Fair Graph Representation Learning with Imbalanced and Biased Data

Yu Wang

Network and Data Science Lab

Vanderbilt University
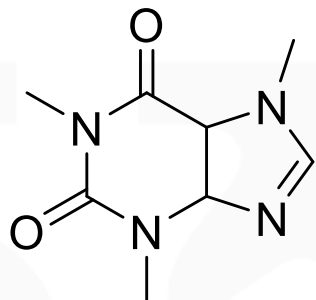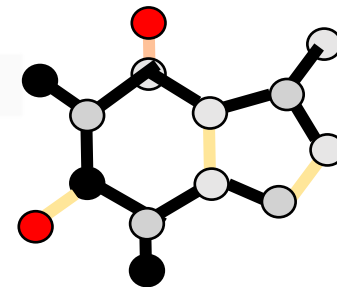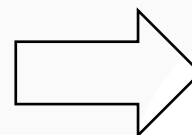
yu.wang.1@Vanderbilt.edu

https://yuwvandy.github.io/
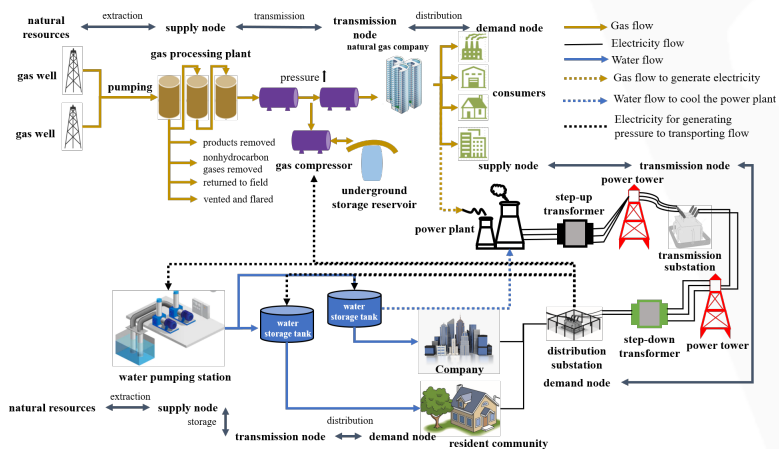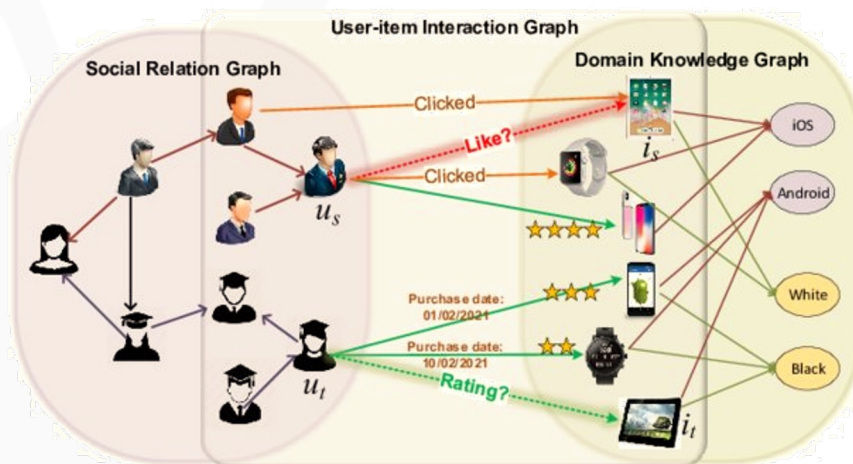
Motivation

