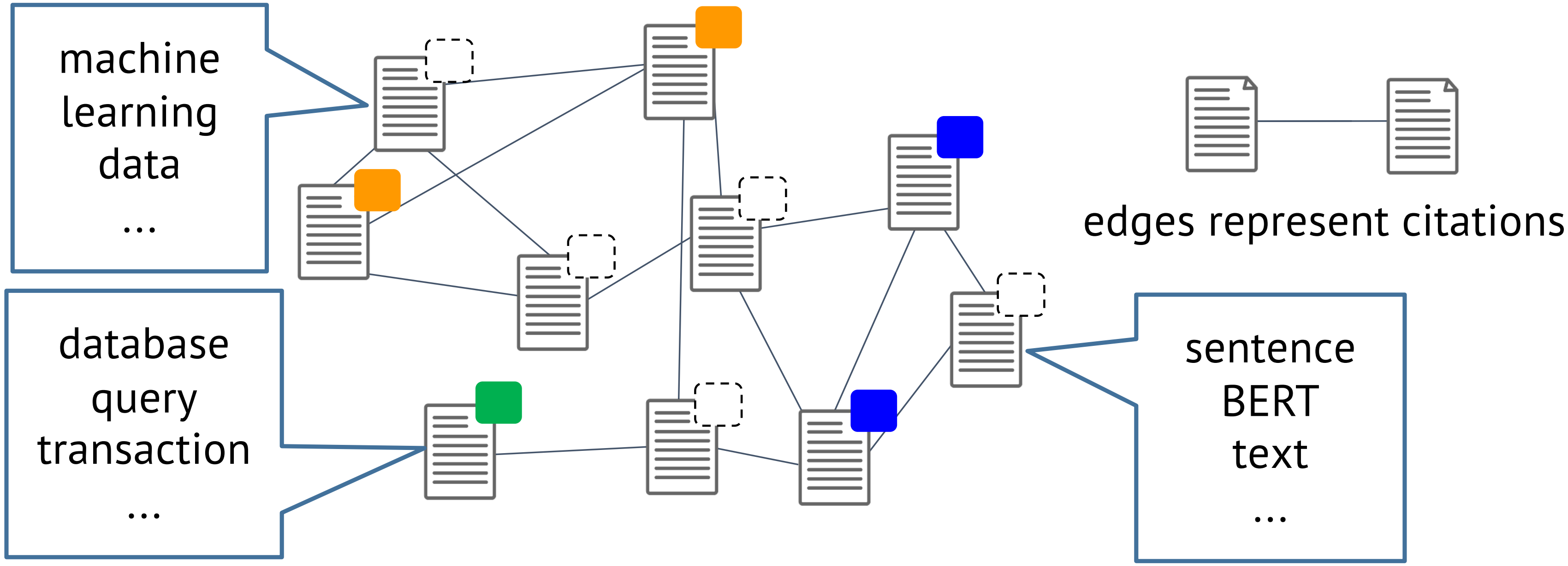
36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

Beyond Real-world Benchmark Datasets: An Empirical Study of Node Classification with GNNs

Seiji Maekawa[1], Koki Noda[2], Yuya Sasaki[1], Makoto Onizuka[1]
[1]Osaka University, [2]TDAI Lab

