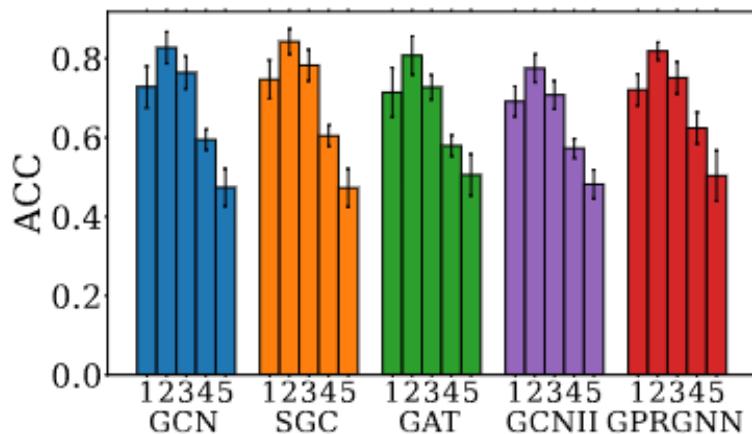
(c) Chameleon ($h=0.22$)



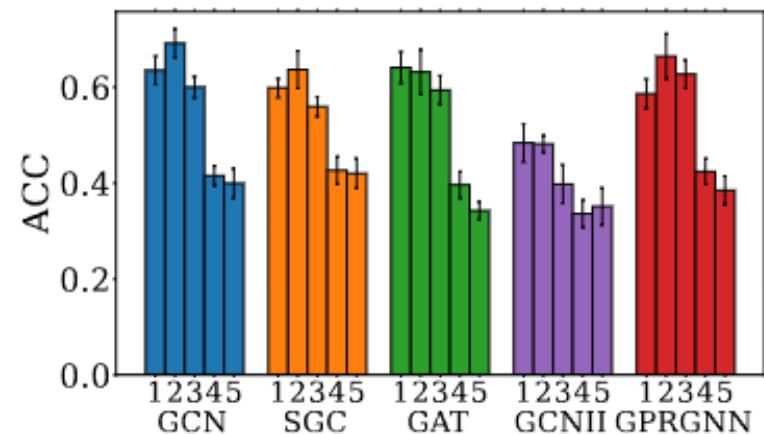
(d) Squirrel ($h=0.25$)

Empirical verification

$$s = \epsilon_m$$



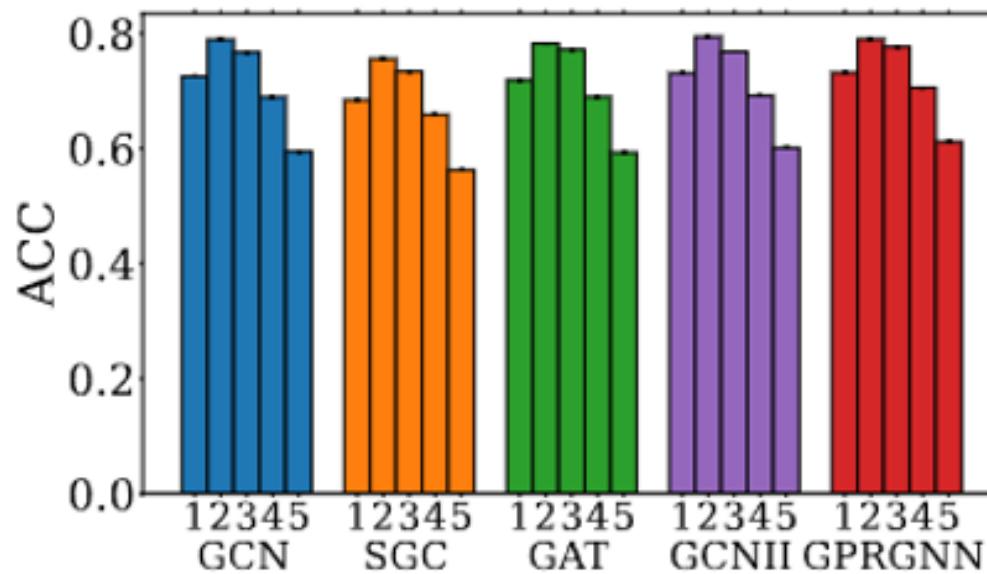
(c) Chameleon ($h=0.22$)



(d) Squirrel ($h=0.25$)

Empirical verification

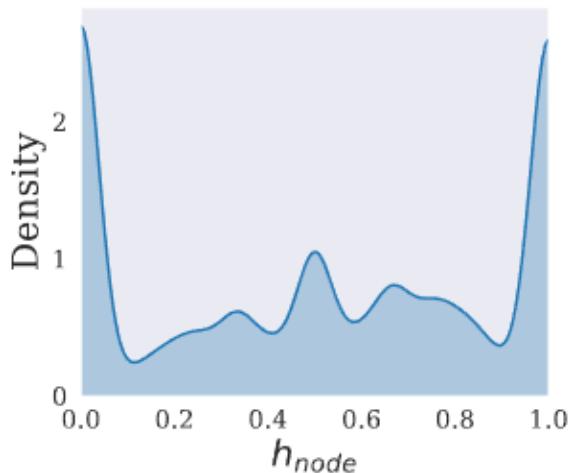
$$s = |h_{\text{tr}} - h_m|$$



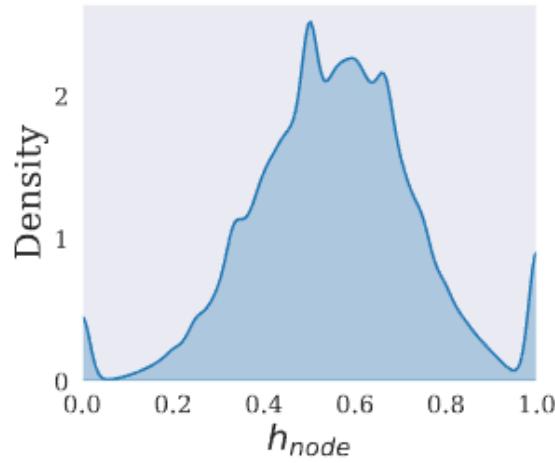
(b) Ogbn-arxiv ($h=0.63$)

Empirical success on more datasets

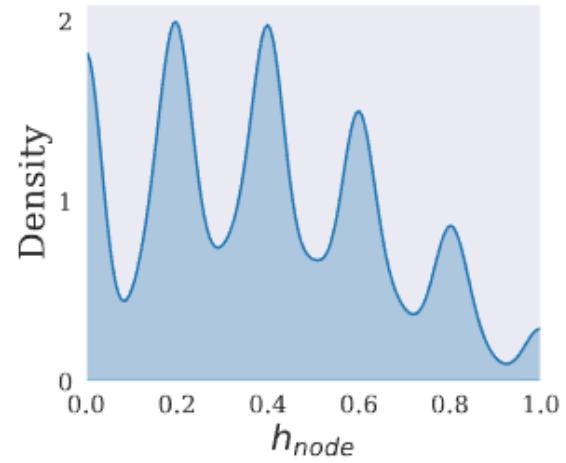
$$s = \epsilon_m + |h_{\text{tr}} - h_m|$$



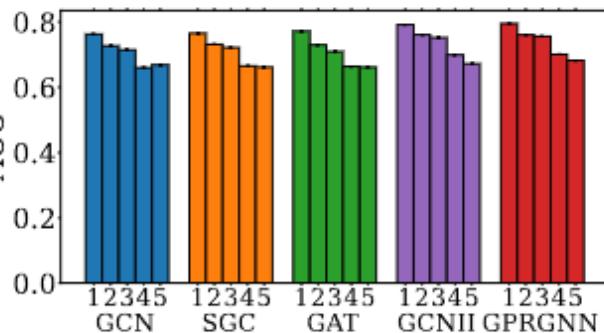
(e) IGB-tiny ($h=0.57$)