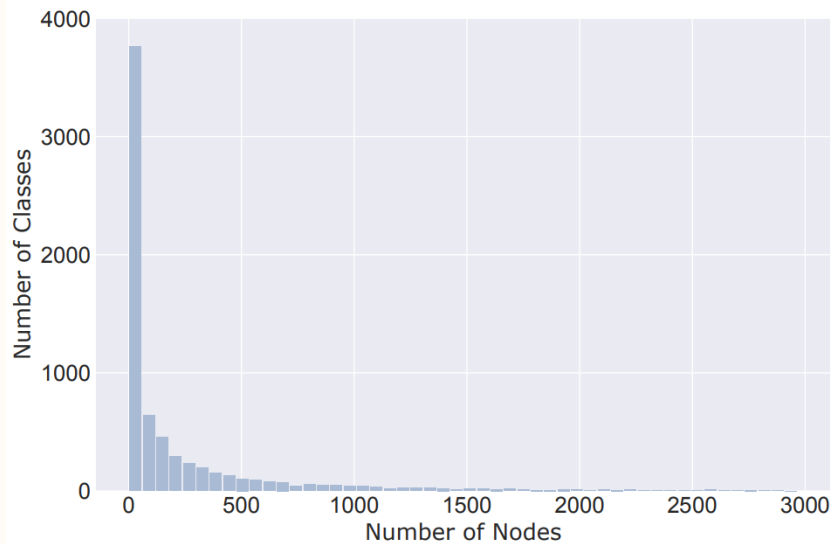Social network

View As

Molecule compounds

Infrastructure network

Online purchasing network

Shouji wang et al.
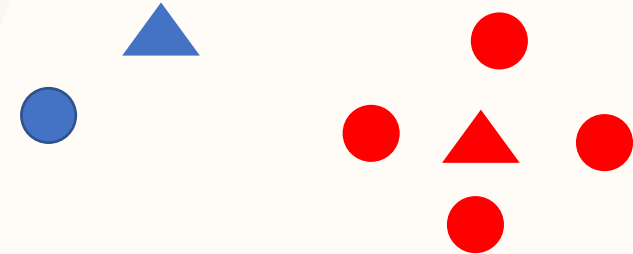
Motivation

Imbalanced Data

Kaize ding 2020

Skewed data distribution in network

Imbalanced training loss

$$\mathcal{L} = \mathcal{L}_1 + \boxed{\mathcal{L}_2}$$
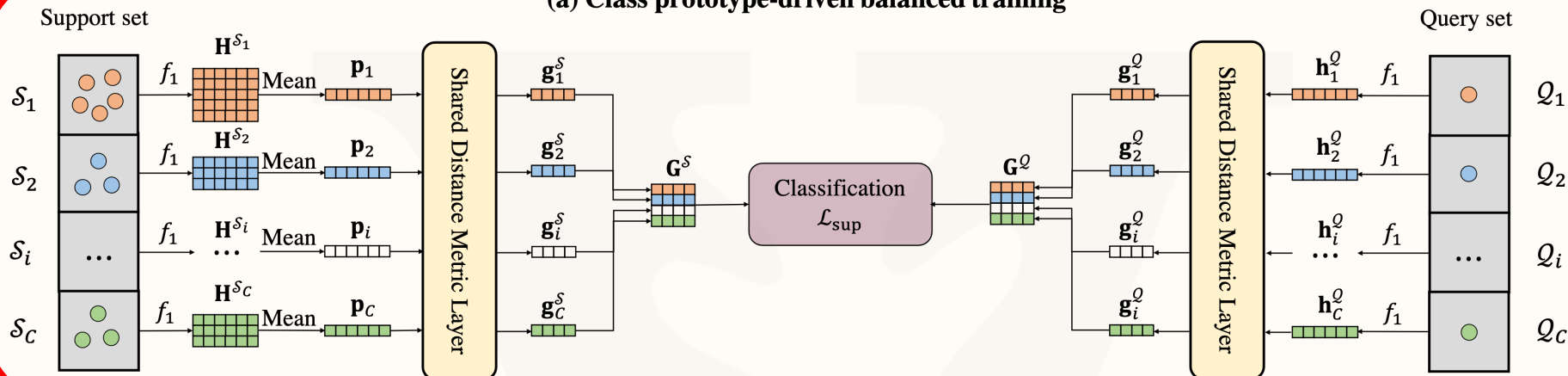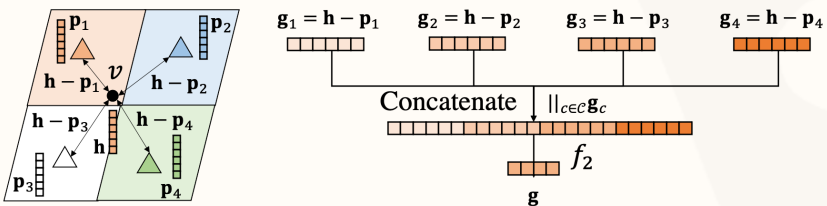
Minority

Poor generalibility

Feature propagation cause more problems!

Imbalanced node classification

(a) Class prototype-driven balanced training

Support set

Query set

(b) Distance Metric Learning

$g_1 = h - p_1 \quad g_2 = h - p_2 \quad g_3 = h - p_3 \quad g_4 = h - p_4$

Concatenate $\|_{c \in \mathcal{C}} g_c$

f_2

g

(c) Imbalanced Label Propagation

If $t = \dfrac{0.35 - 0.23}{0.29} = 0.41 > \eta$?

Assign pseudo label

(d) Self-supervised Learning

$\mathbf{P}$

$\mathcal{L}_{\text{self}_p} = \sum_{i,j} \mathbf{SIM}_{ij} - \text{tr}(\mathbf{SIM})$

$\mathbf{G}$

$\mathcal{L}_{\text{self}_s} = \mathbf{G}^{\mathrm{T}} \mathbf{L} \mathbf{G}$

Sample from the support and query set of each class to balance the training loss
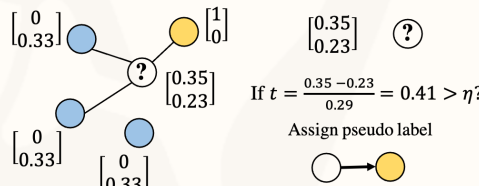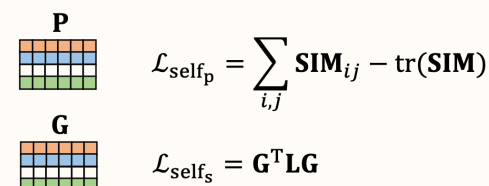
Imbalanced node classification

