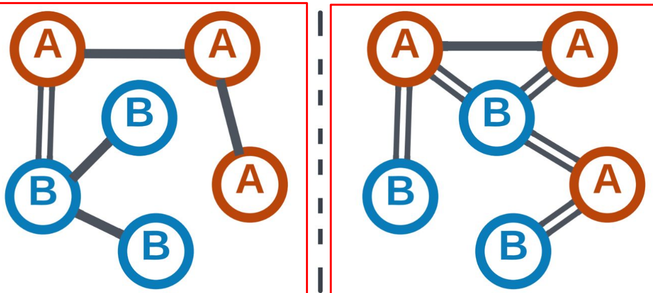
Finding the Missing-half: Graph Complementary Learning for Homophily-prone and Heterophily-prone Graphs

Yizhen Zheng, He Zhang, Vincent Lee, Yu Zheng, Xiao Wang, Shirui Pan

Existing problem



(a) Homophilic-prone Graph

(b) Heterophilic-prone Graph

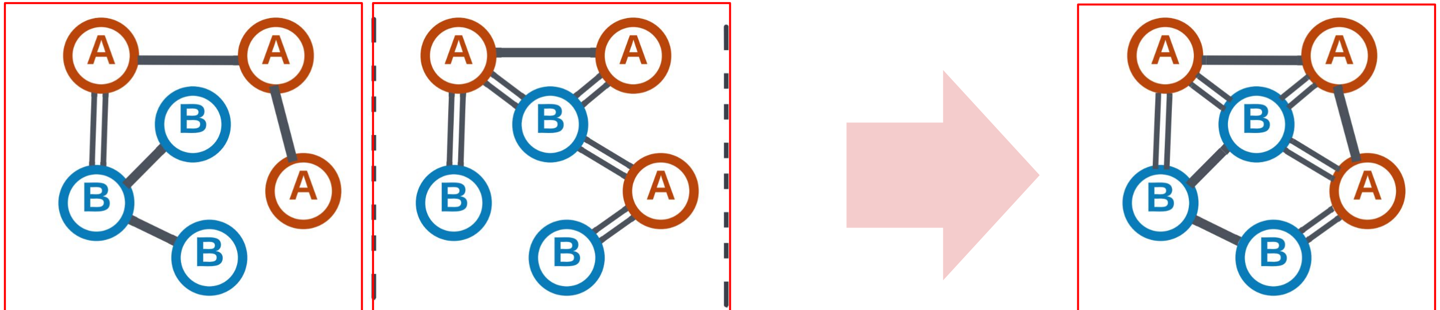
Real-World Graph have only one kind of tendency.

(a) Connecting same class → Homophilic-prone
(b) Connecting different class -> Heterophilic-prone

Existing GNNs only take the original graph during training.

Existing problem

Could we find the missing-half to complement the graph? 🤔



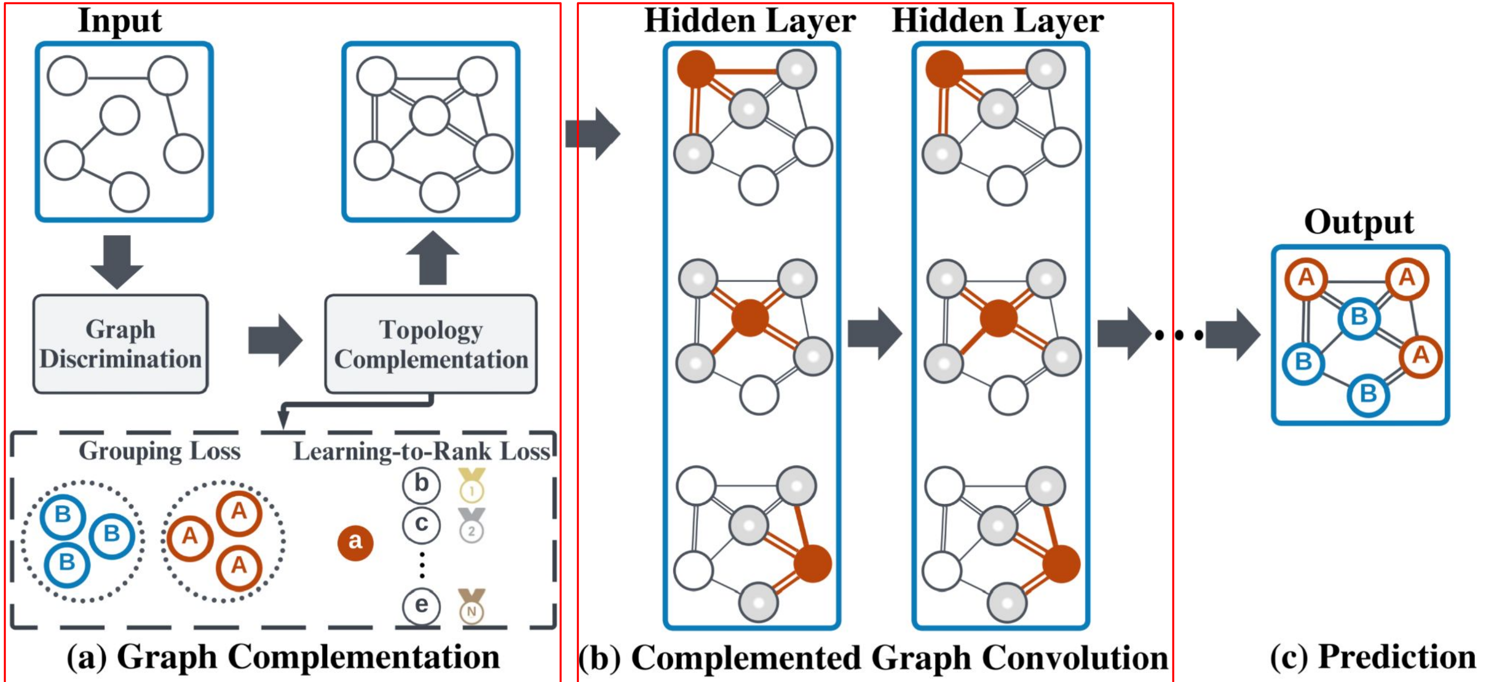
(a) Homophilic-prone Graph

(b) Heterophilic-prone Graph

(c) Complemented Graph

Enriched Structural Information!!