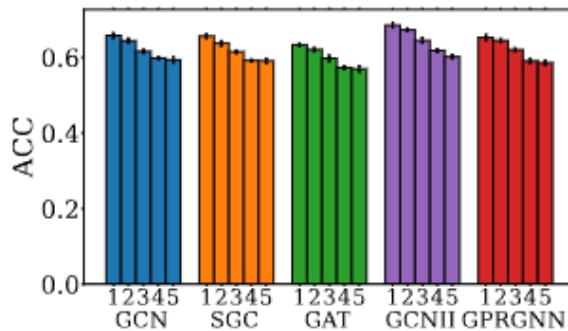
(i) Twitch-gamers ($h=0.56$)



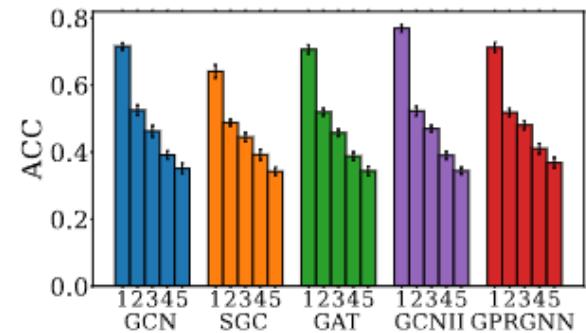
(j) Amazon-ratings ($h=0.38$)



(c) IGB-Tiny ($h=0.57$)

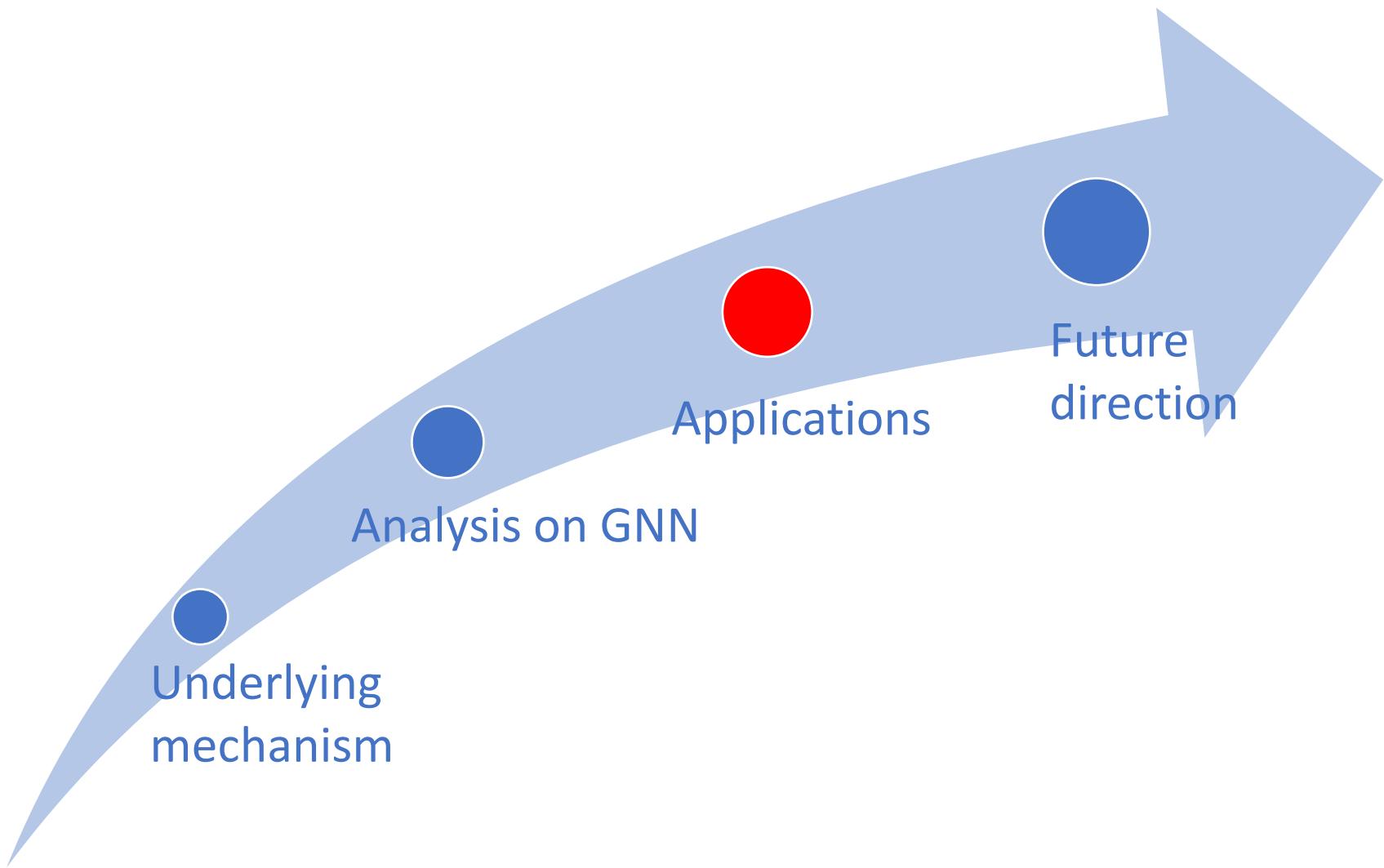


(e) Twitch-gamers ($h=0.56$)



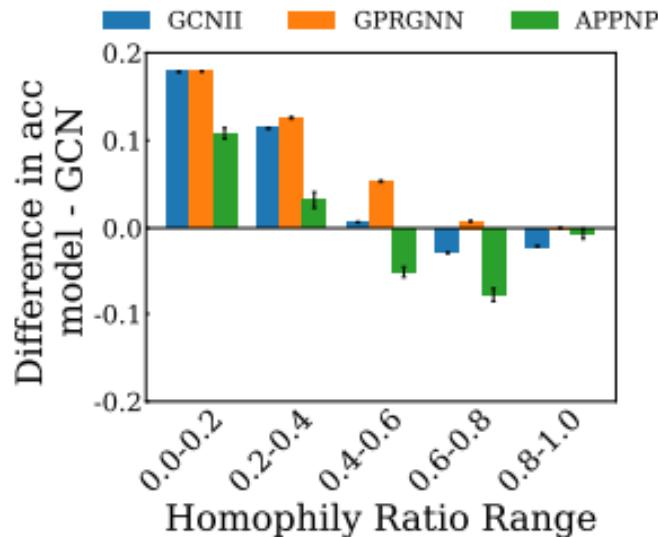
(f) Amazon-ratings ($h=0.38$)

Outline

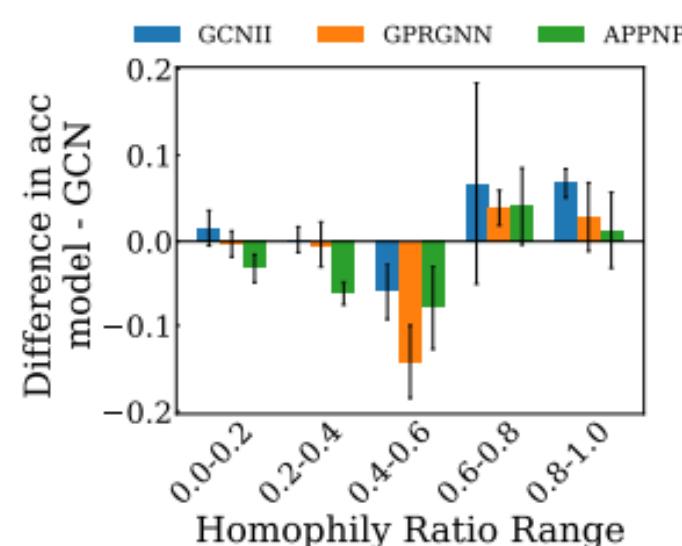


Elucidating the effectiveness of Deeper GNNs

Comparison between GCN and Deeper GNNs



(a) PubMed ($h=0.79$)