(a) Class prototype-driven balanced training

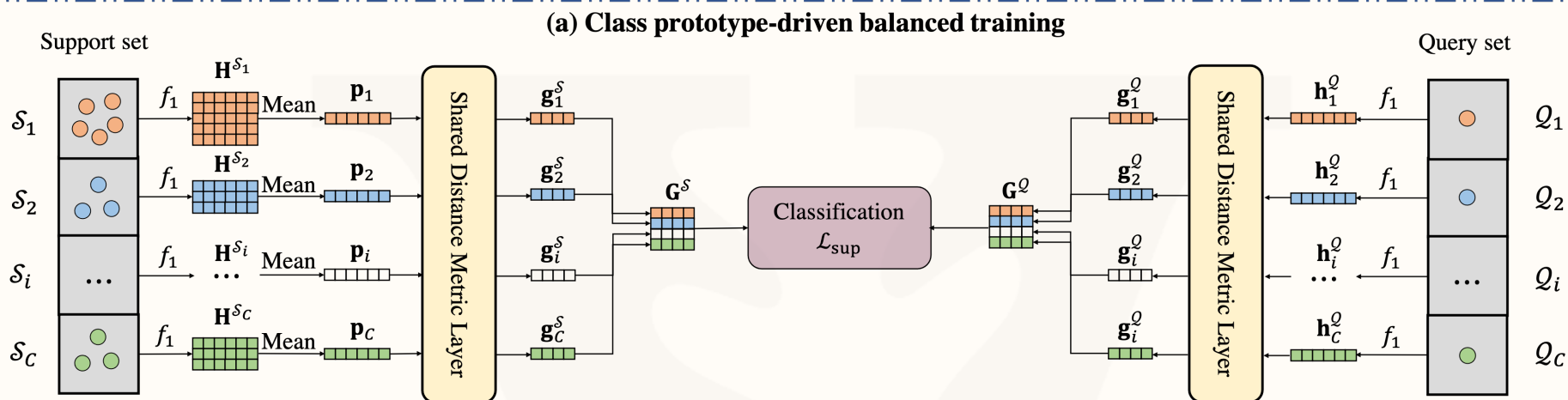(b) Distance Metric Learning

(c) Imbalanced Label Propagation

(d) Self-supervised Learning

Utilize the distance to each prototype to calculate node features

Imbalanced node classification

(a) Class prototype-driven balanced training

(b) Distance Metric Learning

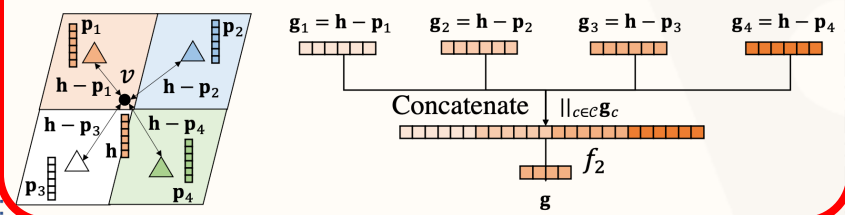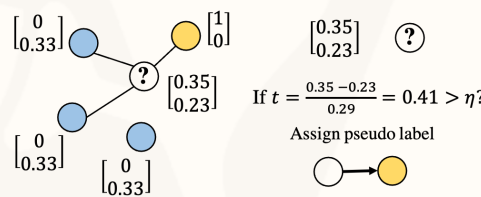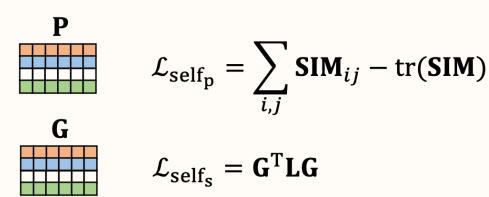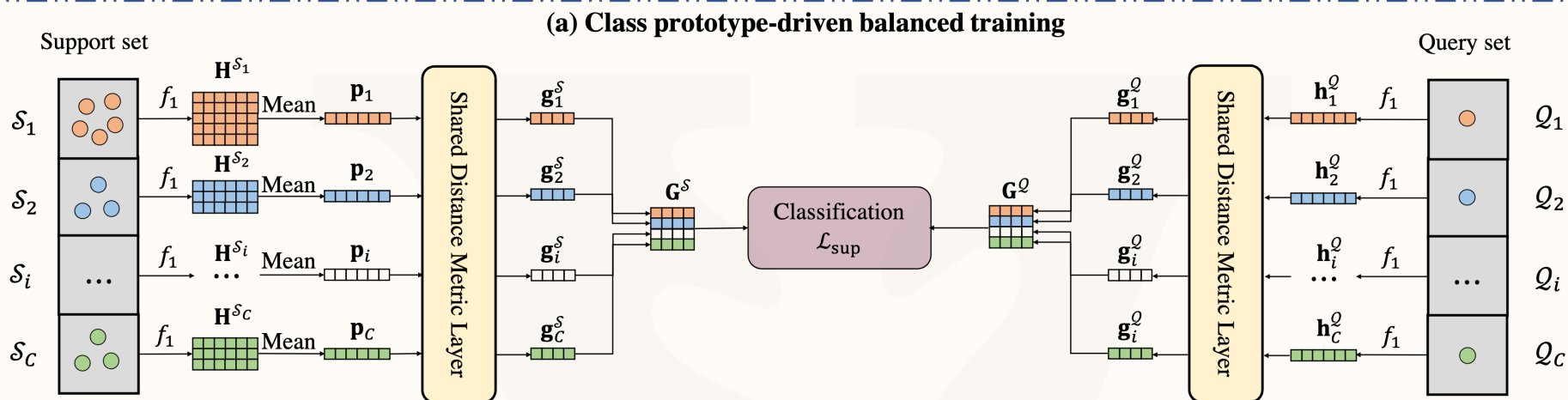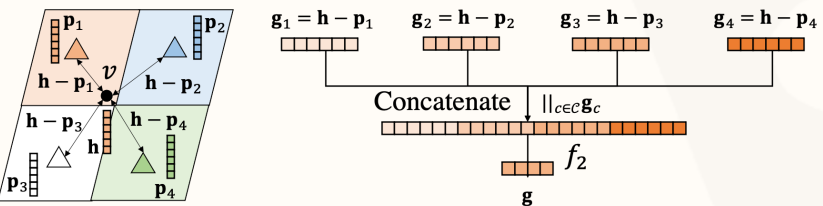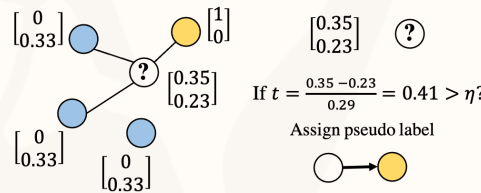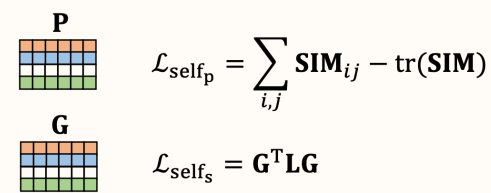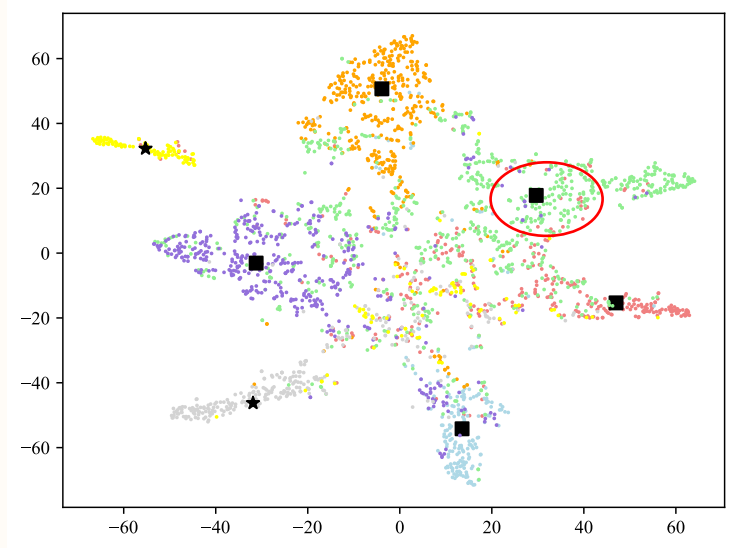(c) Imbalanced Label Propagation