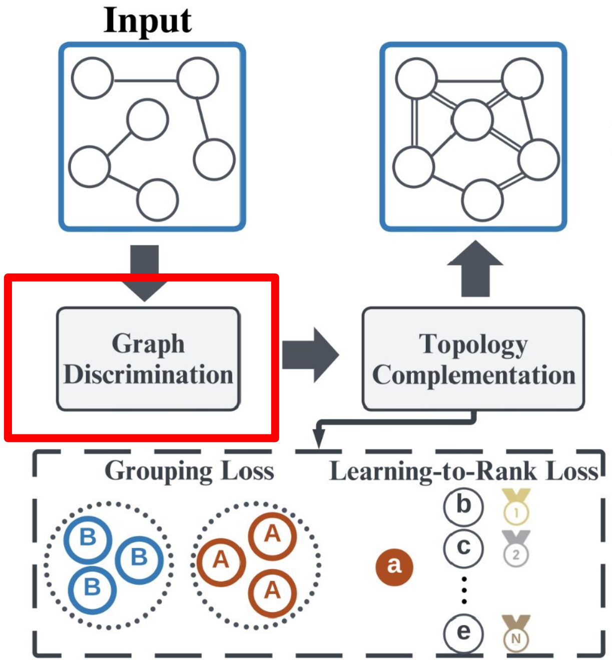
The GOAL Framework



Find the missing-half

New Conv for Complemented Graph

The GOAL Framework

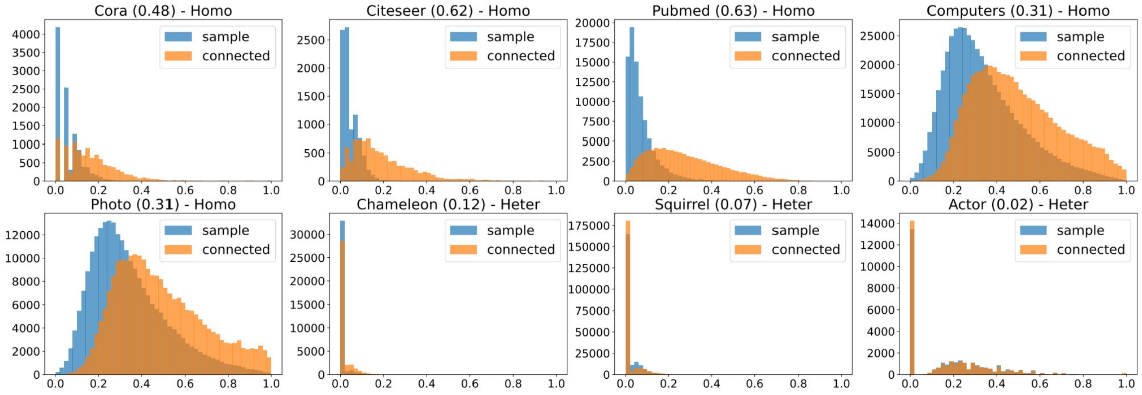


(a) Graph Complementation

Find the missing-half

Graph Discrimination:

Find which part is missing for a graph?



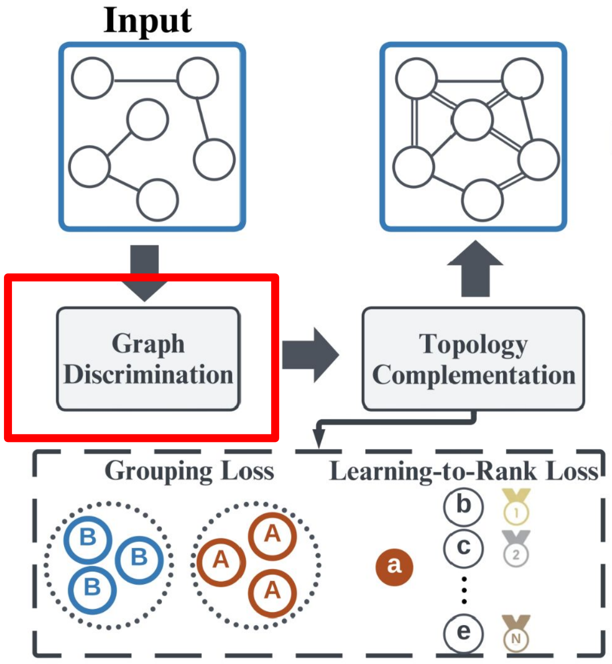
Average Cosine Similarity

(Random sampled pairs)

VS

(Connected node pairs)

The GOAL Framework

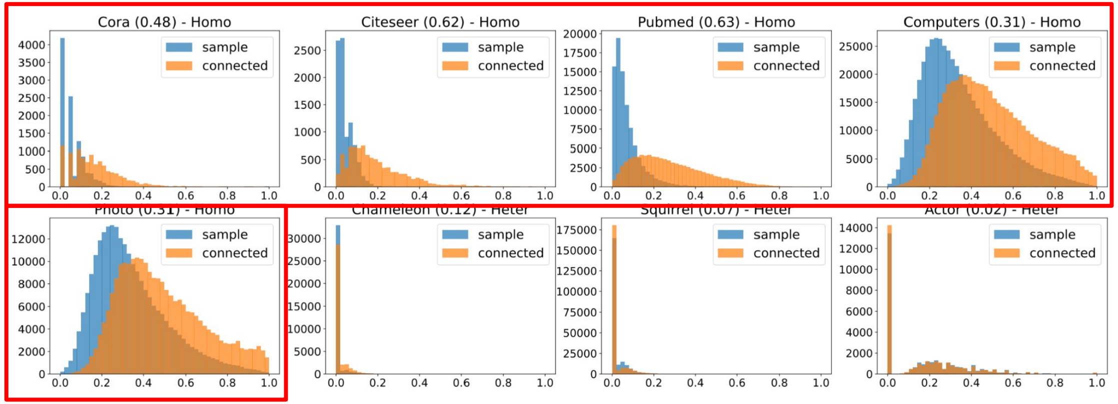


(a) Graph Complementation

Find the missing-half

Graph Discrimination:

