CS461 Project Midterm Review

Oversampling in Heterogeneous Graphs using SMOTE

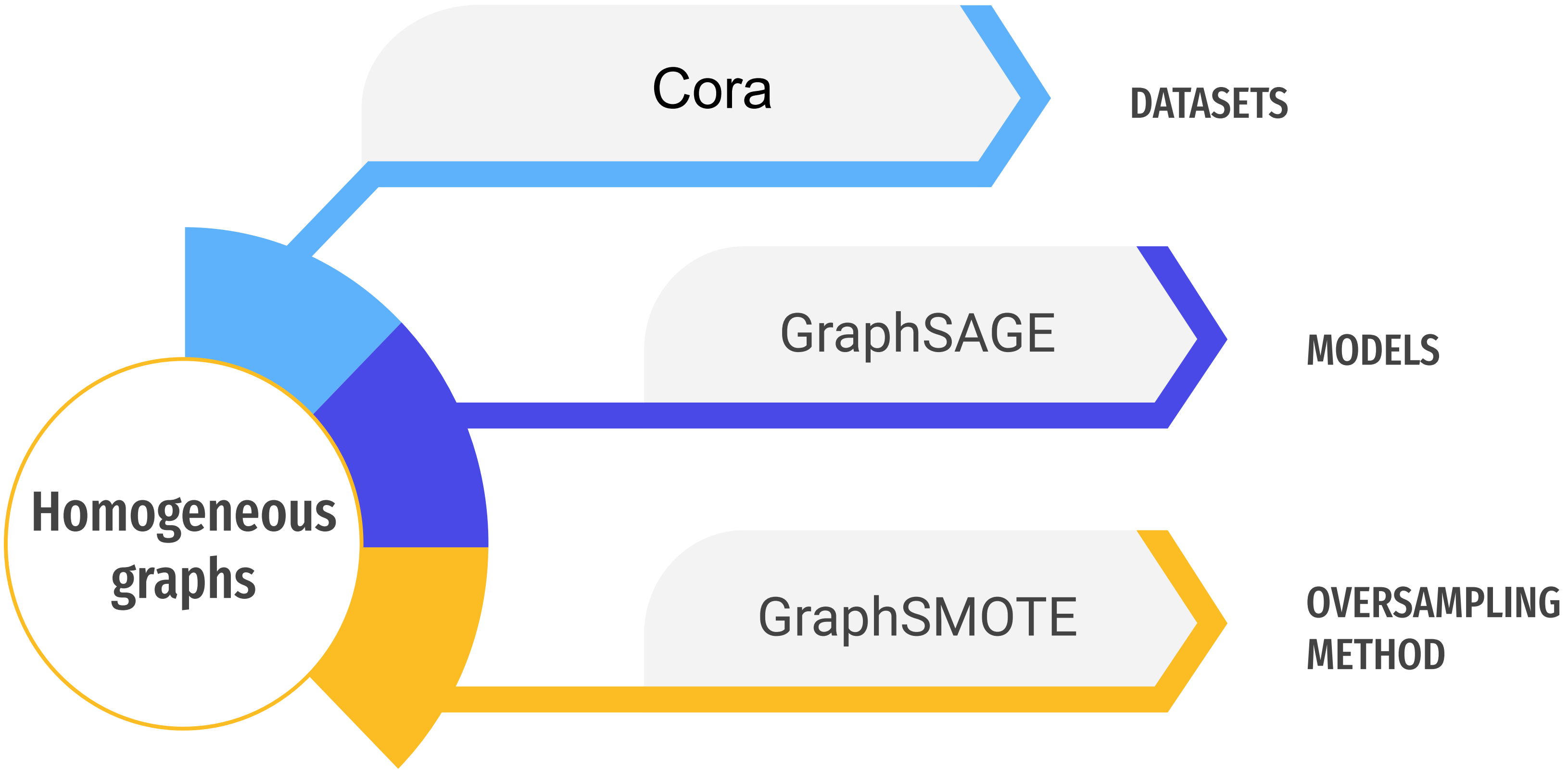
By
Adhilsha A and **Deependra Singh**

AIM & IMPORTANCE

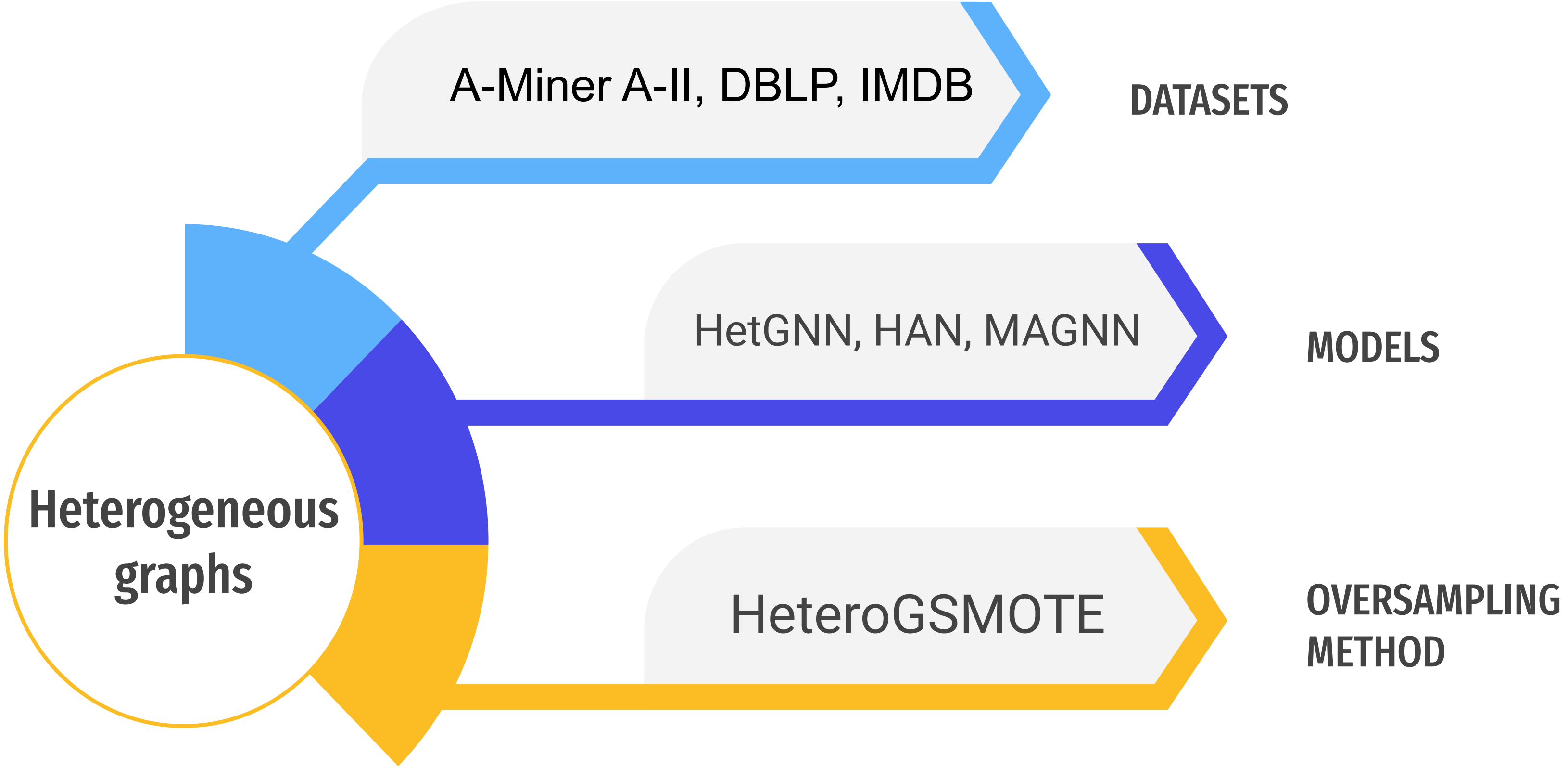
- **AIM**
  - To perform Oversampling on Heterogenous graphs with class imbalance using SMOTE technique to improve downstream tasks.
  - Experimenting the technique on both multiple class imbalances and metapaths to gain more insight.

- **IMPORTANCE**
  - Heterogeneous graphs represent a wide range of real-life data.
  - Heterogeneous graphs with imbalanced class distributions are challenging problem cause of the obvious lack of data, affecting the model's overall performance as well as the performance on the minority class.
  - Current SMOTE technique was applied on homogeneous graphs and hence did not address different node and edge types.