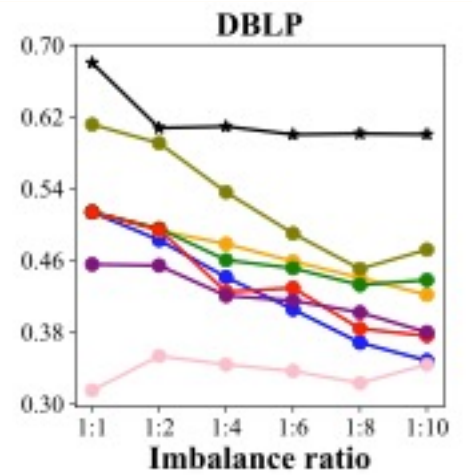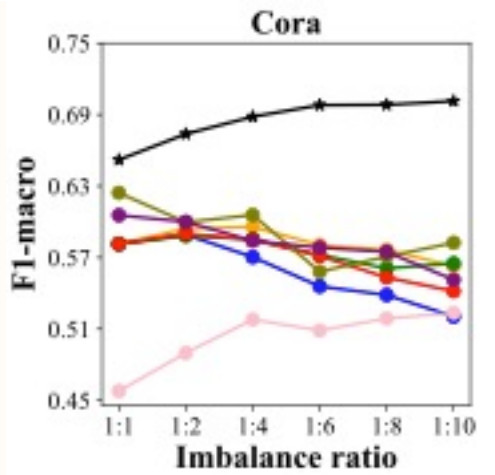(d) Self-supervised Learning

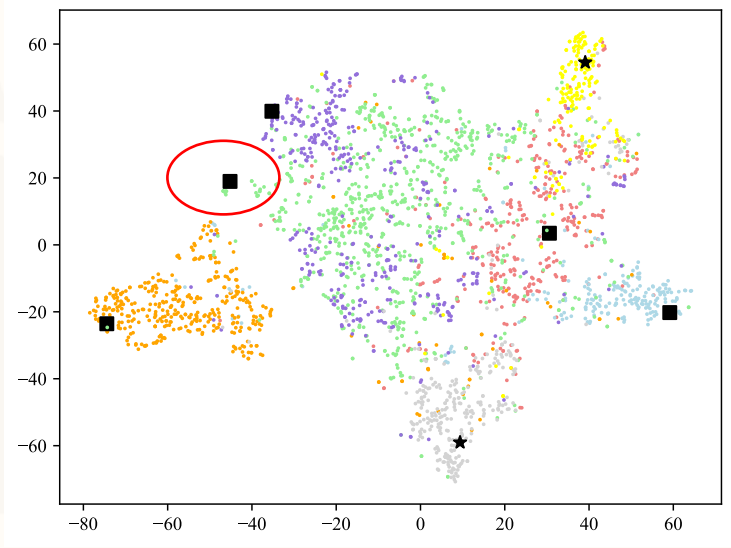Imbalanced label propagation to boost the existing training data