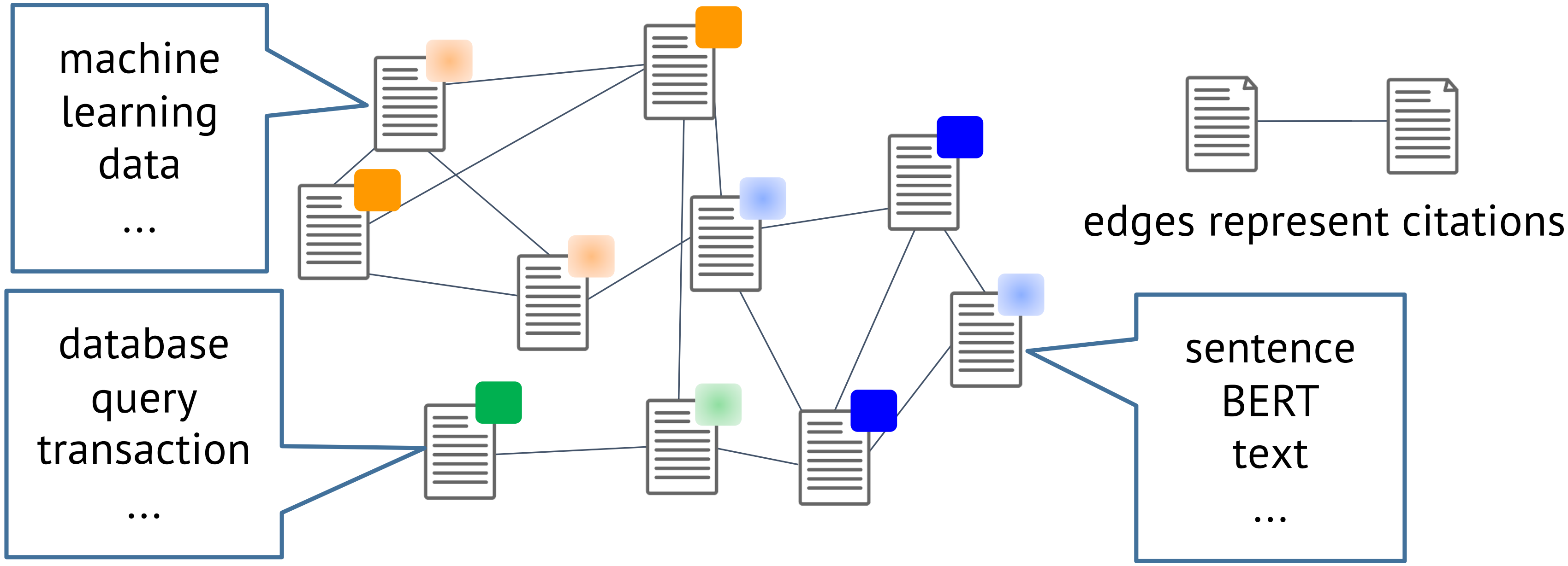
[code] [paper]

Node Classification



Given a partially labeled network,

Node Classification



Given a partially labeled network, predict the labels of the rest of nodes.

Graph Neural Networks (GNNs) are powerful tools for node classification.

Limitations of Existing Evaluation of GNNs

Towards practical use cases of GNNs, researchers and developers need to deeply understand the strengths and weaknesses of GNNs from **various aspects**.

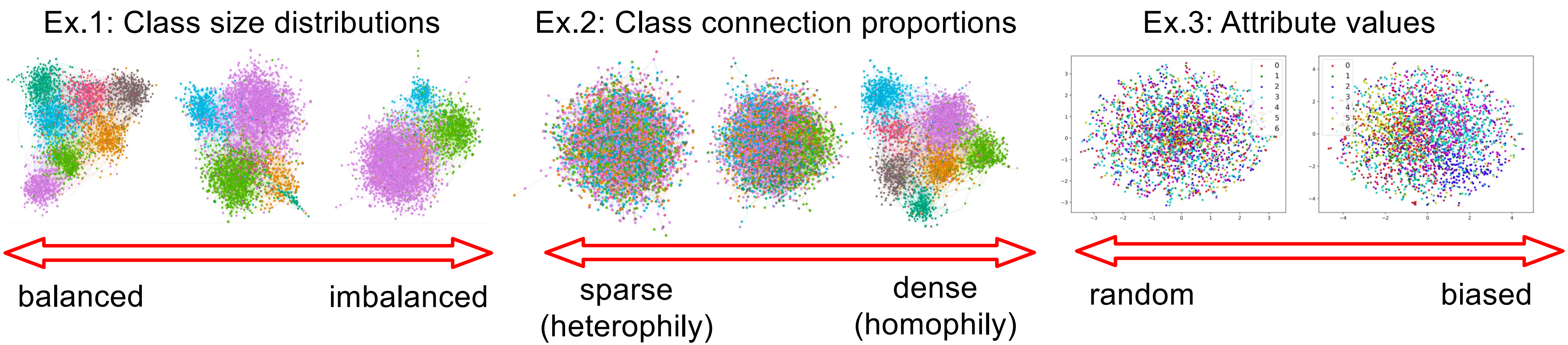
However, to assess GNNs, most existing works use well-known but **limited benchmark datasets**, such as Cora, Citeseer, and PubMed [1] (relatively balanced classes, strong homophily, and small-scale).

It is important to conduct extensive experiments using various graphs with **different characteristics**.

[1] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. "Revisiting semi-supervised learning with graph embeddings". In ICML. PMLR, 2016.