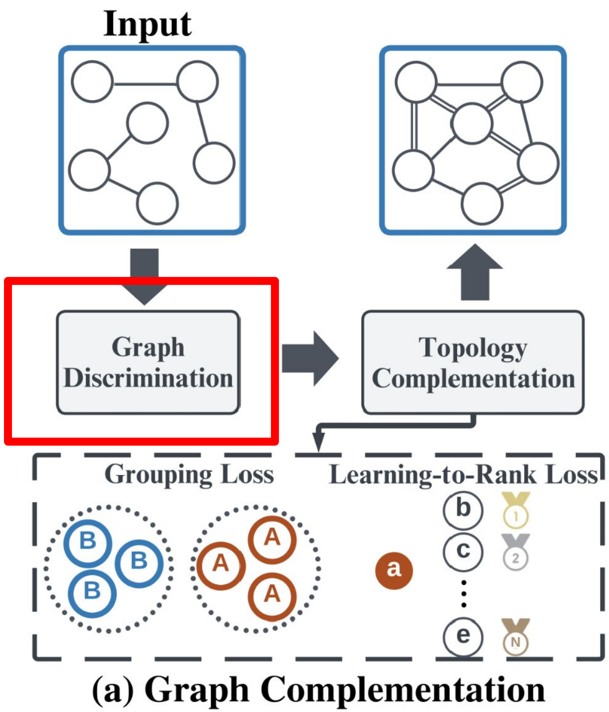
Find which part is missing for a graph?



Homophilic-prone graph

Connected different from Sampled!

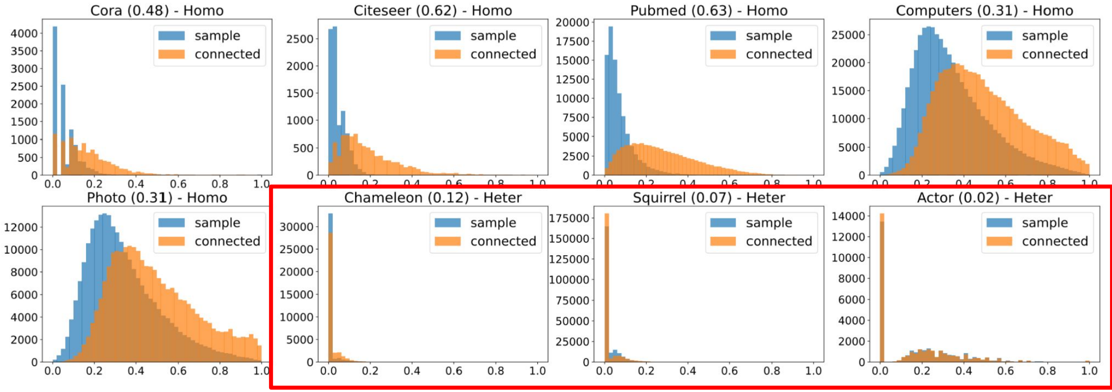
The GOAL Framework



Find the missing-half

Graph Discrimination:

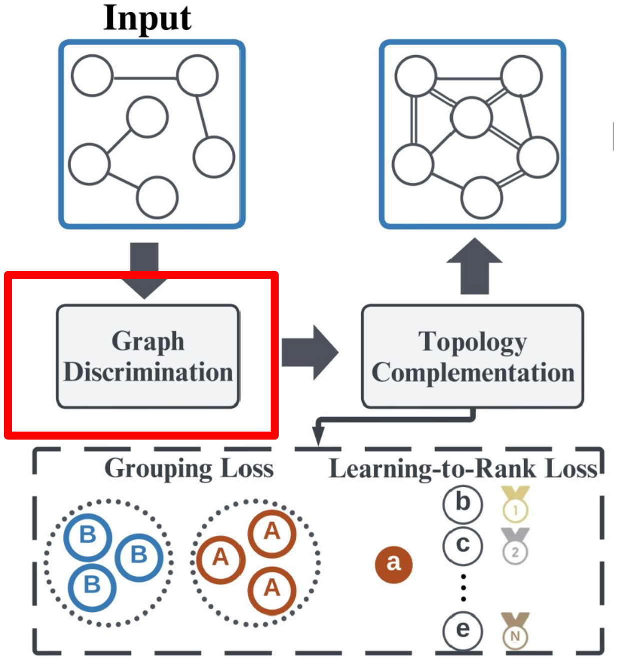
Find which part is missing for a graph?



Heterophilic-prone graph

Connected similar to Sample!

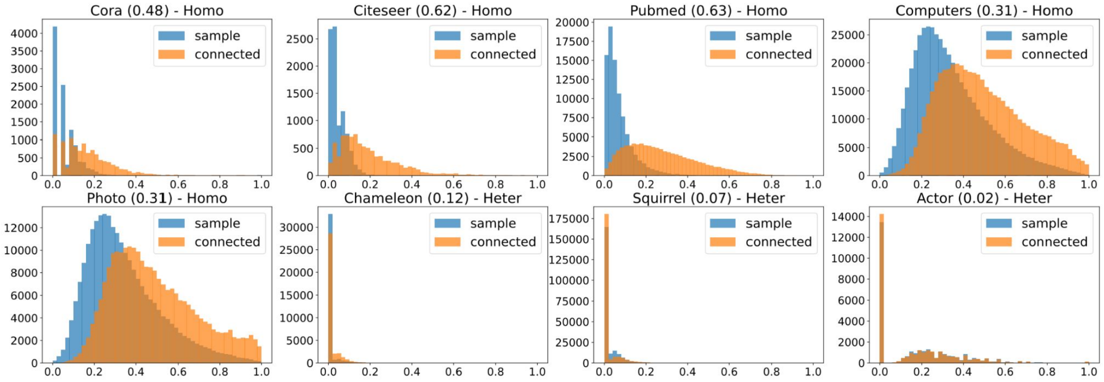
The GOAL Framework



(a) Graph Complementation

Find the missing-half

Graph Discrimination:
Find which part is missing for a graph?