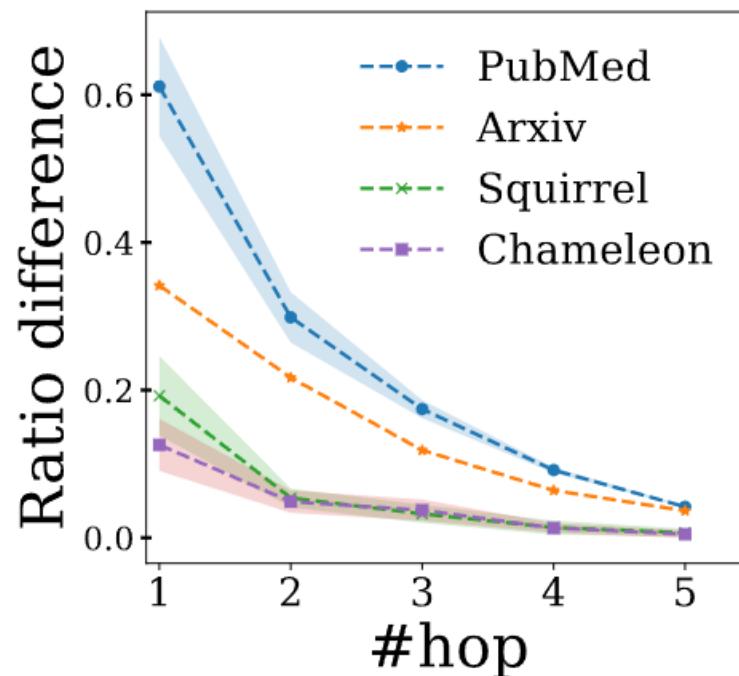
(c) Chameleon ($h=0.22$)

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

Model
mechanism



Data
mechanism

When they fit, GNNs can perform well

When not, GNNs underperform

Some Suggestions & look ahead

Before start, think carefully on your data property

Remember the drawback of GNNs. Not always good!

Solve the drawback of GNNs

Inspire new principled applications on GNNs

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.

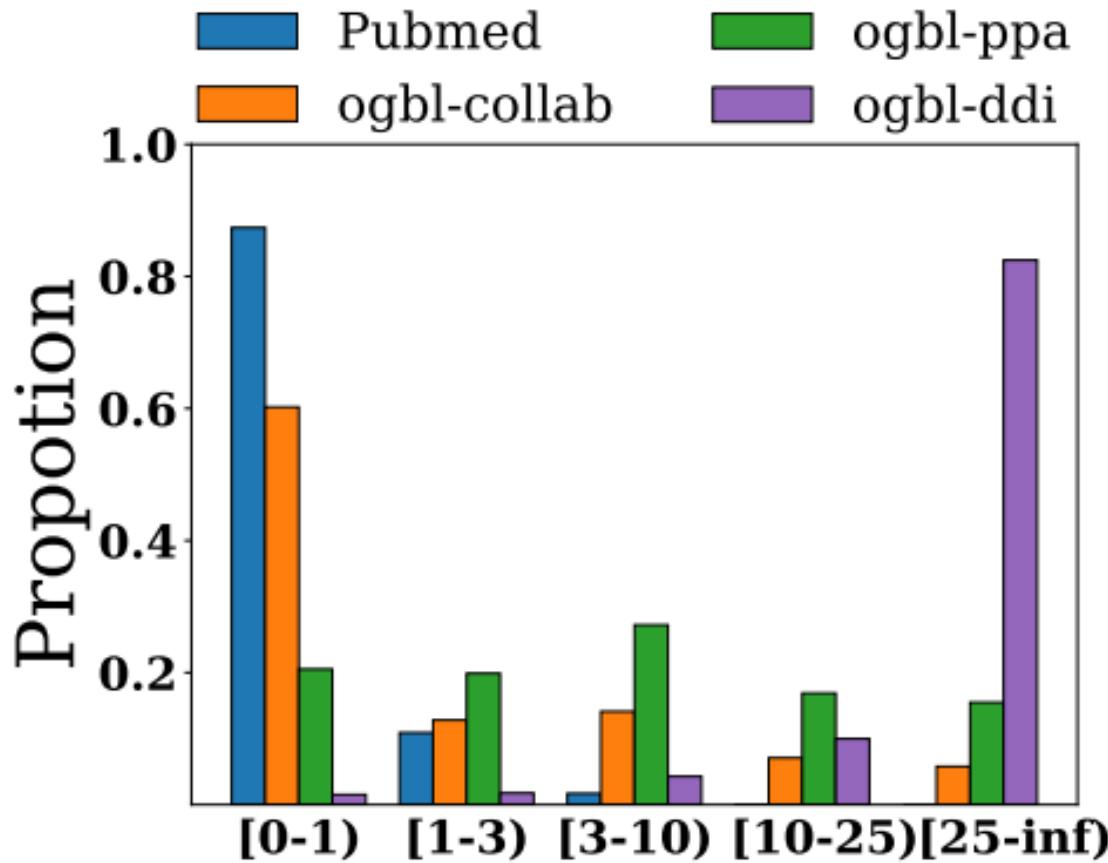


Figure 1: Distribution disparity of Common Neighbors across datasets.



(c) OGBL-DDI



(d) OGBL-PPA

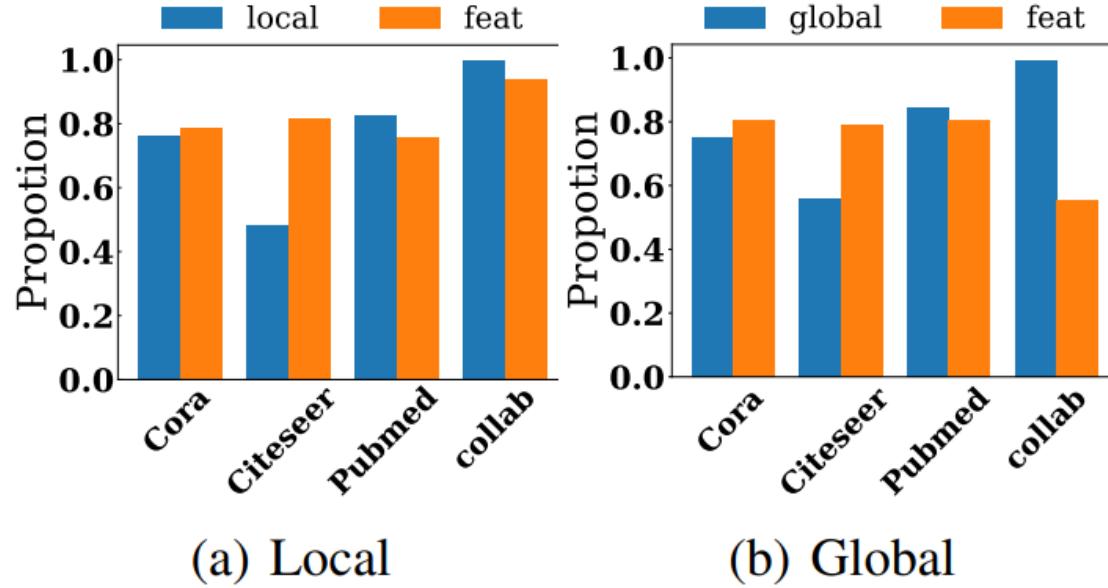


Figure 4: Distinctions between FP heuristic and structural heuristics. Each bar depicts the proportion of node pairs accurately predicted by the heuristic on one factor but not on the other one.