Self-supervised learning to smooth distance metric representation of neighboring nodes and separate prototype representations
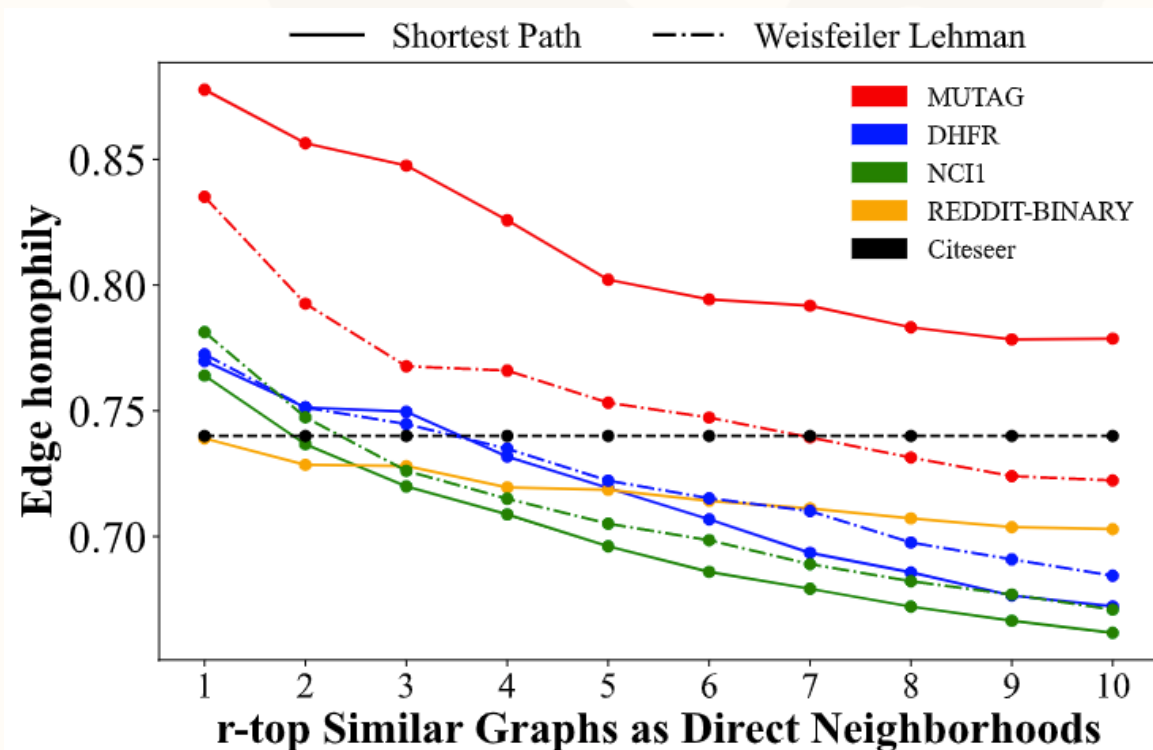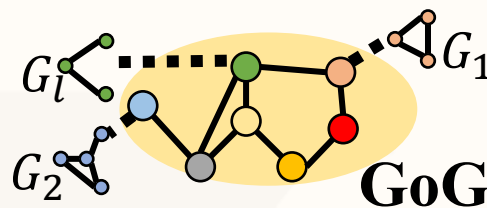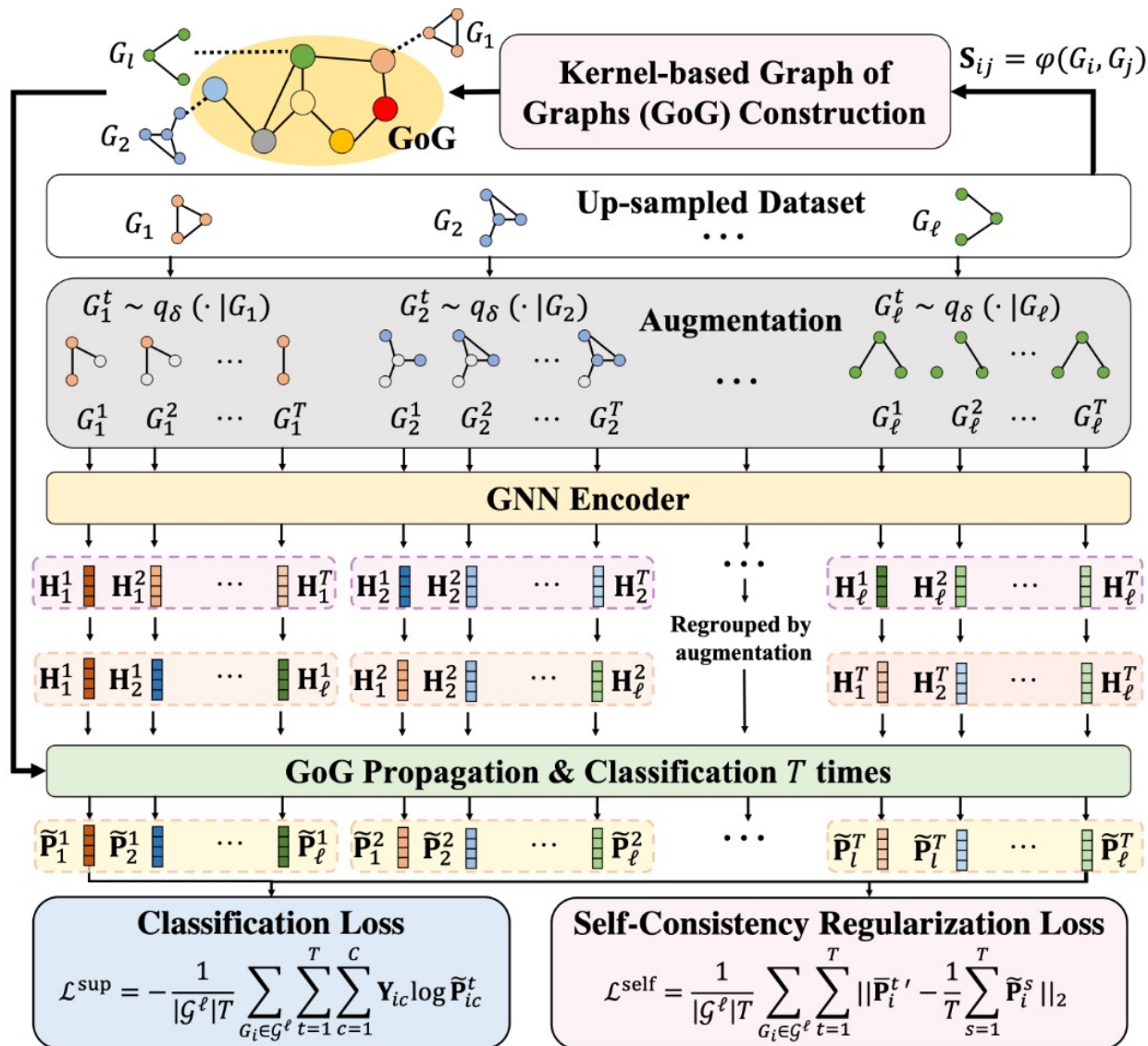
Imbalanced node classification