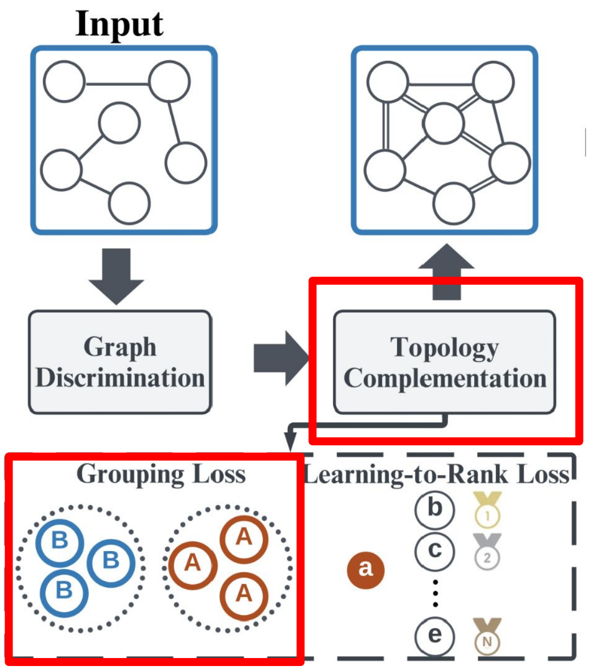
K-S statistics (evaluation of distribution difference) > 0.2

→ homophilic-prone graph

→ need to find heterophilic-prone structure

→ Vice versa

The GOAL Framework

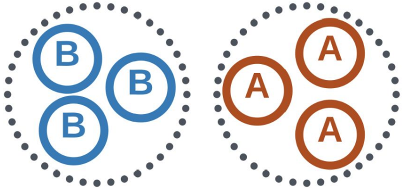


(a) Graph Complementation

Find the missing-half

Topology Complementation:
Find the missing-half graph structure.

Grouping Loss

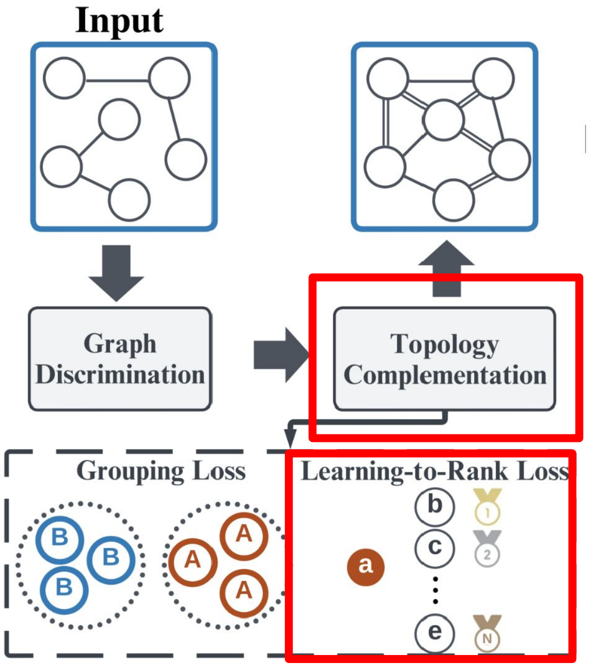


Maximise same class node similarity
Minimise different class node similarity

$$\mathcal{L}_{pos} = -log(sig(\frac{\sum_{(i,j)\in\mathbb{P}} \mathbf{Z}_{gc}^{T}[v_i] \cdot \mathbf{Z}_{gc}[v_j]}{|\mathbb{P}|}) + \epsilon),$$

$$\mathcal{L}_{neg} = -log(1 - sig(\frac{\sum_{(i,j)\in\mathbb{N}} \mathbf{Z}_{gc}^{T}[v_i] \cdot \mathbf{Z}_{gc}[v_j]}{|\mathbb{N}|} + \epsilon)),$$

The GOAL Framework



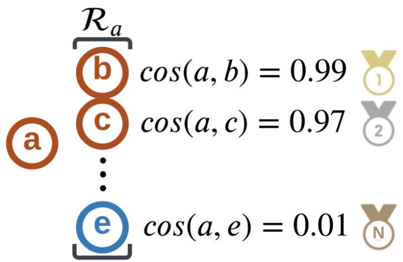
(a) Graph Complementation

Find the missing-half

Topology Complementation:

Find the missing-half graph structure.

Learning-to-rank loss



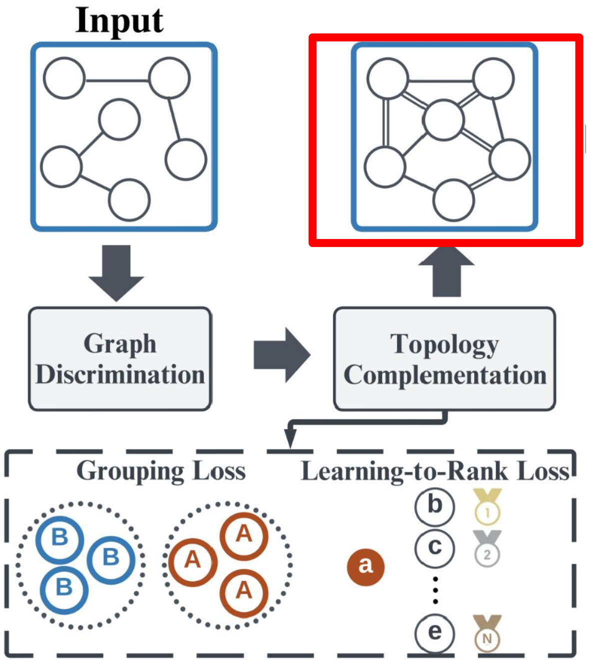
For node a

→ most similar k same class nodes

→ least similar k different class nodes

A 2k list → put nodes in correct rank

The GOAL Framework



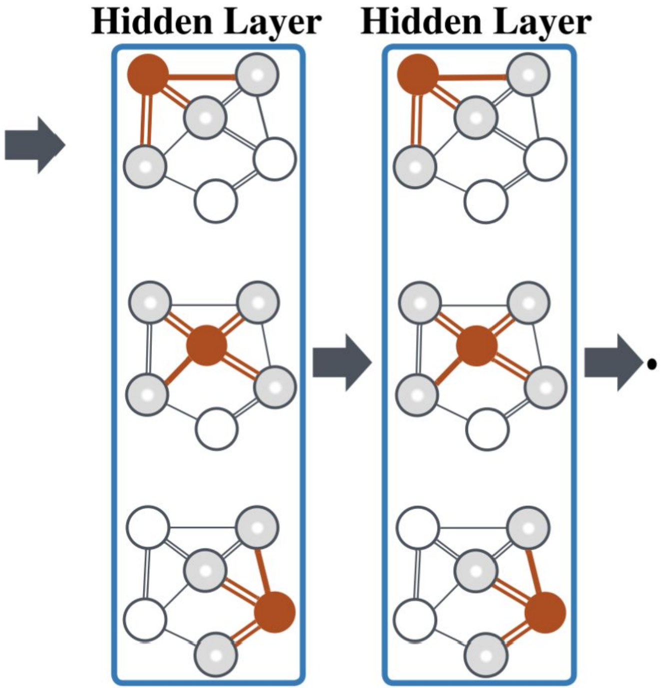
(a) Graph Complementation

Find the missing-half

Topology Complementation:
Find the missing-half graph structure.