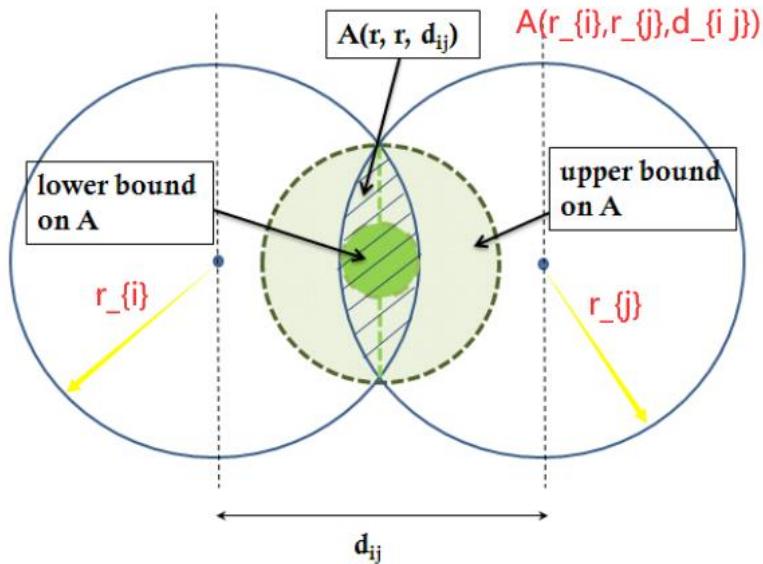
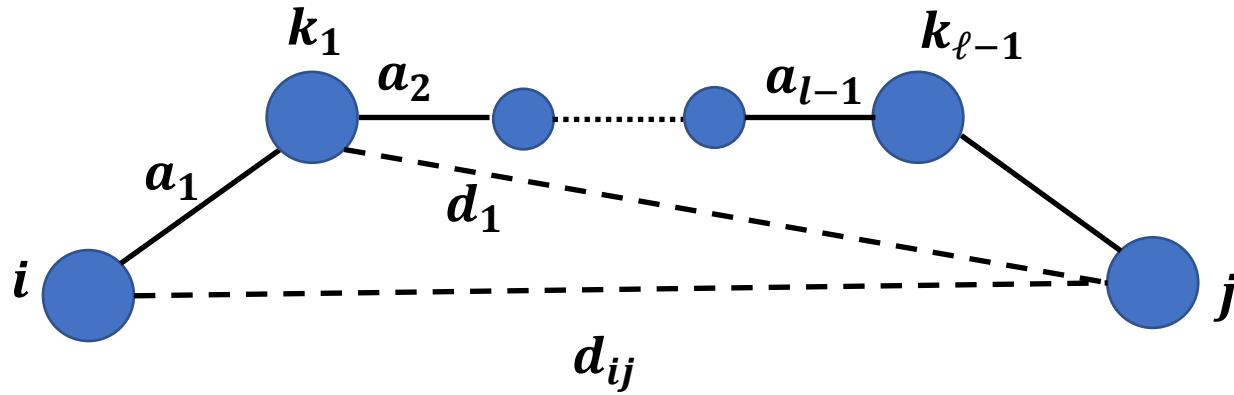


Figure 1: Common neighbors of two nodes must lie in the intersection $A(r, r, d_{ij})$.

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



$$\begin{aligned}
P_\ell(i, j) &= P(i \sim k_1 \sim \dots \sim k_{\ell-1} \sim j \mid d_{ij}) \\
&= P(a_1 \leq r'_1 \cap \dots \cap a_{\ell-1} \leq r'_{\ell-1} \mid d_{ij}) \\
&= \int_{d_1, \dots, d_{\ell-2}} P(a_1 \leq r'_1, \dots, a_{\ell-1} \leq r'_{\ell-1}, d_1, \dots, d_{\ell-2} \mid d_{ij}) \\
&= \int_{d_{\ell-2} = (d_{ij} - \sum_{n=0}^{\ell-3} r_n)_+}^{r_{\ell-1} + r_\ell} \dots \int_{d_1 = (d_{ij} - r_0)_+}^{\sum_{m=2}^\ell r_m} P(a_1 \leq r'_1, d_1 \mid d_{ij}) \dots P(a_{\ell-1} \leq r'_{\ell-1}, a_\ell \leq r'_\ell \mid d_{\ell-2}) \\
&\leq A\left(r'_1, \sum_{m=2}^\ell r_m, d_{ij}\right) \times A\left(r'_2, \sum_{m=3}^\ell r_m, (d_{ij} - r_0)_+\right) \times \dots \times A\left(r'_{\ell-1}, r_\ell, (d_{ij} - \sum_{n=0}^{\ell-3} r_n)_+\right) \\
&\leq \prod_{p=1}^{\ell-1} A\left(r'_p, \sum_{m=p+1}^\ell r_m, \left(d_{ij} - \sum_{n=0}^{p-2} r_n\right)_+\right)
\end{aligned} \tag{9}$$

Table 3: Selected Datasets Statistics.

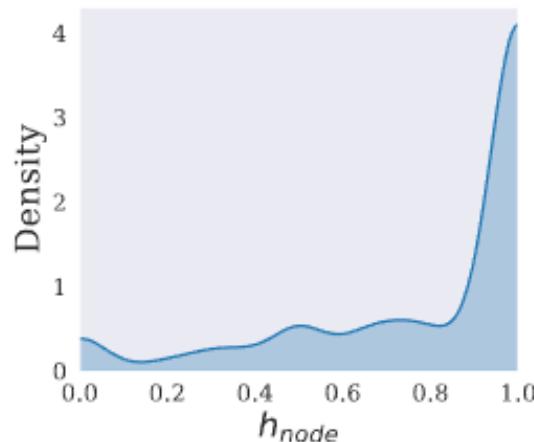
	Power	Reddit	Amazon Photo	Flicker
#Nodes	4,941	232,965	7,650	334,863
#Edges	6,594	114,615,892	238,162	899,756
#Feature	NA	602	745	500
Mean Degree	2.67	98.38	6.21	5.69
Split Ratio	80/10/10	80/10/10	80/10/10	80/10/10
Domains	Transport	Social	Web	Social

Table 4: Web Domain Datasets Statistics.

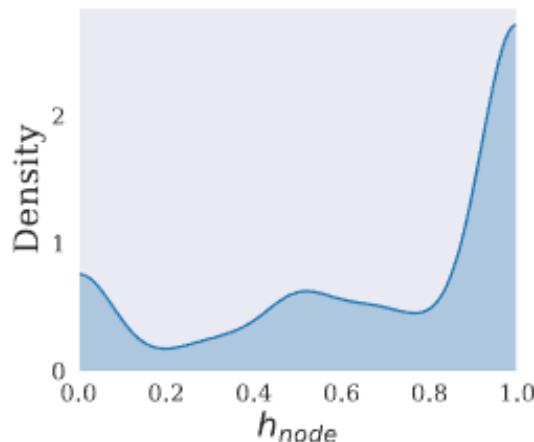
	PB	Email-Enron	Amazon Photo	Amazon	Google
#Nodes	1,222	36,692	7,650	334,863	875,713
#Edges	16,714	183,831	238,162	899,756	5,105,039
Mean Degree	27.36	10.02	6.21	5.69	116.49
Split Ratio	80/10/10	80/10/10	80/10/10	80/10/10	80/10/10
Domains	Web	Web	Web	Web	Web

Appendix for Node classification

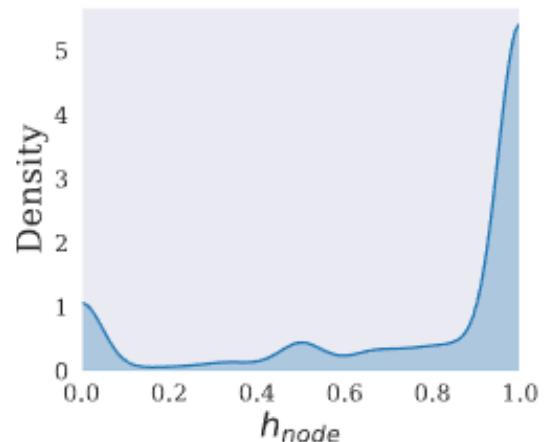
Homophily ratio distribution



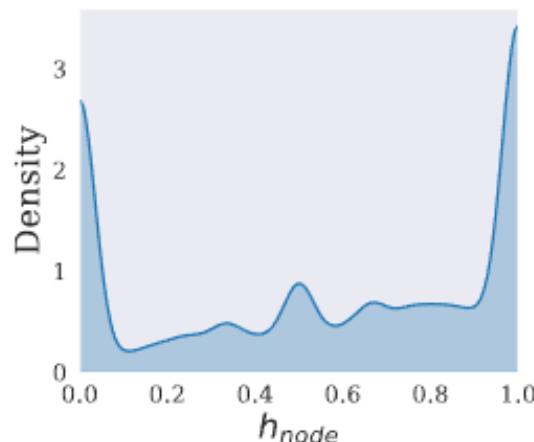
(a) Cora ($h=0.81$)