Our model

GCN

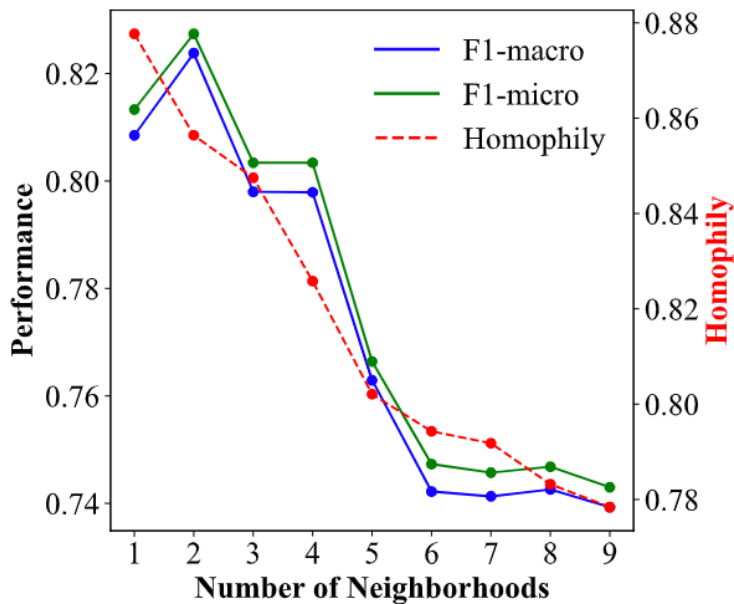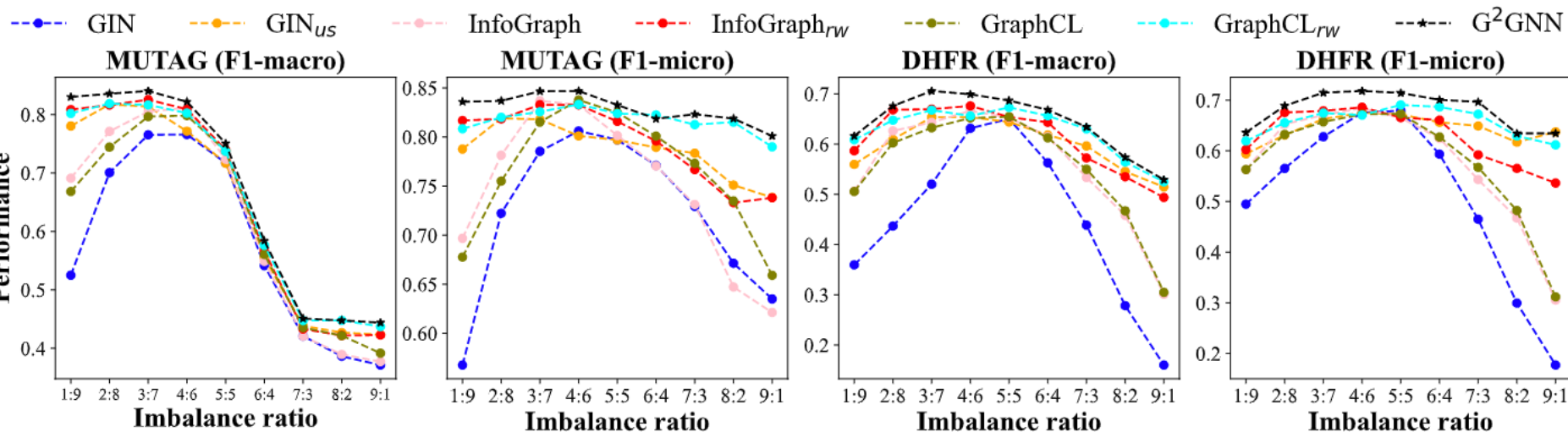# Imbalanced graph classification

Construct a graph of graphs

Imbalanced graph classification

$$\mathbf{S}_{ij} = \varphi(G_i, G_j)$$

Kernel-based Graph of Graphs (GoG) Construction

GoG

Up-sampled Dataset

$$G_1^t \sim q_\delta(\cdot|G_1) \qquad G_2^t \sim q_\delta(\cdot|G_2) \qquad G_\ell^t \sim q_\delta(\cdot|G_\ell)$$

Augmentation

$$G_1^1 \quad G_1^2 \quad \cdots \quad G_1^T \qquad G_2^1 \quad G_2^2 \quad \cdots \quad G_2^T \qquad G_\ell^1 \quad G_\ell^2 \quad \cdots \quad G_\ell^T$$