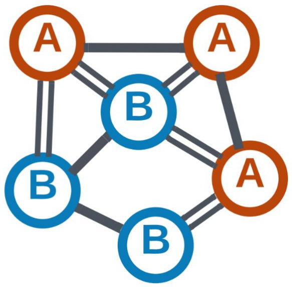
Two losses to train a MLP
→
Generate node embeddings
→
Find each node most similar/least similar neighbours (cosine sim)
→
Connect them
→
Done!

The GOAL Framework



(b) Complemented Graph Convolution

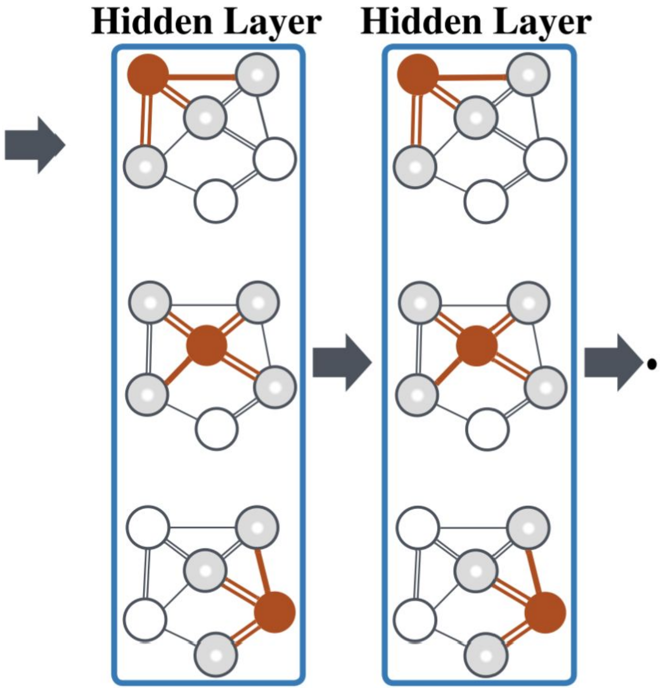
New Conv for Complemented Graph



(c) Complemented Graph

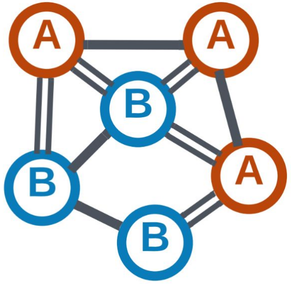
How to design a new graph convolution → handle complemented graphs with both homophily- and heterophily-prone topology? 🤔

The GOAL Framework



(b) Complemented Graph Convolution

New Conv for Complemented Graph

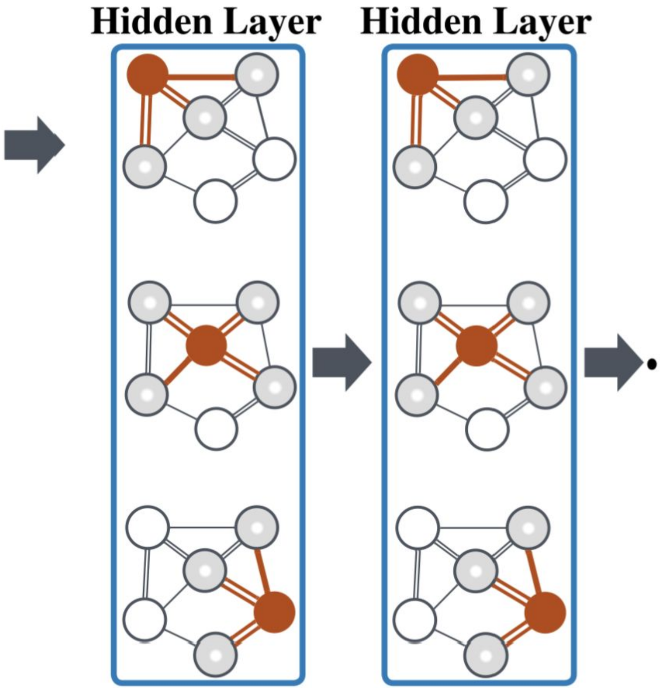


(c) Complemented Graph

Straightforward Solution:

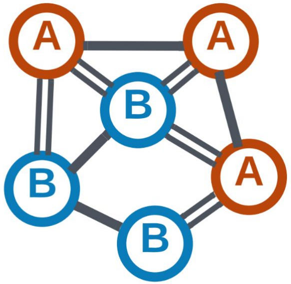
Maximise similarity for **homophilic-connected** nodes
Minimise similarity for **heterophilic-connected** nodes

The GOAL Framework



(b) Complemented Graph Convolution

New Conv for Complemented Graph



(c) Complemented Graph

Two Objectives

Maximise sim for **homophilic-connected**

$$\mathcal{O}_o = \min_{\mathbf{H}}\{tr(\mathbf{H}^T(\mathbf{I} - \hat{\mathbf{A}}_o)\mathbf{H})\}$$