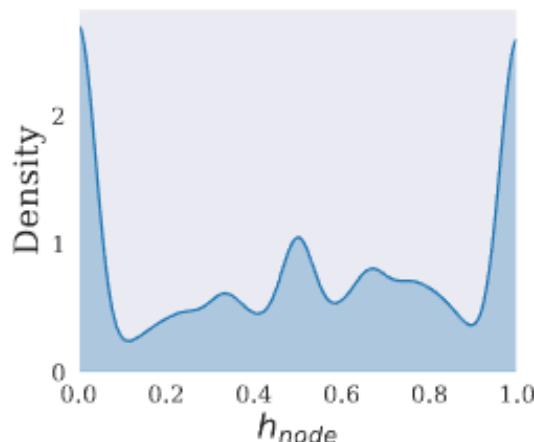
(b) CiteSeer ($h=0.71$)



(c) PubMed ($h=0.79$)

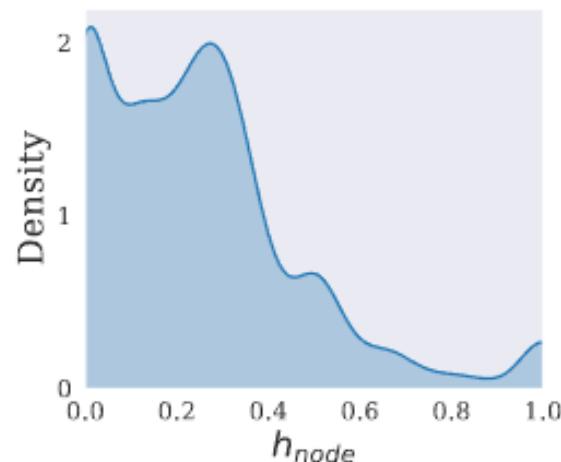


(d) Ogbn-Arxiv ($h=0.63$)

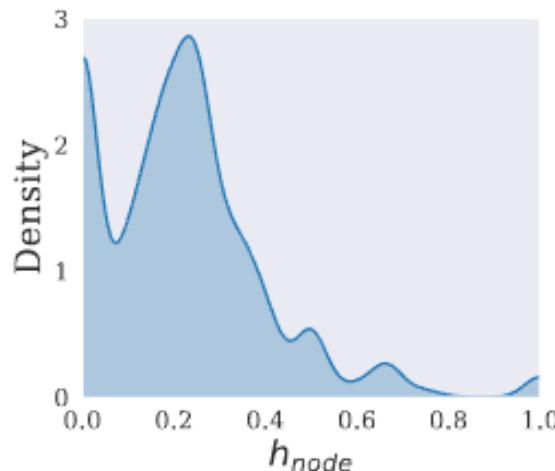


(e) IGB-tiny ($h=0.57$)

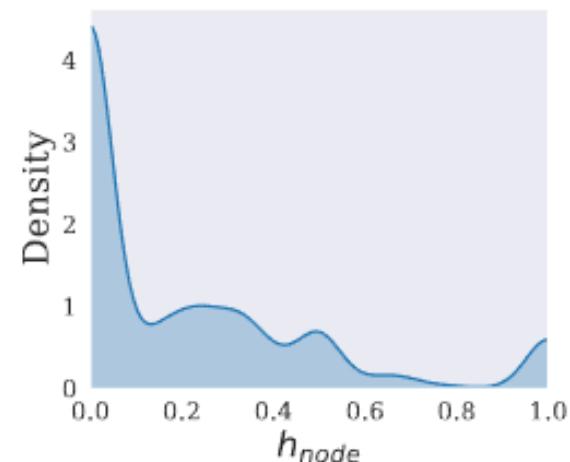
Homophily ratio distribution



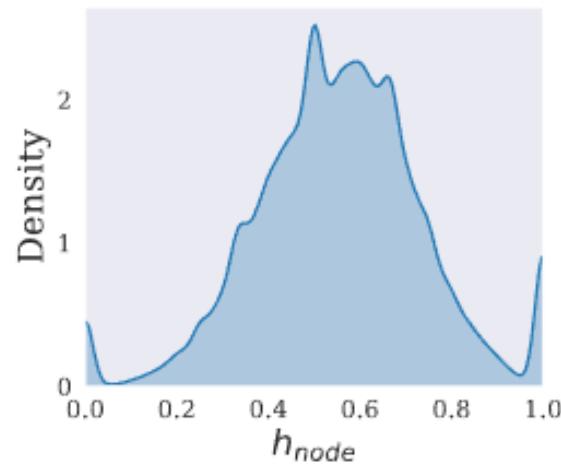
(f) Chameleon ($h=0.25$)



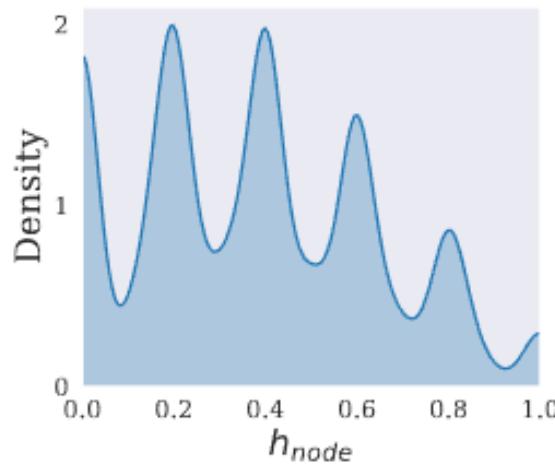
(g) Squirrel ($h=0.22$)



(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($\mu_1, \mu_2, (p^{(1)}, q^{(1)}), (p^{(2)}, q^{(2)}), \text{Pr(homo)}$)). *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 $V_m \sim \text{CSBM}(\mu_1, \mu_2, p^{(i)}, q^{(i)})$, 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 [13] with the following components: (1) a one-hop mean aggregation function g with $g(X, G)$ denoting the output. (2) MLP feature transformation $f(g_i(X, G); W_1, W_2, \dots, W_L)$, where f is a ReLU-activated L -layer MLP with W_1, \dots, W_L as parameters for each layer. The largest width of all the hidden layers is denoted as b .*

Addition theoretical analysis on linear separability

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

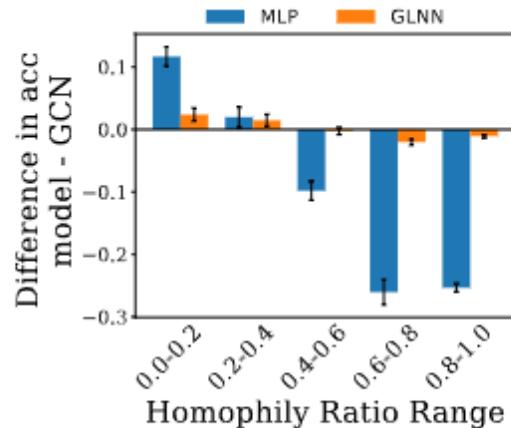
Improved linear separability can be found with $d_i > 100$

Empirical verification

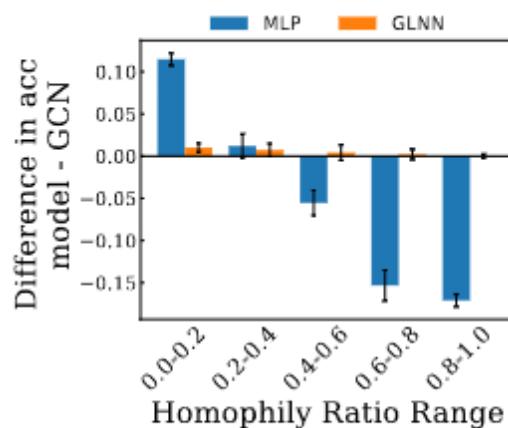
Table 6: The performance of logistic regression algorithm on homophilic nodes, heterophilic nodes, and a mixture of homophilic and heterophilic nodes. The results on the first row and first column correspond to the performance on homophilic nodes and heterophilic nodes, solely.