Homogeneous Graphs



Heterogeneous Graphs datasets and models



Homogeneous graphs & Heterogeneous graphs

DATASET

Cora

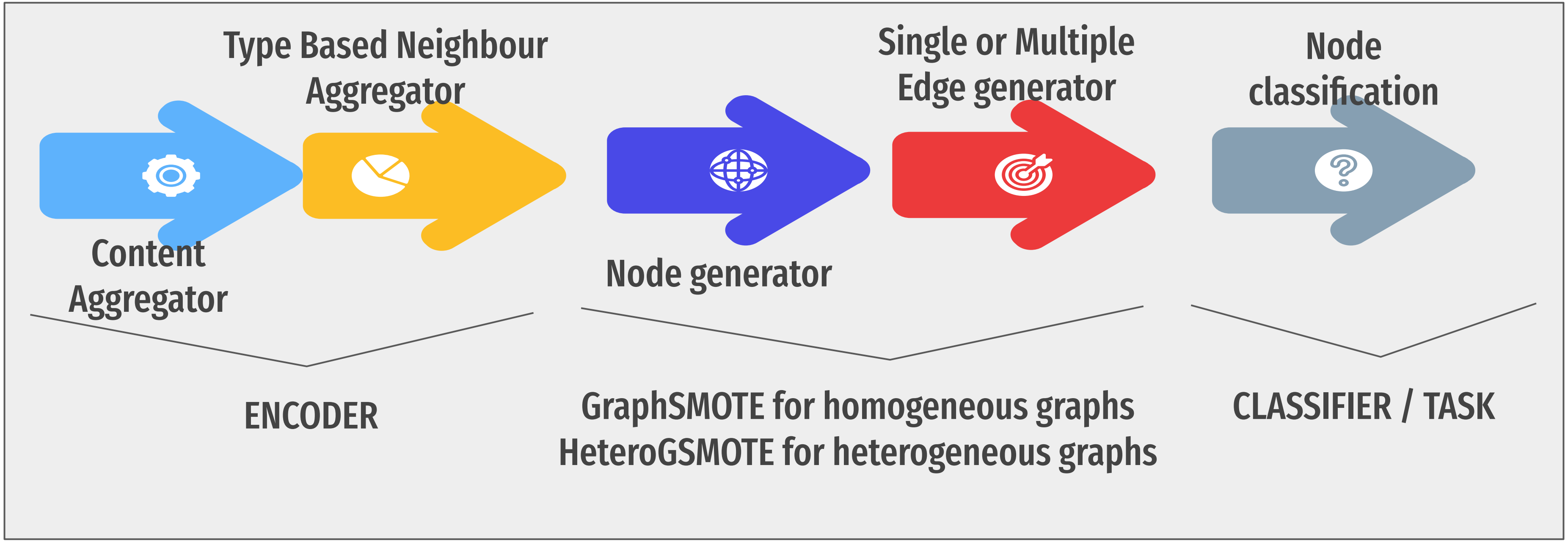
- 2708 paper nodes
- 1433 dimensional Attributes
- 7 classes
- 5429 paper-paper citation edges

DATASET

A-Miner A-II

- 28645 Author nodes
- 21044 Paper nodes
- 18 Venue nodes
- 69311 A-P edges
- 46391 P-P edges
- 21044 P-V edges
- Abstract and title embeddings of papers
- 4 classes

Experiment Implementation



$$\alpha^{v,i} = \frac{exp\left\{LeakyReLU(u^T[f_i \bigoplus f_1(v)])\right\}}{\sum_{f_j \in \mathcal{F}(v)} exp\left\{LeakyReLU(u^T[f_j \bigoplus f_1(v)])\right\}}$$

Results for GraphSMOTE

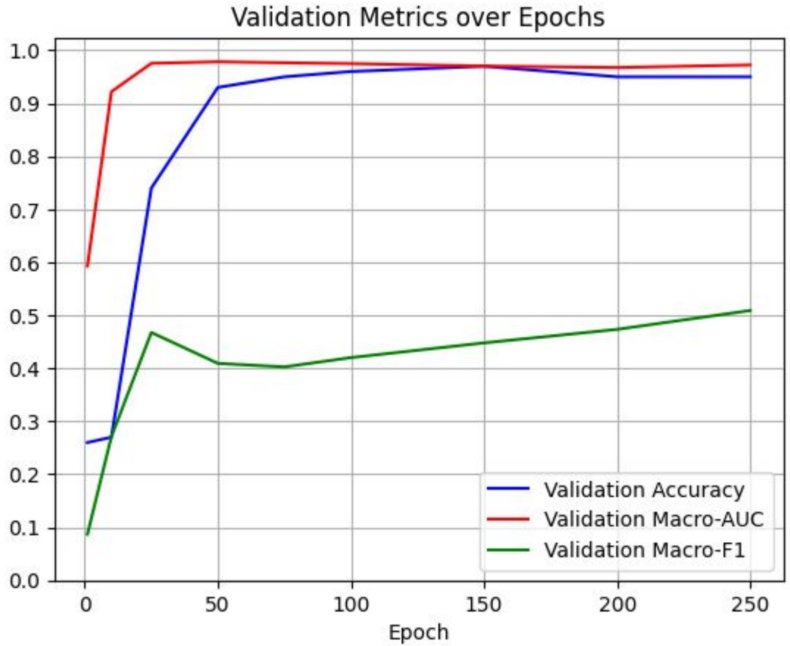
Method	Accuracy	Macro-Avg AUC-ROC	Macro Avg F1
1. No Smote	0.3636	0.7815	0.0051
1. With Smote	0.5481	0.8392	0.0641
2. No Smote	0.6494	0.9208	0.61347
2. With Smote	0.6545	0.9137	0.6007