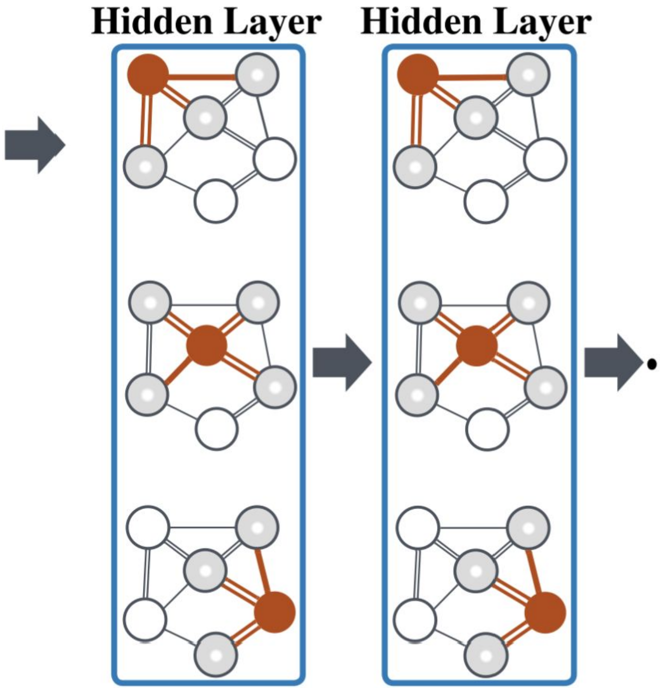
$$= \frac{1}{2}\sum_{(i,j)\in\mathcal{E}}^{E} \boxed{\hat{\mathbf{A}}_o}[i,j] \parallel \mathbf{H}_i - \mathbf{H}_j \parallel^2,$$

Minimise sim for **heterophilic-connected**

$$\mathcal{O}_t = \min_{\mathbf{H}}\{tr(\mathbf{H}^T(\mathbf{I} + \hat{\mathbf{A}}_t)\mathbf{H})\}$$

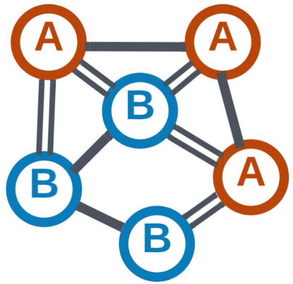
$$= \frac{1}{2}\sum_{(i,j)\in\mathcal{E}}^{E} \boxed{\hat{\mathbf{A}}_t}[i,j] \parallel \mathbf{H}_i + \mathbf{H}_j \parallel^2,$$

The GOAL Framework



(b) Complemented Graph Convolution

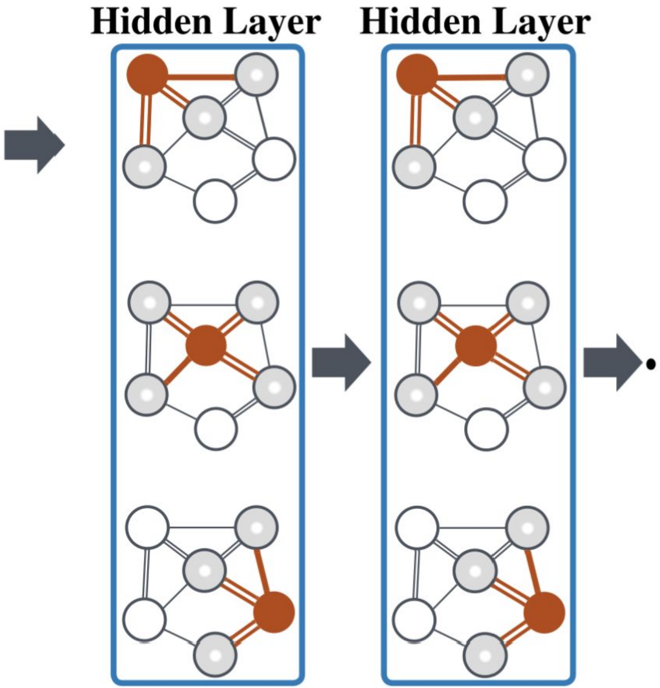
New Conv for Complemented Graph



(c) Complemented Graph

A step forward: **any other objective**? 🤔

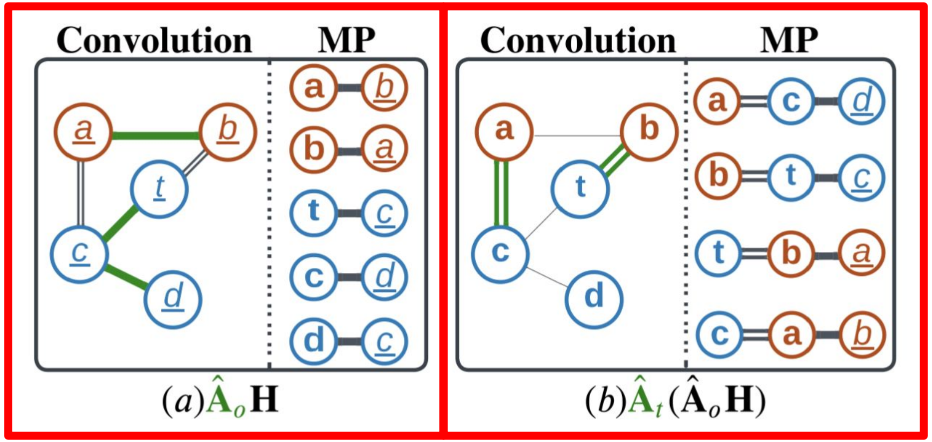
The GOAL Framework



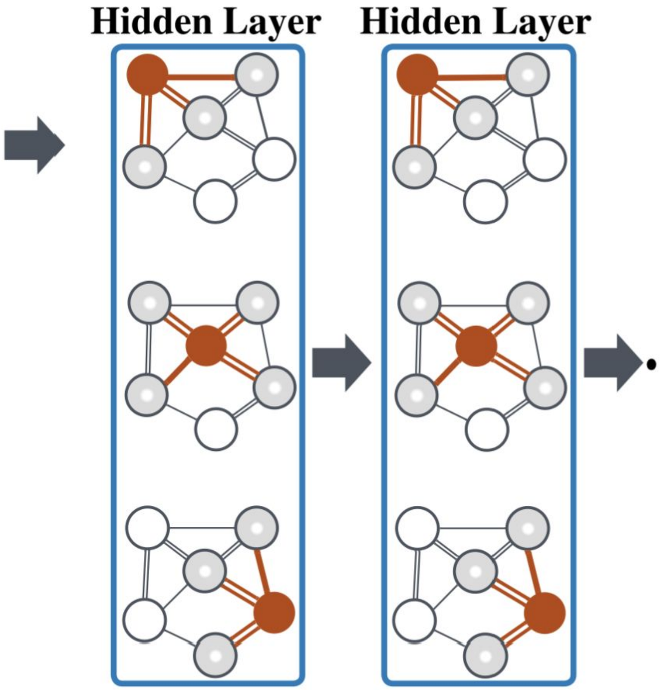
(b) Complemented Graph Convolution

New Conv for Complemented Graph

$\hat{\mathbf{A}}_t \hat{\mathbf{A}}_o$ The **heterophilic neighbors of homophilic neighors** → still heterophilic-prone