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

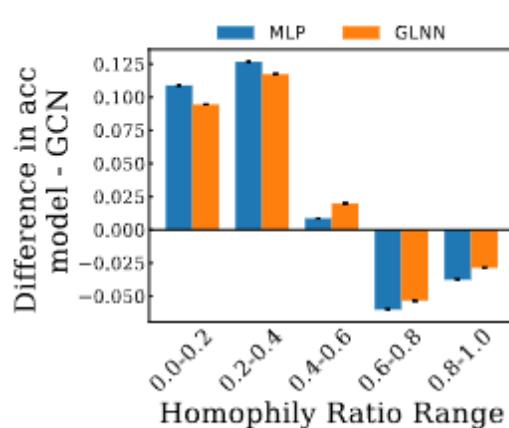
Additional comparison between GCN and MLP



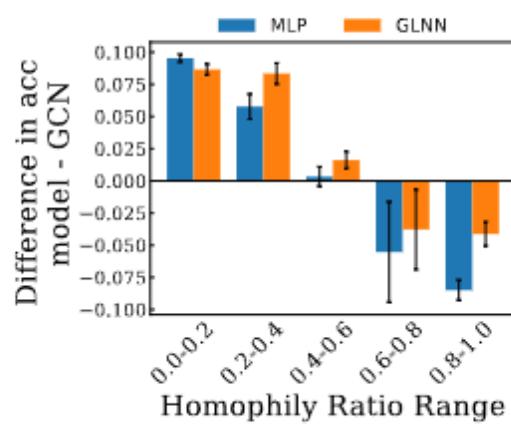
(a) Cora ($h=0.81$)



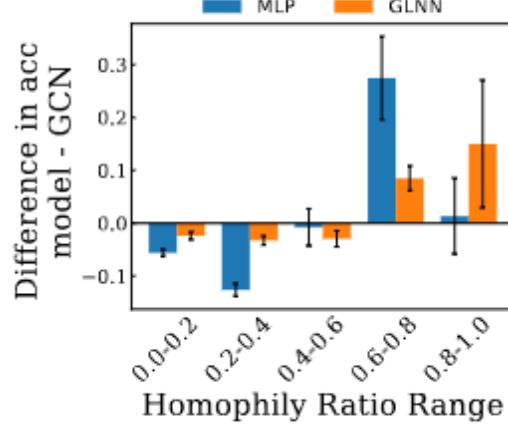
(b) CiteSeer ($h=0.71$)



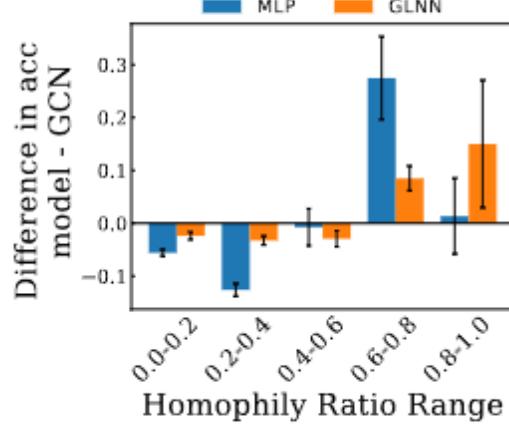
(c) IGB-tiny ($h=0.58$)



(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

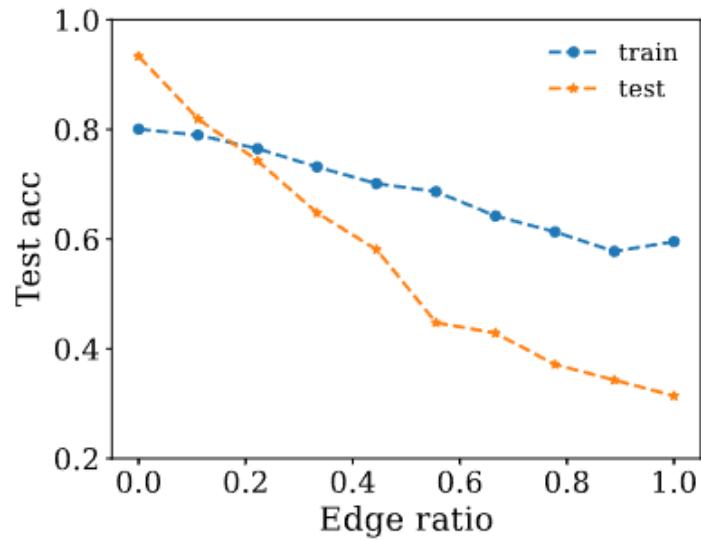
$\mathcal{D}_0 : \text{Categorical}([0, 0.5, 0, 0, 0, 0.5]),$
 $\mathcal{D}_1 : \text{Categorical}([0.5, 0, 0.5, 0, 0, 0]),$
 $\mathcal{D}_2 : \text{Categorical}([0, 0.5, 0, 0.5, 0, 0]),$
 $\mathcal{D}_3 : \text{Categorical}([0, 0, 0.5, 0, 0.5, 0]),$
 $\mathcal{D}_4 : \text{Categorical}([0, 0, 0, 0.5,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

nodes with **high**
homophily

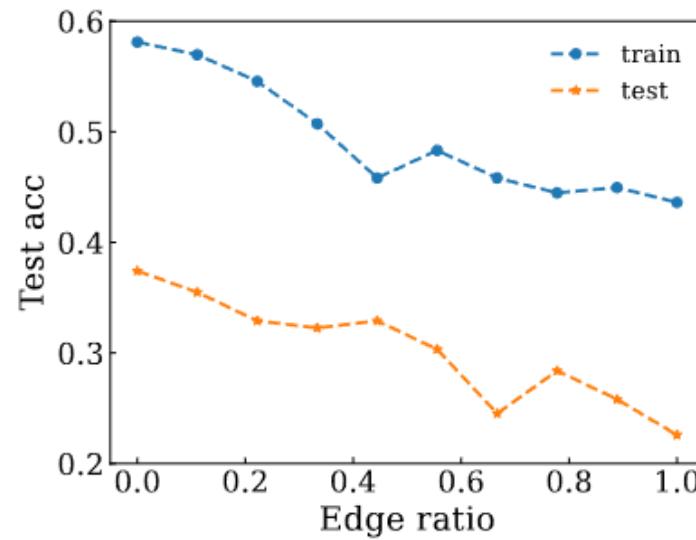
Adding K across-class edges

nodes with
low homophily

Additional synthetic analysis

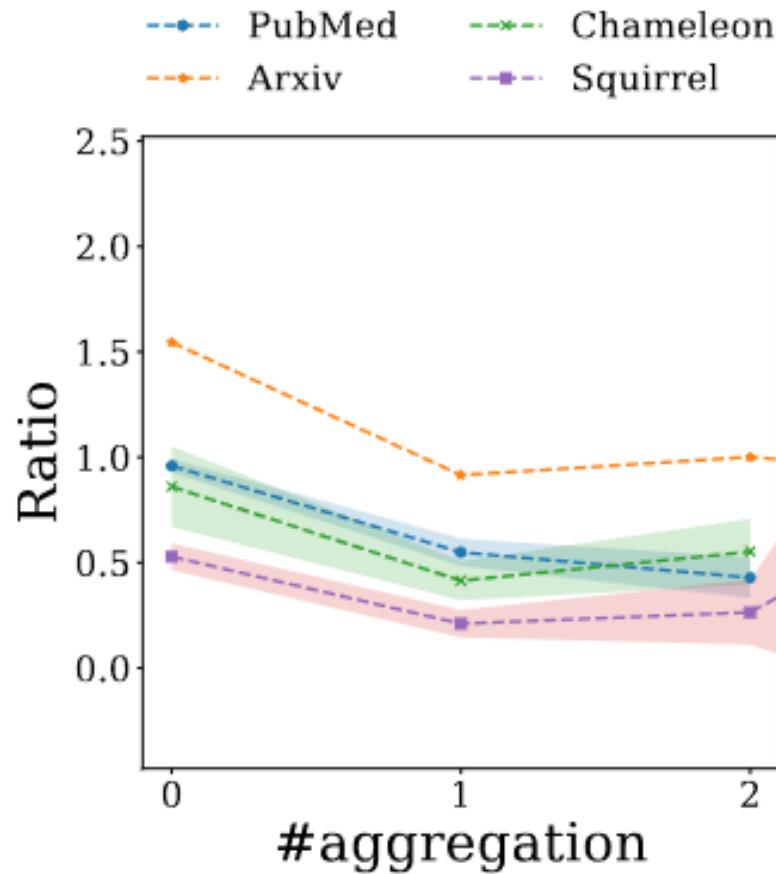


(a) Synthetic graphs generated from Cora with the targeted heterophilic edge algorithm