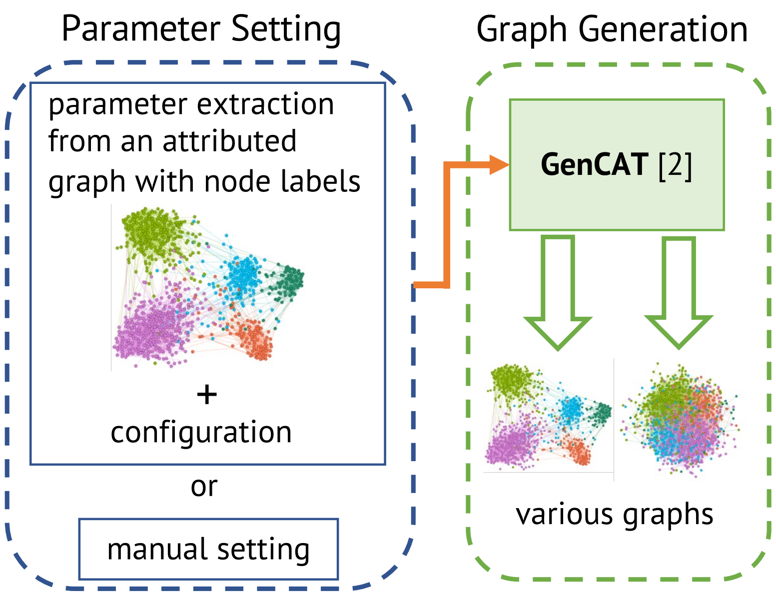
Approach to Comprehensive Evaluations of GNNs

We empirically study the performance of GNNs with various graphs by synthetically **changing one or a few target characteristic(s)** of graphs.



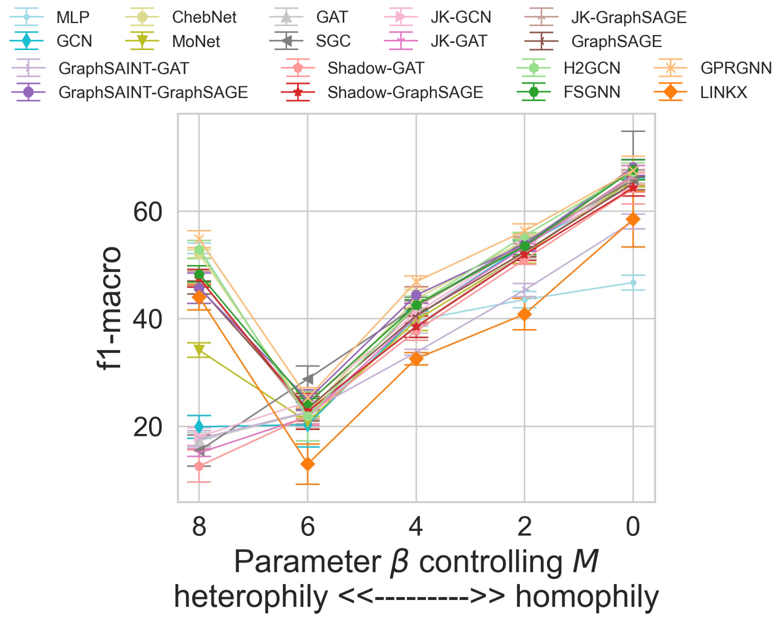
Approach to Comprehensive Evaluations of GNNs

We empirically study the performance of GNNs with various graphs by synthetically **changing one or a few target characteristic(s)** of graphs.



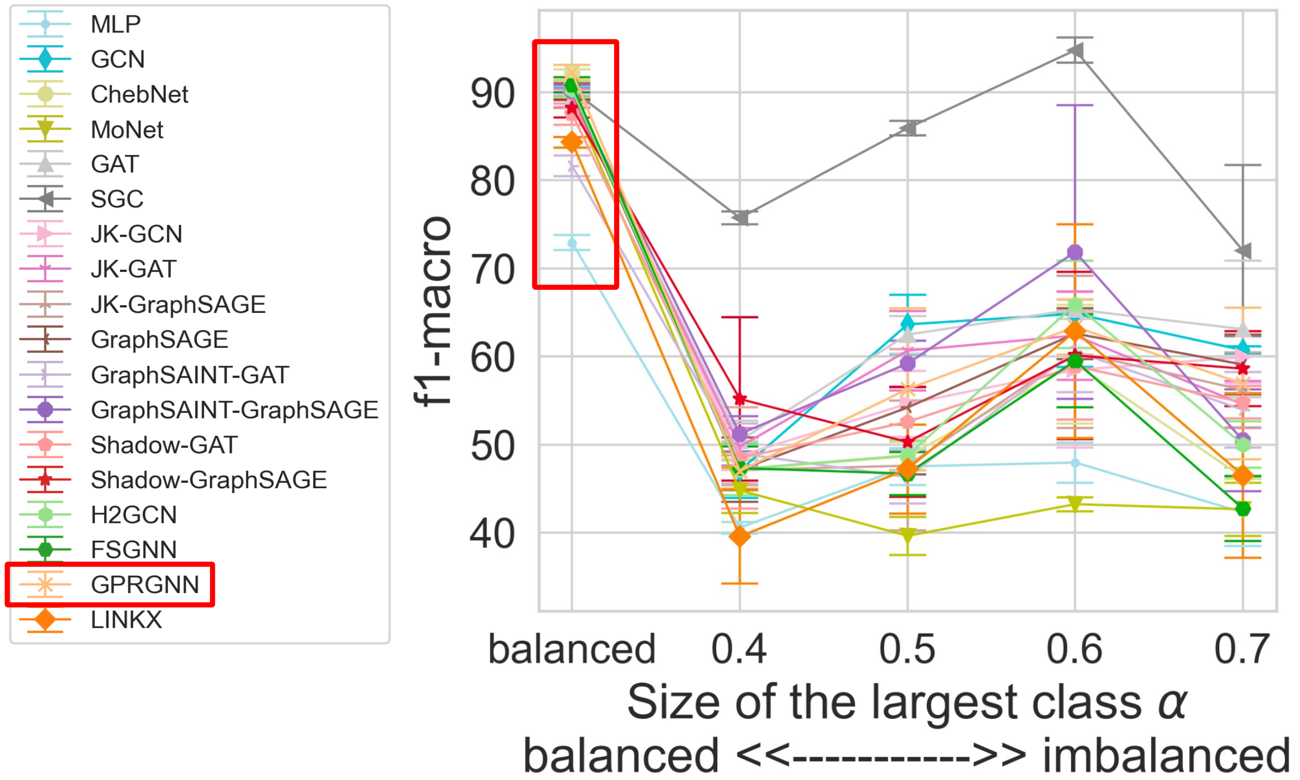
Ex. The parameters are extracted from ***Cora*** and then users can configure the parameters.