GNN Encoder

$$\mathbf{H}_1^1 \; \mathbf{H}_1^2 \quad \cdots \quad \mathbf{H}_1^T \qquad \mathbf{H}_2^1 \; \mathbf{H}_2^2 \quad \cdots \quad \mathbf{H}_2^T \qquad \mathbf{H}_\ell^1 \; \mathbf{H}_\ell^2 \quad \cdots \quad \mathbf{H}_\ell^T$$

Regrouped by augmentation

$$\mathbf{H}_1^1 \; \mathbf{H}_2^1 \quad \cdots \quad \mathbf{H}_\ell^1 \qquad \mathbf{H}_1^2 \; \mathbf{H}_2^2 \quad \cdots \quad \mathbf{H}_\ell^2 \qquad \mathbf{H}_1^T \; \mathbf{H}_2^T \quad \cdots \quad \mathbf{H}_\ell^T$$

GoG Propagation & Classification T times

$$\widetilde{\mathbf{P}}_1^1 \; \widetilde{\mathbf{P}}_2^1 \quad \cdots \quad \widetilde{\mathbf{P}}_\ell^1 \qquad \widetilde{\mathbf{P}}_1^2 \; \widetilde{\mathbf{P}}_2^2 \quad \cdots \quad \widetilde{\mathbf{P}}_\ell^2 \qquad \widetilde{\mathbf{P}}_l^T \; \widetilde{\mathbf{P}}_l^T \quad \cdots \quad \widetilde{\mathbf{P}}_\ell^T$$

Classification Loss

$$\mathcal{L}^{\text{sup}} = -\frac{1}{|\mathcal{G}^\ell|T} \sum_{G_i \in \mathcal{G}^\ell} \sum_{t=1}^{T} \sum_{c=1}^{C} \mathbf{Y}_{ic} \log \widetilde{\mathbf{P}}_{ic}^t$$

Self-Consistency Regularization Loss

$$\mathcal{L}^{\text{self}} = \frac{1}{|\mathcal{G}^\ell|T} \sum_{G_i \in \mathcal{G}^\ell} \sum_{t=1}^{T} ||\bar{\mathbf{P}}_i^{t\,\prime} - \frac{1}{T}\sum_{s=1}^{T} \widetilde{\mathbf{P}}_i^s ||_2$$
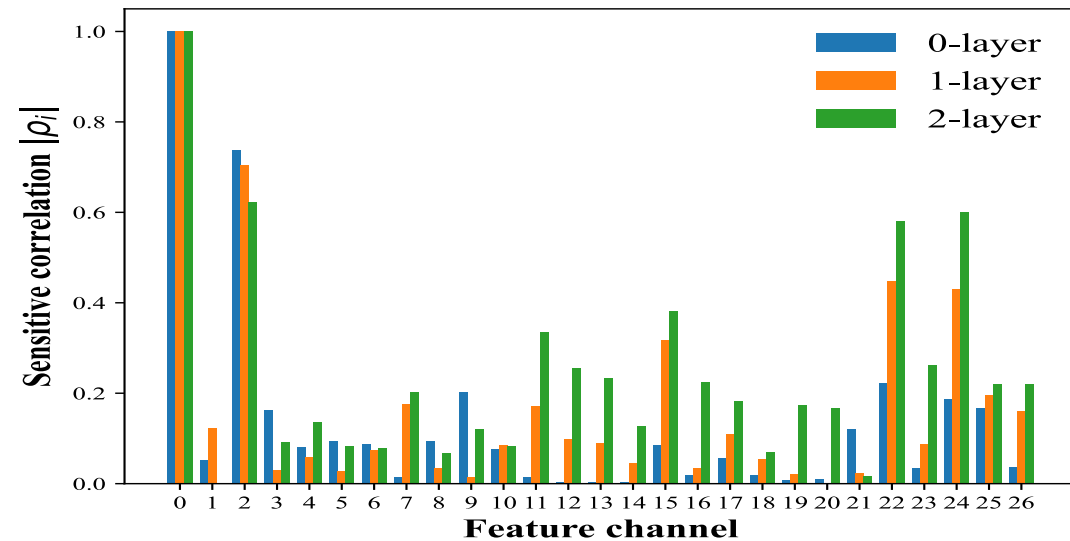
Imbalanced graph classification
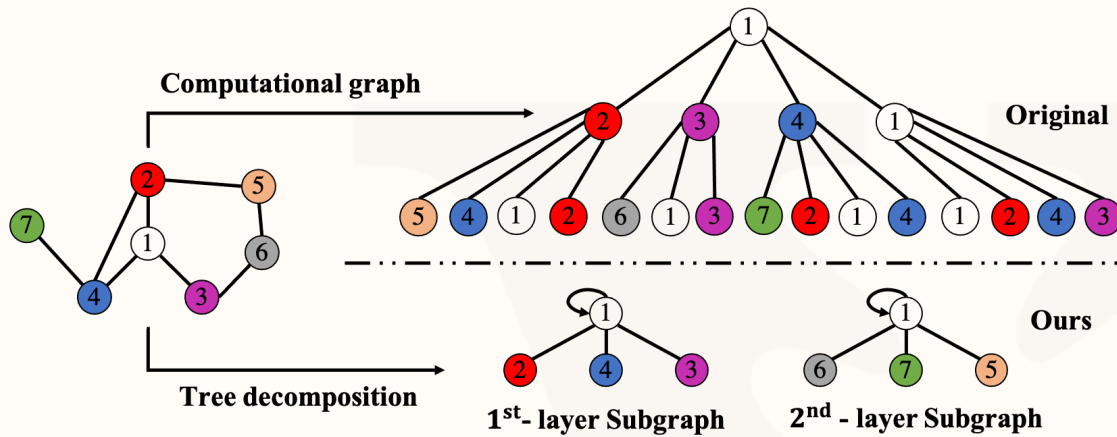
Fair node representation learning
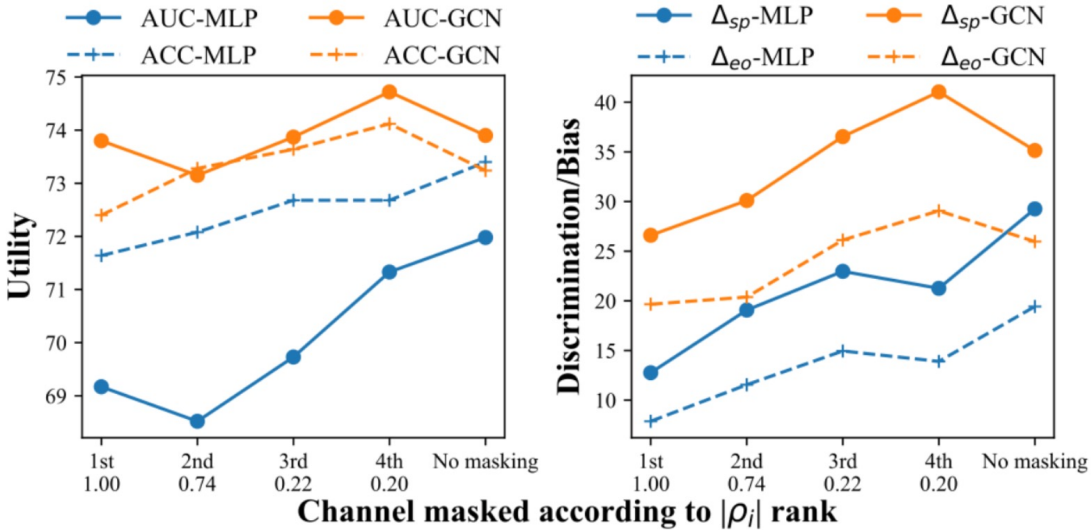
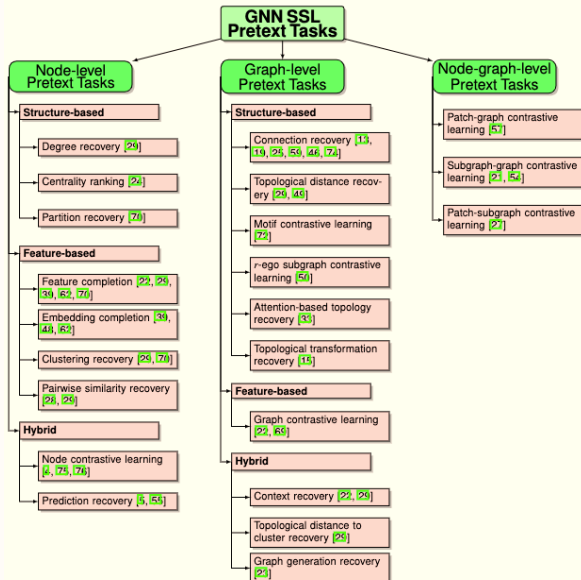Discrimination is correlated to correlations