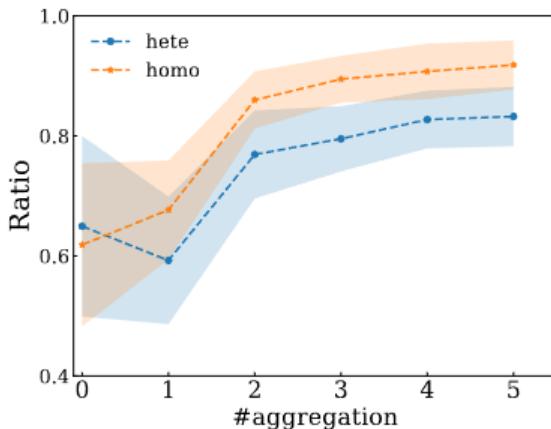
(b) Synthetic graphs generated from Squirrel with the targeted homophilious edge algorithm

Additional discriminative analysis on GCN

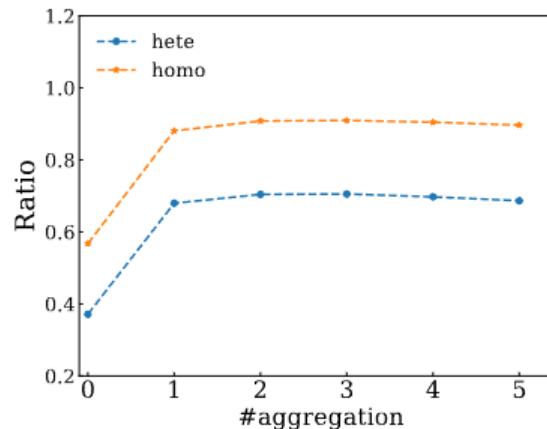


Additional local discriminative analysis

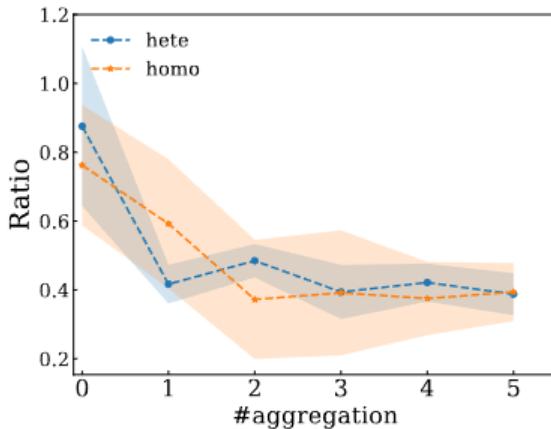
$$r = \frac{\sum_{v \in V_{\text{te}}} \mathbb{1} [\exists c \in \mathcal{C}, |\mathcal{M}_v^c| > \frac{k}{2}]}{|V_{\text{te}}|}$$



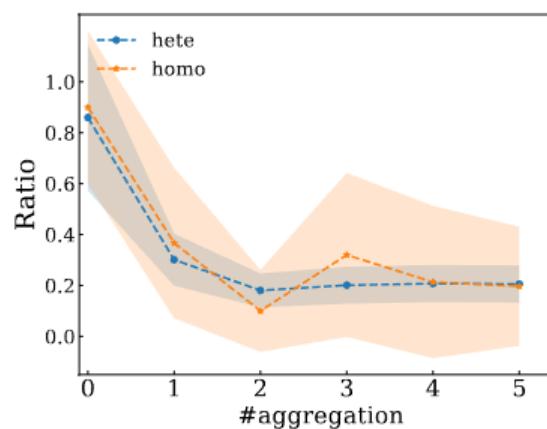
(c) PubMed ($h=0.79$)



(d) Ogbn-arxiv ($h=0.63$)



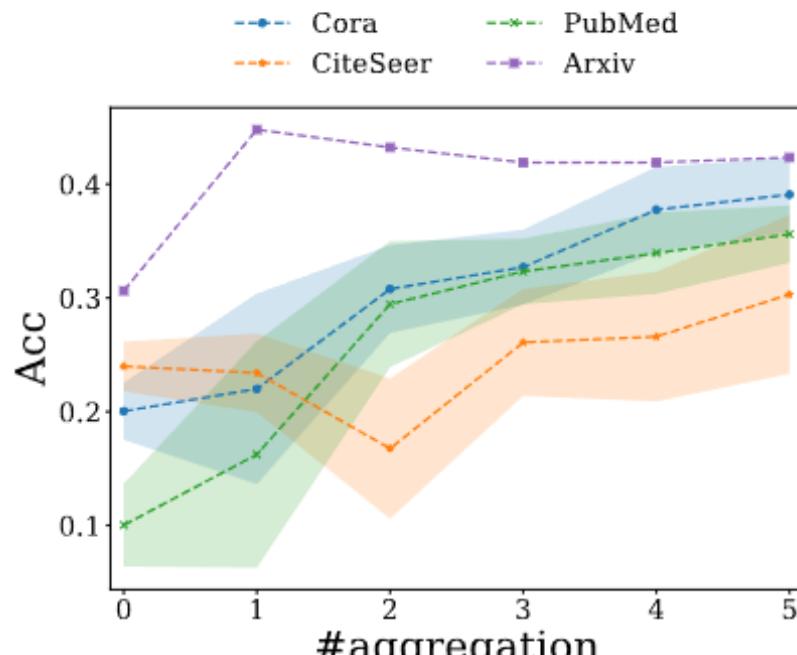
(e) Chameleon ($h=0.25$)



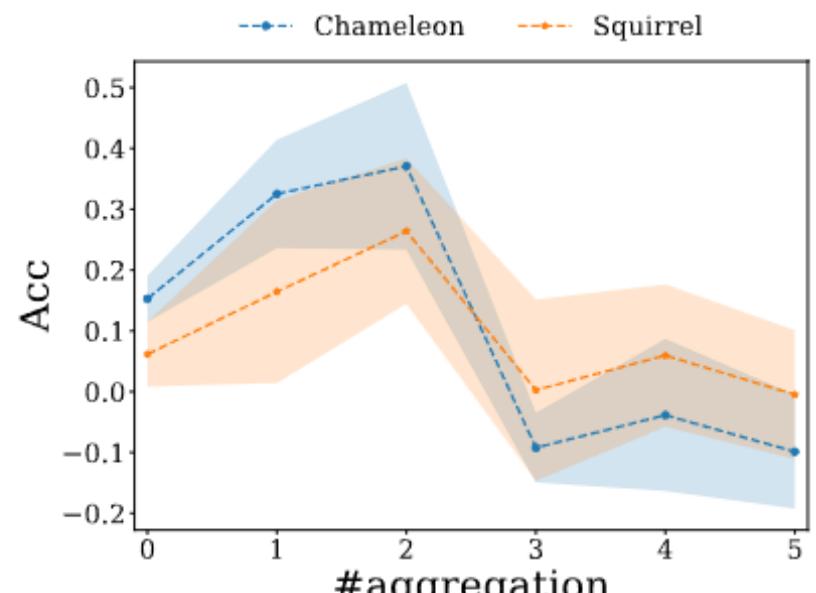
(f) Squirrel ($h=0.22$)