Evaluating graph neural network models with generated graphs having various characteristics.



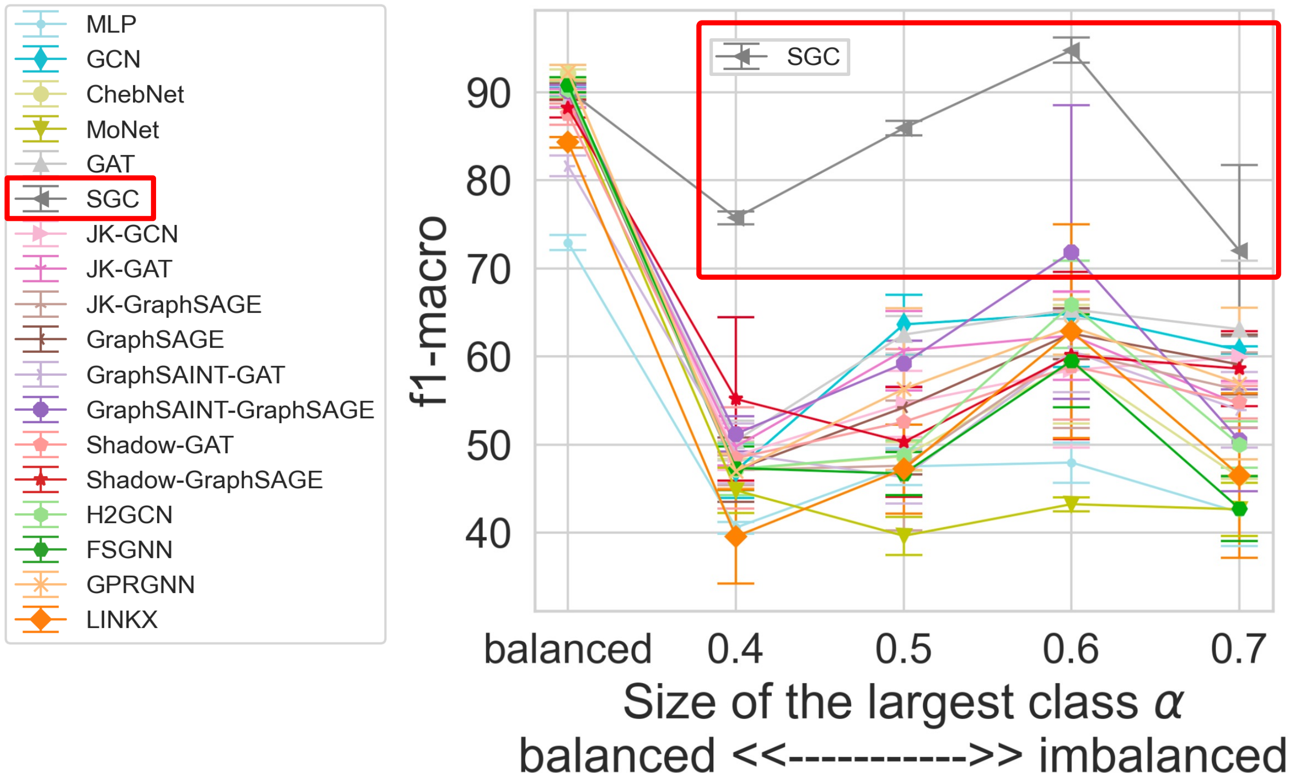
[2] Seiji Maekawa, Yuya Sasaki, George Fletcher, and Makoto Onizuka. GenCAT: Generating Attributed Graphs with Controlled Relationships between Classes, Attributes, and Topology. arXiv preprint, 2021.