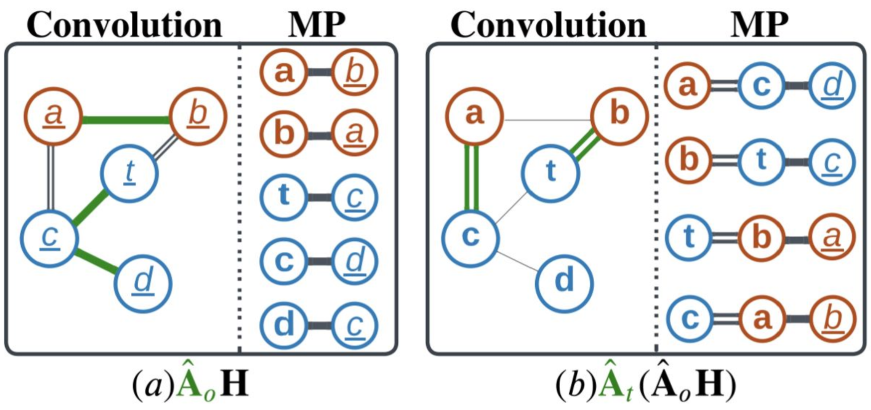


The GOAL Framework



(b) Complemented Graph Convolution

New Conv for Complemented Graph

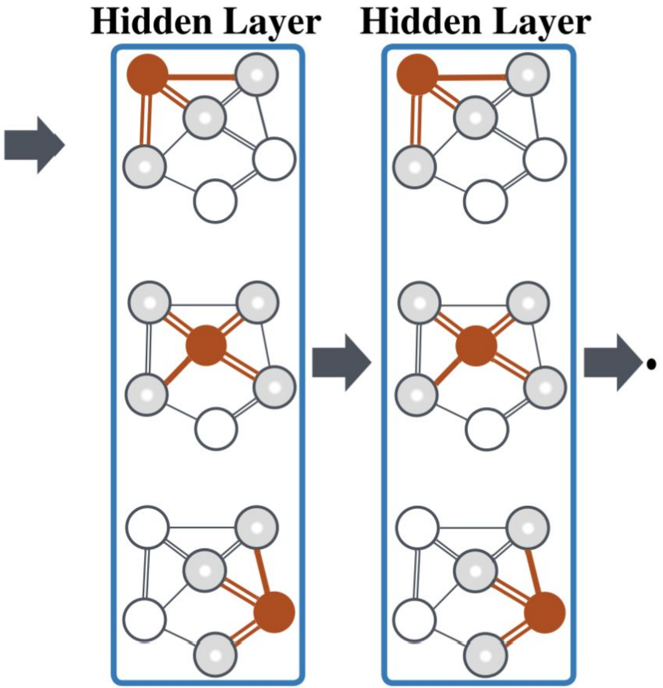


Minimise sim for extended heterophilic neighbors

$$\mathcal{O}_c = \min_{\mathbf{H}}\{tr(\mathbf{H}^T(\mathbf{I}+\hat{\mathbf{A}}_{to})\mathbf{H})\}; \hat{\mathbf{A}}_{to} = \hat{\mathbf{A}}_t\hat{\mathbf{A}}_o$$

$$= \frac{1}{2}\sum_{(i,j)\in\mathcal{E}}^{E} \hat{\mathbf{A}}_{to}[i,j] \parallel \mathbf{H}_i + \tilde{\mathbf{H}}_j \parallel^2 .$$

The GOAL Framework



(b) Complemented Graph Convolution

New Conv for Complemented Graph

Combine all objectives → we finally derive the new convolution

$$\mathbf{H}^{l+1} = \sigma((\alpha \mathbf{I} + \beta \hat{\mathbf{A}}_o - \gamma \hat{\mathbf{A}}_t - \delta \hat{\mathbf{A}}_{to})\mathbf{H}^l \mathbf{W}^l), \quad (8)$$

Experiment

Table 1. Node classification for eight baselines on eight datasets.