Additional local discriminative analysis

$$\text{Acc}_{\text{local}} = \frac{\sum_{v \in V_{\text{agree}}} \mathbb{1} [c_v = c_{N_v}]}{|V_{\text{agree}}|}$$



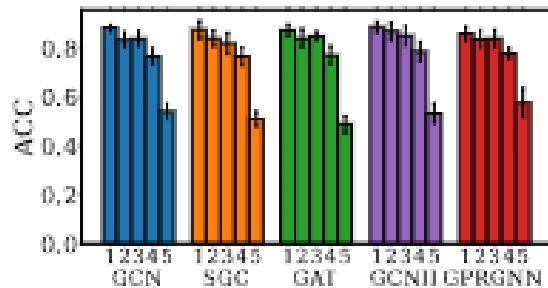
(a) Homophilic graphs



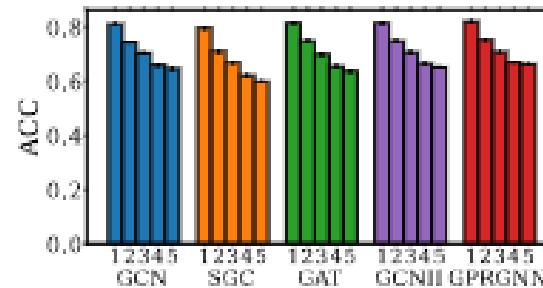
(b) Heterophilic graphs

Additional higher-order performance disparity

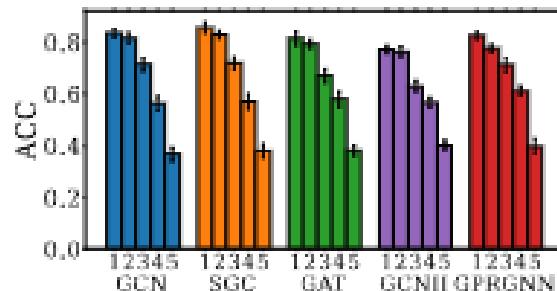
$$s = \epsilon_m + |h_{\text{tr}} - h_m|$$



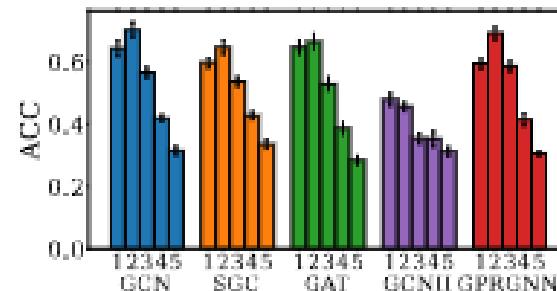
(a) PubMed



(b) Ogbn-arxiv



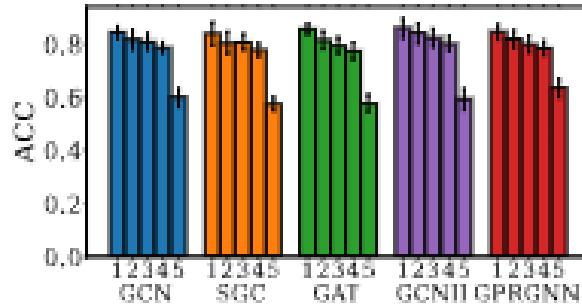
(c) Chameleon



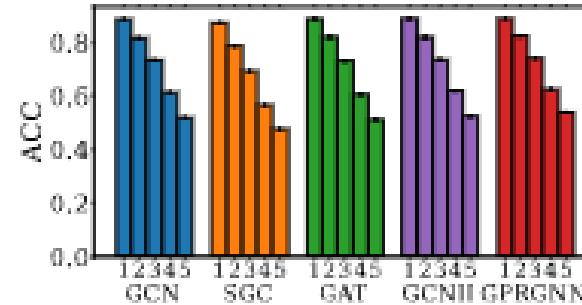
(d) Squirrel

Additional higher-order performance disparity

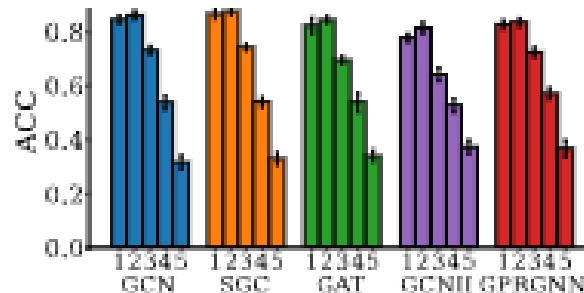
$$s = \epsilon_m$$



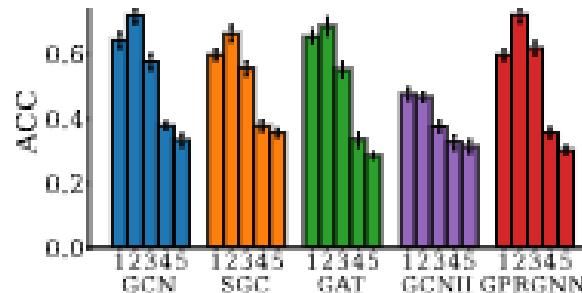
(a) PubMed



(b) Ogbn-arxiv



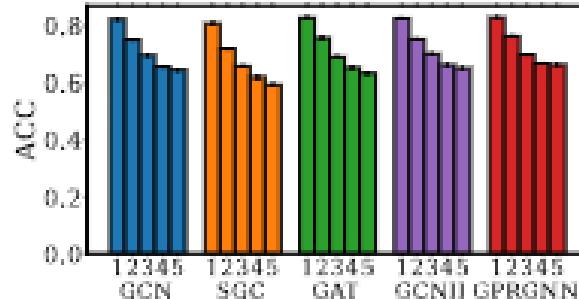
(c) Chameleon



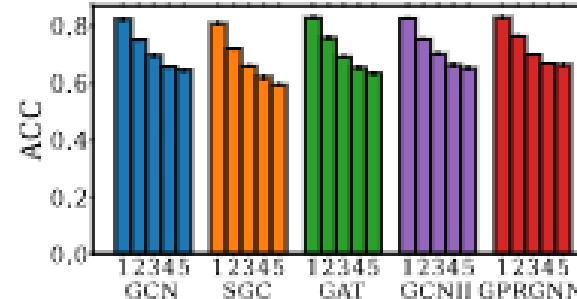
(d) Squirrel

Additional higher-order performance disparity

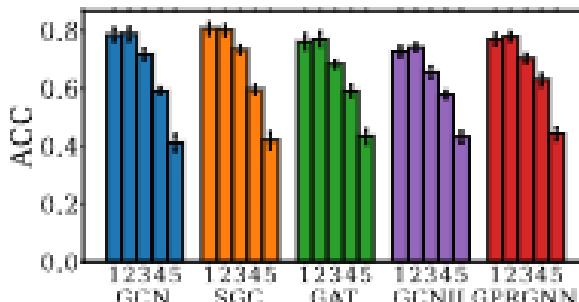
$$s = |h_{\text{tr}} - h_m|$$



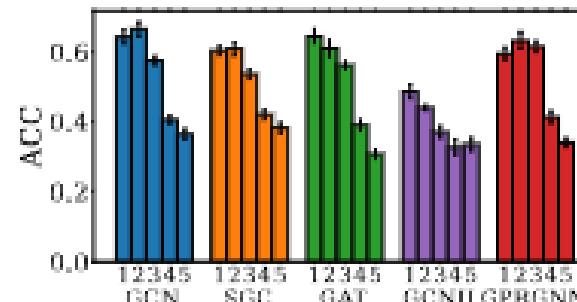
(a) PubMed



(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 [6]

	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