Empirical Study 1: Class Size Distributions



The very recent method GPRGNN achieves the best score in the balanced setting.

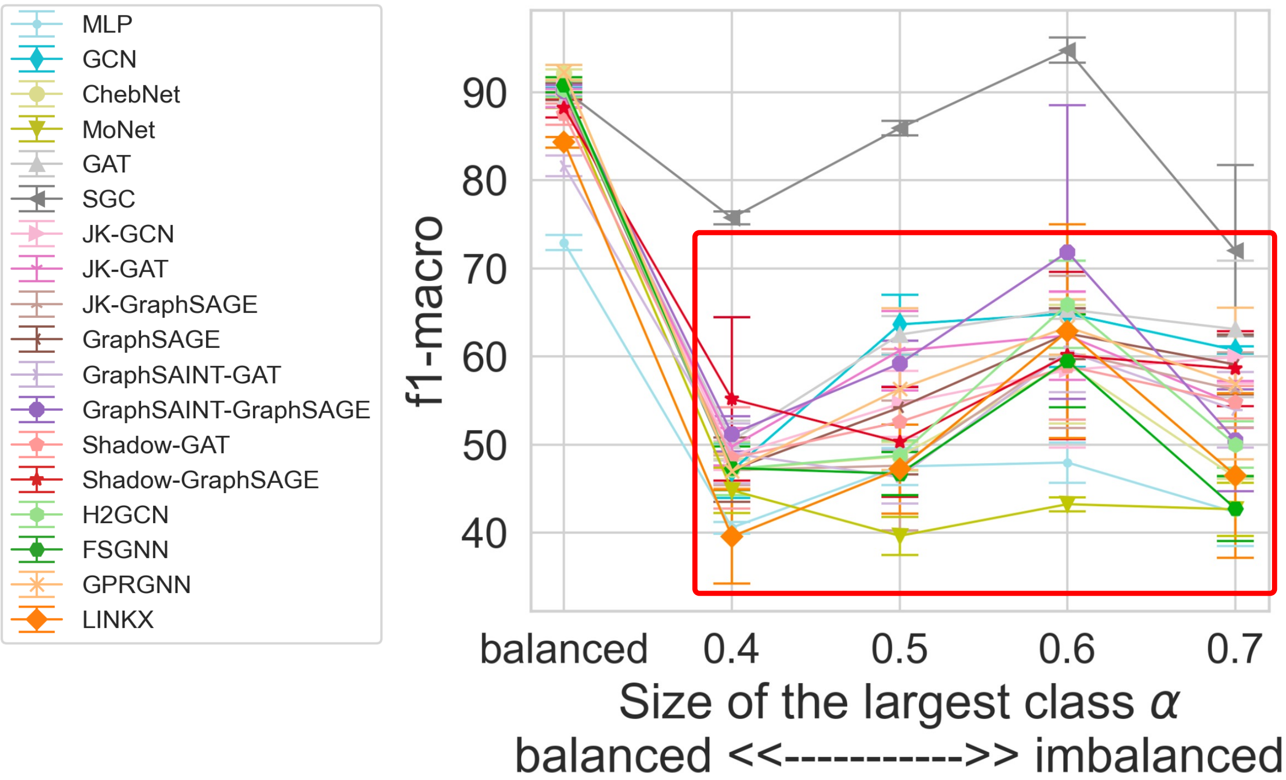
Empirical Study 1: Class Size Distributions



The very recent method GPRGNN achieves the best score in the balanced setting.

Interestingly, a linear model (SGC) achieves the best scores in the imbalanced settings.