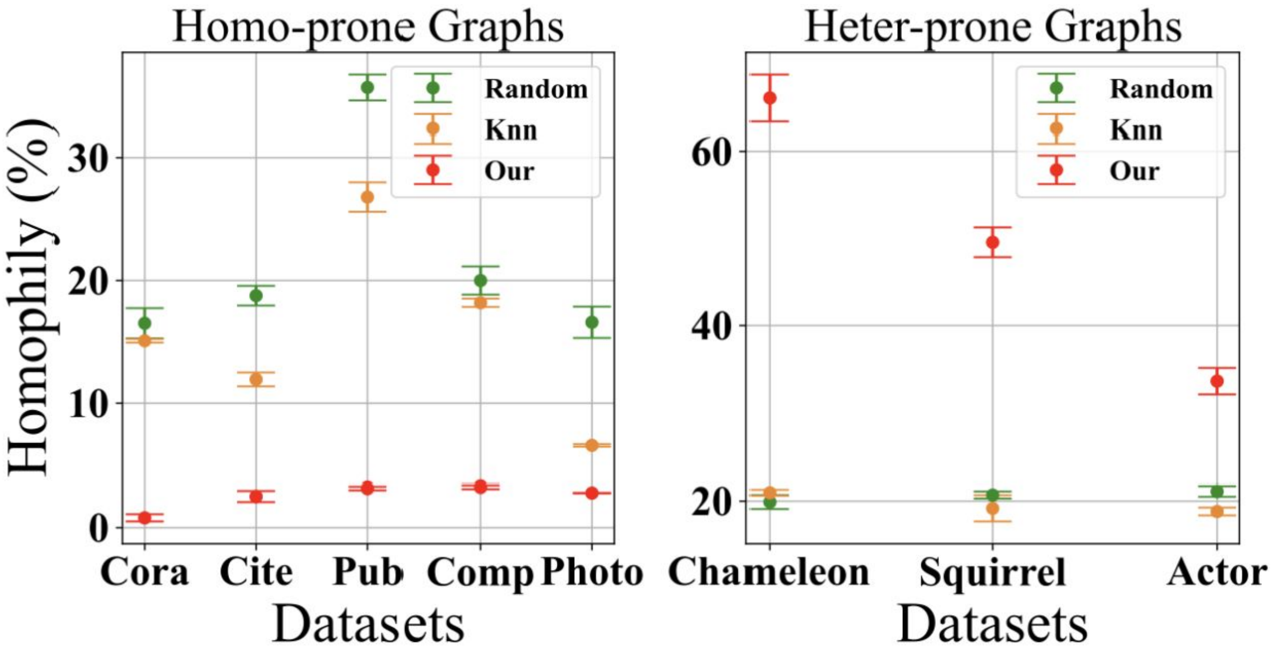
	Cora	Citeseer	Pubmed	Computer	Photo	Chameleon	Squirrel	Actor
MLP	72.09 ± 0.32	71.67 ± 0.40	87.47 ± 0.14	83.59 ± 0.89	90.49 ± 0.20	46.55 ± 0.42	30.67 ± 0.52	28.75 ± 0.88
GCN	87.50 ± 1.04	75.11 ± 1.12	87.20 ± 0.52	83.55 ± 0.38	89.30 ± 0.82	62.72 ± 2.09	47.26 ± 0.34	29.98 ± 1.18
GAT	88.25 ± 1.22	75.75 ± 1.23	85.88 ± 0.38	85.36 ± 0.50	90.81 ± 0.22	62.19 ± 3.78	51.80 ± 1.04	28.17 ± 1.19
APPNP	88.36 ± 0.61	76.03 ± 1.27	86.21 ± 0.25	88.32 ± 0.36	94.44 ± 0.36	50.88 ± 1.18	33.58 ± 1.00	29.82 ± 0.82
GraphSage	88.01 ± 1.29	75.17 ± 1.35	87.39 ± 0.84	88.54 ± 0.69	94.23 ± 0.62	58.82 ± 2.29	41.19 ± 0.75	31.76 ± 0.73
ChebyNet	87.49 ± 0.90	75.50 ± 0.87	89.05 ± 0.29	89.77 ± 0.36	95.02 ± 0.41	59.98 ± 1.54	40.18 ± 0.55	35.85 ± 1.05
GPR-GNN	88.65 ± 0.75	75.70 ± 0.81	88.53 ± 0.30	87.63 ± 0.48	94.60 ± 0.30	67.96 ± 2.55	49.52 ± 5.00	30.78 ± 0.61
JKNET	86.99 ± 1.60	75.38 ± 1.30	88.64 ± 0.51	86.97 ± 0.56	92.68 ± 0.58	64.63 ± 3.08	44.91 ± 1.94	28.48 ± 1.25
GOAL	$\mathbf{88.75 \pm 0.87}$	$\mathbf{77.15 \pm 0.95}$	$\mathbf{89.25 \pm 0.55}$	$\mathbf{91.33 \pm 0.38}$	$\mathbf{95.60 \pm 0.44}$	$\mathbf{71.65 \pm 1.66}$	$\mathbf{60.53 \pm 1.60}$	$\mathbf{36.46 \pm 1.02}$

Achieve best performance on all datasets.

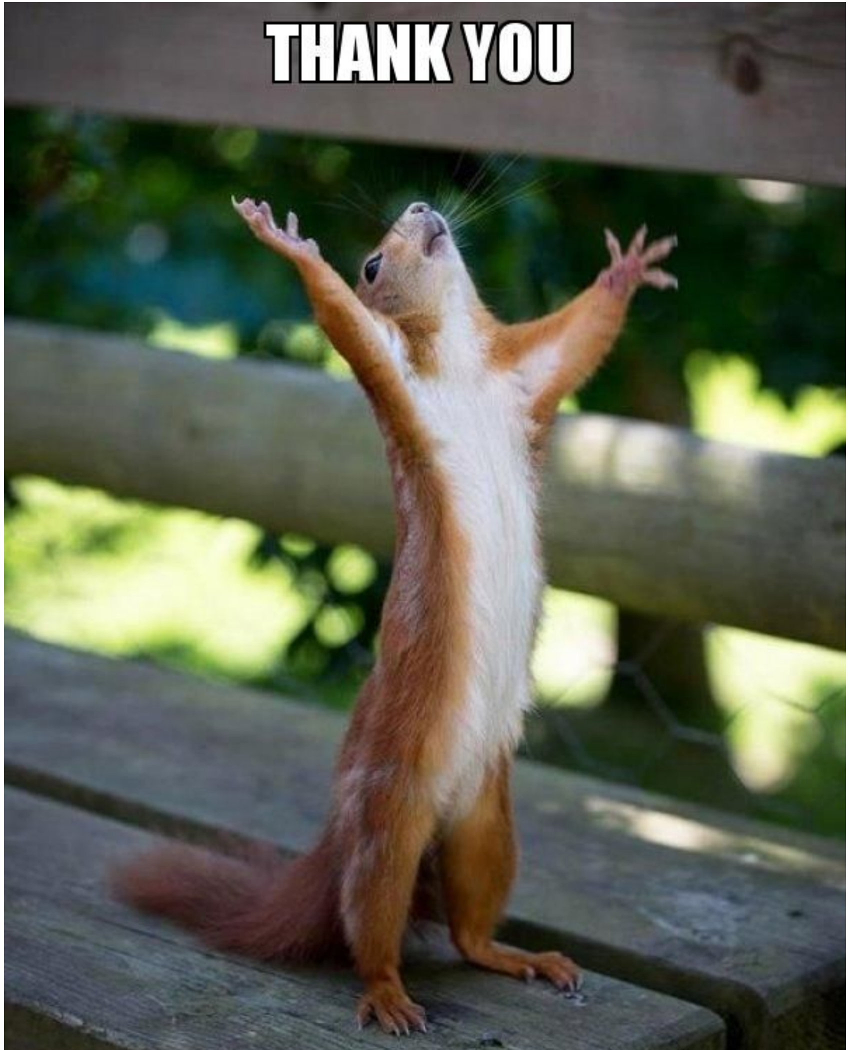
Notably → almost 9% better than the best baseline on Squirrel.

Experiment

ICML
International Conference
On Machine Learning



The generated topology is way better than random and knn connection



ICML
International Conference On Machine Learning