Feature propagation changes correlations.
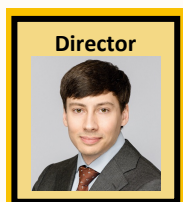
Other works

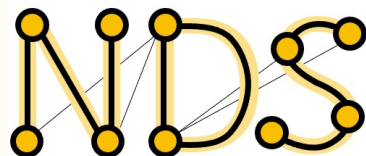Wang, Yu, and Tyler Derr. Tree Decomposed Graph Neural Network *CIKM 2021*

Wang, Yu, Wei Jin, and Tyler Derr. "Graph Neural Networks: Self-supervised Learning." Graph Neural Networks: Foundations, Frontiers, and Applications. Springer, Singapore, 2022. 391-420.
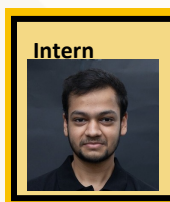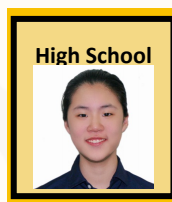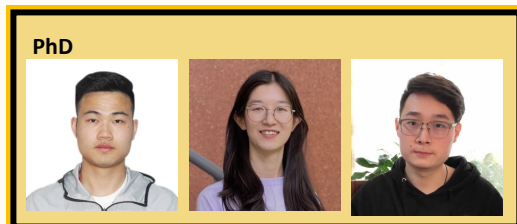
Acknowledgement

Director

Intern

High School

Network and Data Science Lab

PhD
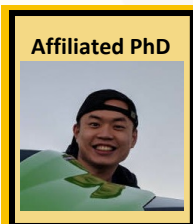
Affiliated PhD

MS
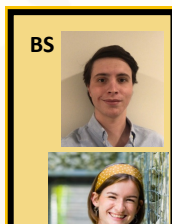
BS
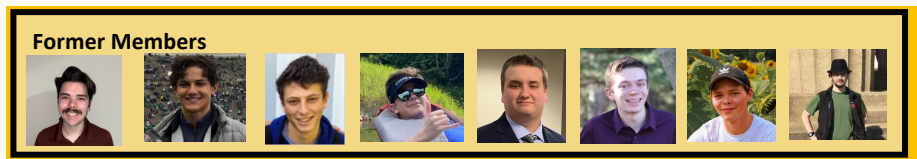
Former Members

More about me!

https://yuwvandy.github.io/

https://nds-vu.github.io/