Empirical Study 1: Class Size Distributions

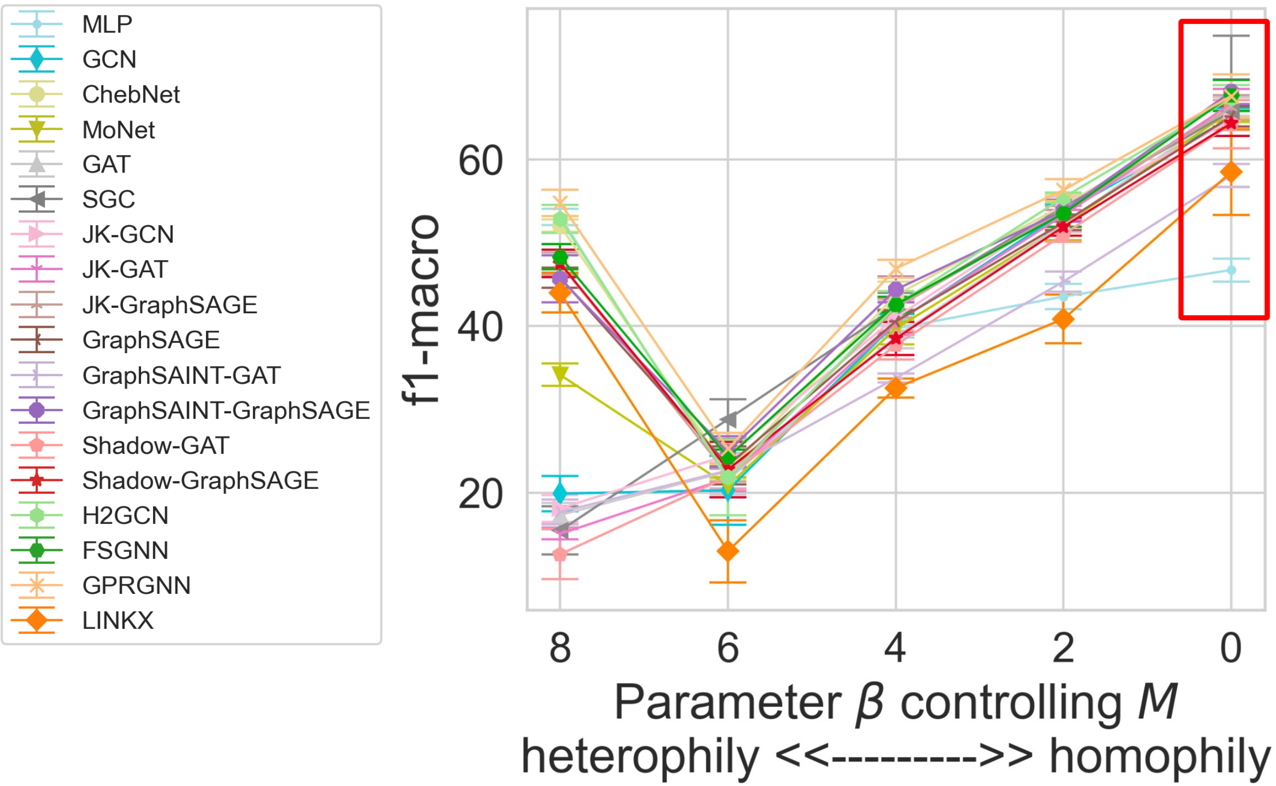


The very recent method GPRGNN achieves the best score in the balanced setting.

Interestingly, a linear model (SGC) achieves the best scores in the imbalanced settings.