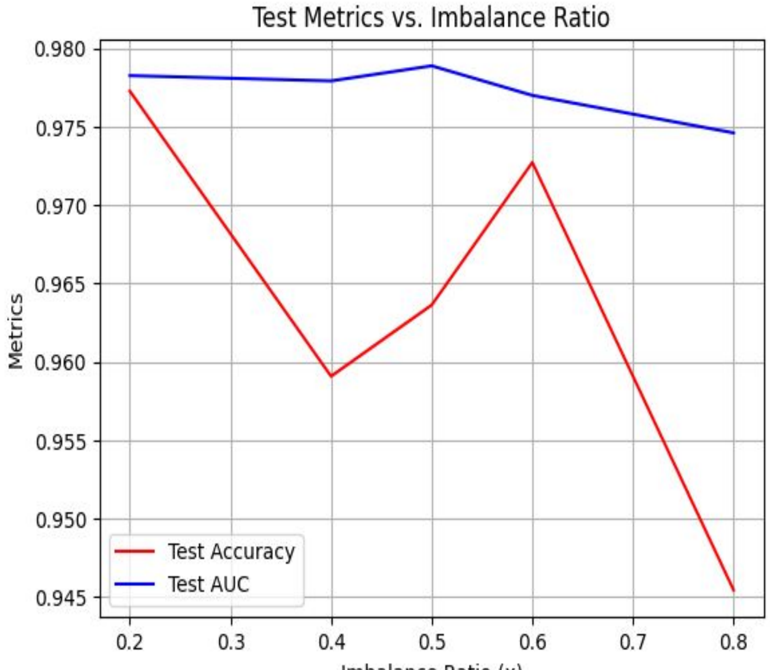
Table 1. Results for GraphSMOTE

Results for HeteroGSMOTE

Method	Accuracy	Macro AUC-ROC	Macro F1
With Smote	0.964	0.979	0.430
Without Smote	0.945	0.977	0.458

Table 2. Results Comparison





Further plans

01

Run our model in different settings, like using LSTM instead of FC, using entire data instead of masking, and introducing class imbalances in multiple classes within the dataset.

02

We also plan to do a comparative study of our model with other baseline models.

03

Additionally, we aspire to enrich our research by incorporating new and intriguing heterogeneous datasets, particularly those that provide labels for all nodes within the graph structure.

04

This will enable us to explore the potential benefits of meta-path-based oversampling, with a focus on elucidating their impact on model performance and generalization.

References

1. **HetGNN**: Zhang, C., Song, D., Huang, C., Swami, A., and Chawla, N. V. Heterogeneous graph neural network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 11. ACM, 2019. doi: 10.1145/3292500.333
2. **MagNN**: Fu, X., Zhang, J., Meng, Z., and King, I. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In Proceedings of The Web Conference 2020, pp. 11. ACM, 2020. doi: 10.1145/3366423.3380297
3. **HAN**: Wang, X., Ji, H., Shi, C., Wang, B., Cui, P., Yu, P., and Ye, Y. Heterogeneous graph attention network. In Sartor, J. B., D'Hondt, T., and De Meuter, W. (eds.), Proceedings of WWW 2019, pp. 4. ACM, 2019. doi: 10.475/123 4.
4. **Review**: Shi, C. Heterogeneous graph neural networks. In Wu, L., Cui, P., Pei, J., and Zhao, L. (eds.), Graph Neural Networks: Foundations, Frontiers, and Applications, pp. 351–369. Springer Singapore, Singapore, 2022.
5. **GraphSMOTE**: Zhao, T., Zhang, X., and Wang, S. Graphsmote: Imbalanced node classification on graphs with graph neural networks. In Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining, pp. 9. ACM, 2021. doi: 10.1145/3437963.3