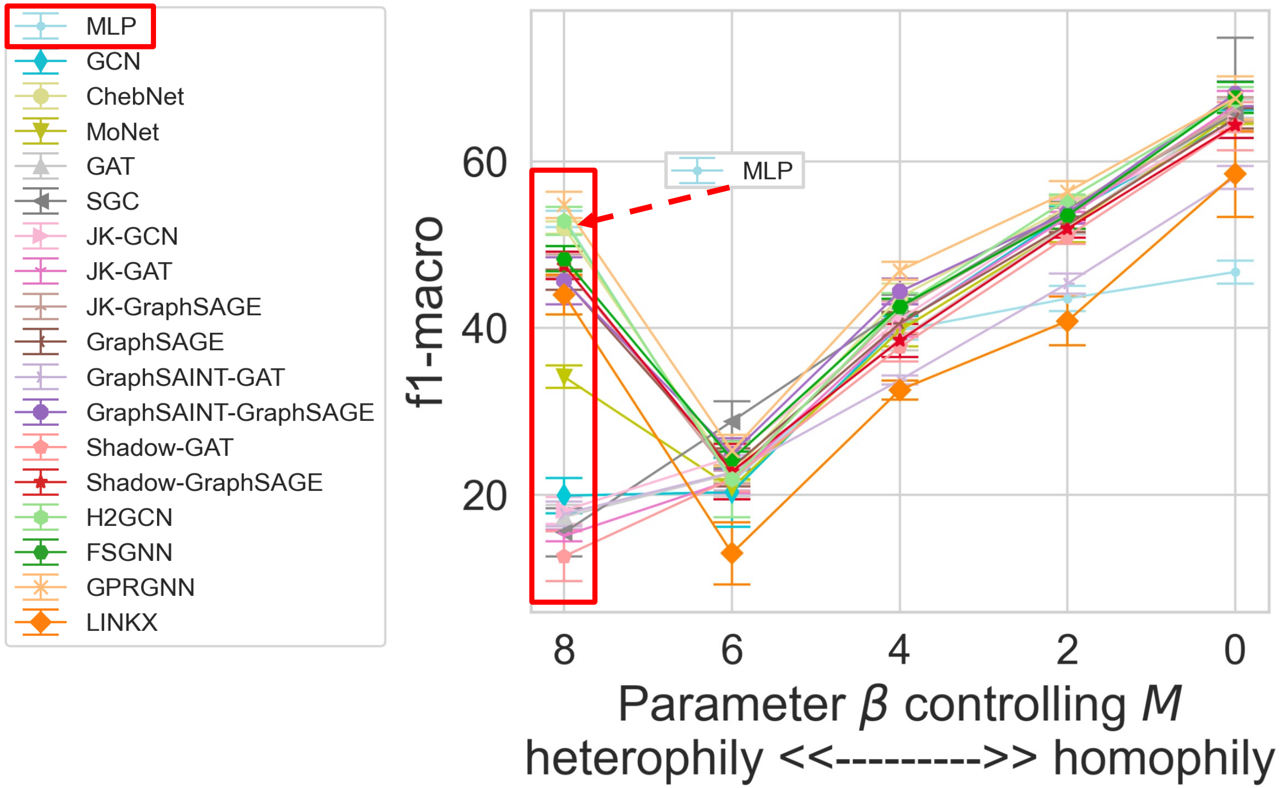
Other complicated GNNs tend to overfit major classes.

Empirical Study 2: Class Connection Proportions



All models work well in the homophily setting.

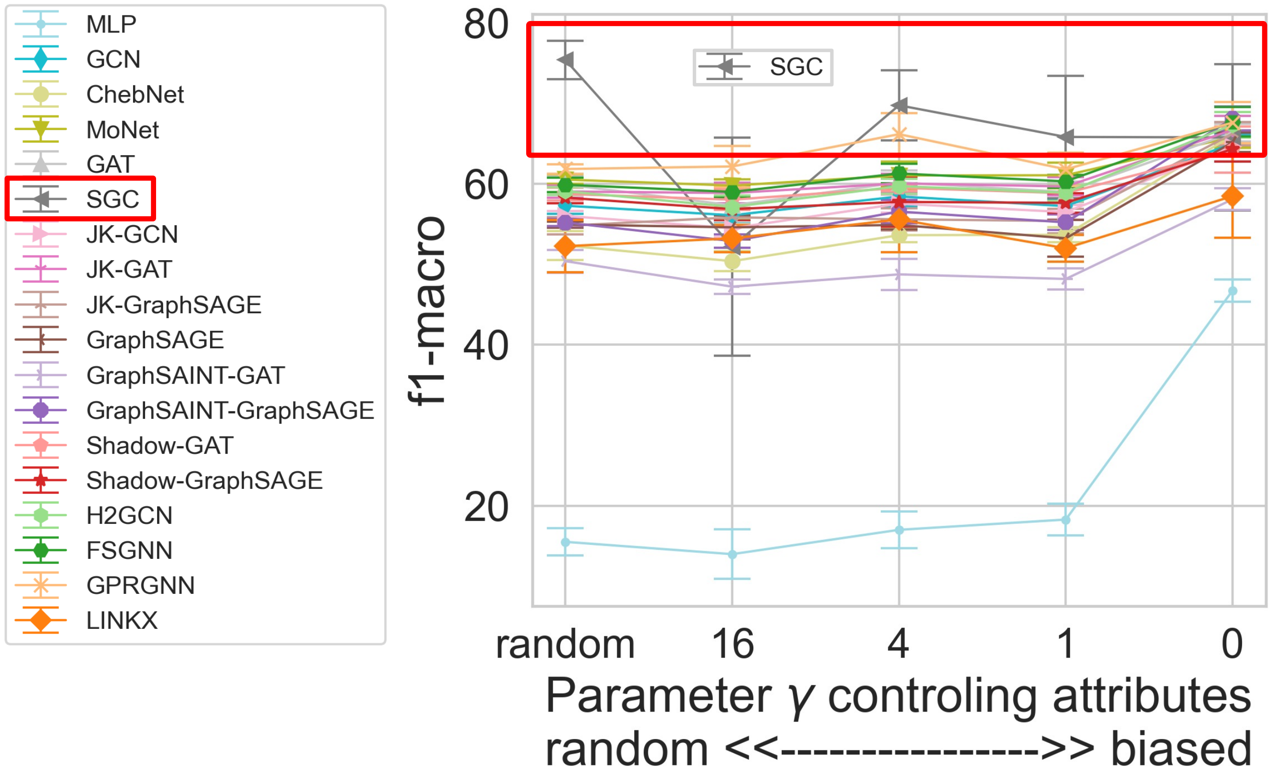
Empirical Study 2: Class Connection Proportions



All models work well in the homophily setting.

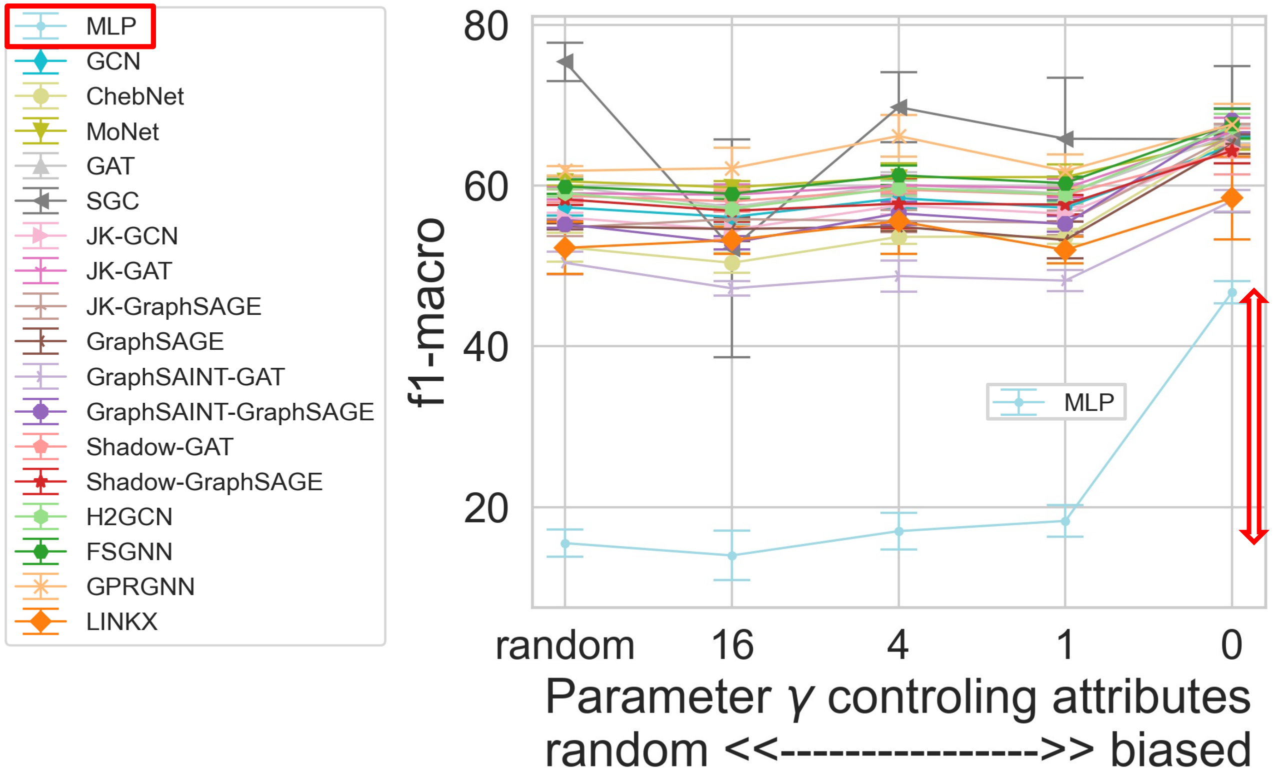
In the heterophily setting, a graph-agonistic model (MLP) achieves comparable results to SOTA GNNs.

Empirical Study 3: Attribute Values



SGC works well since it does not over-fit large classes.

Empirical Study 3: Attribute Values



SGC works well since it does not over-fit large classes.

MLP achieves a large performance gain, which indicates some **overlap between** the contributions of the **topology and attributes**.

[code] [paper]

Contributions

We conducted empirical studies of GNNs for node classification and clarified the limitations and opportunities of the current GNNs.

Open Questions;

- *Class imbalance.*
  How can we develop GNNs that work well in complicated settings such as the combinations of class imbalance, heterophily property, and large-scale graphs?
- *Heterophily setting.*
  How can we develop GNNs that can capture the class structure from heterophilic graphs while achieving the state-of-the-art performance on homophilic graphs?

We hope this work provides interesting insights for future research.

Thank you for your attention!