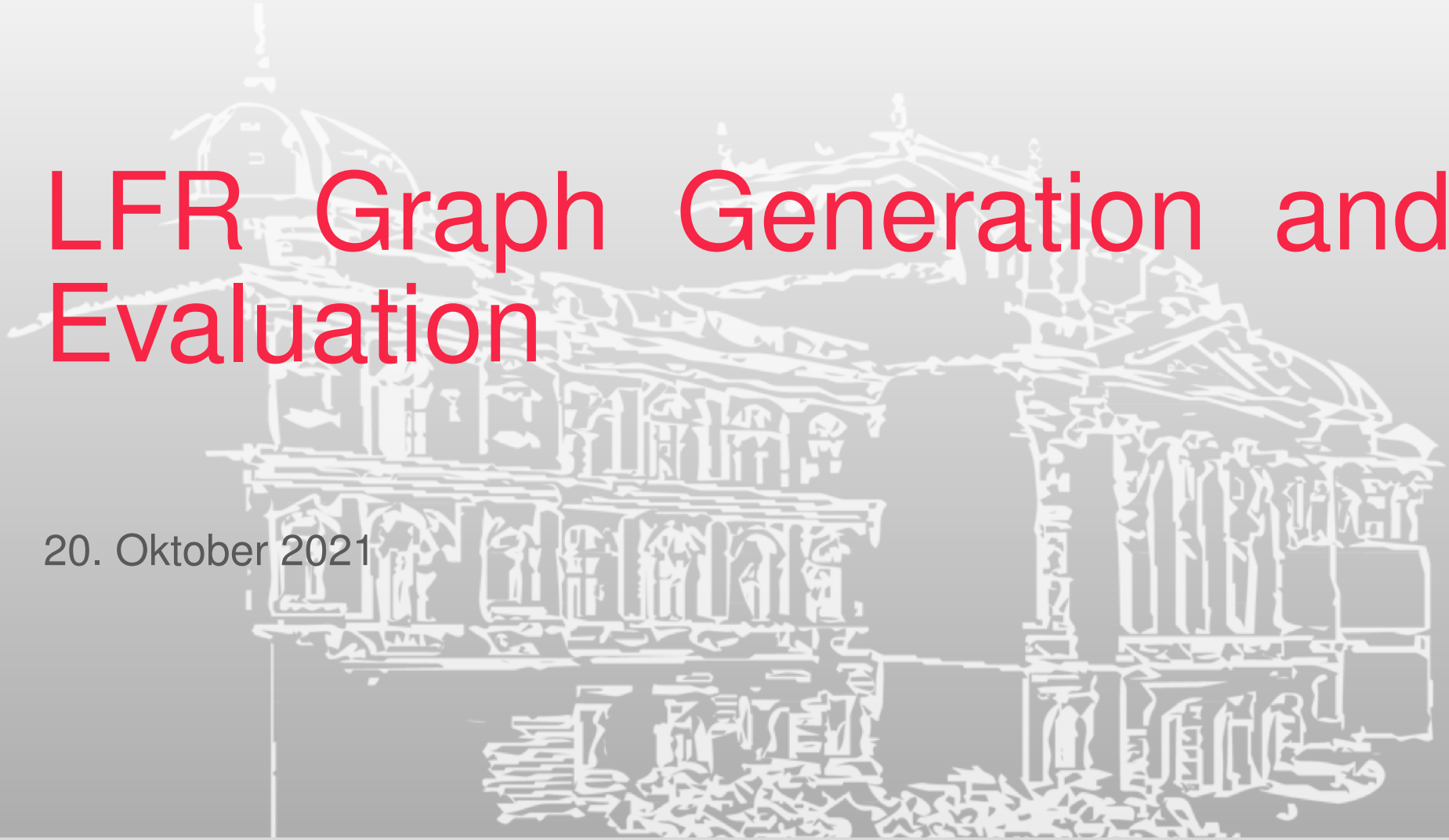
TU Graz

LFR Graph Generation and Evaluation



20. Oktober 2021

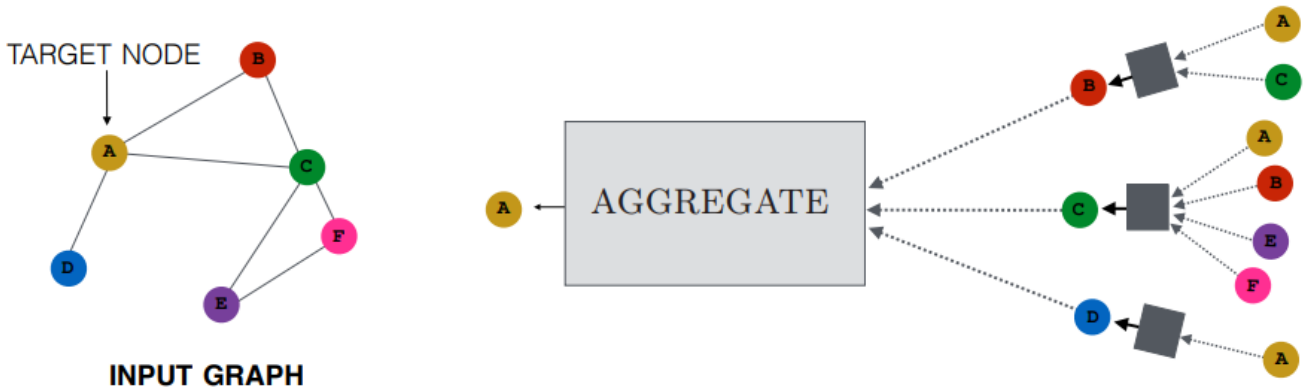
Content

Machine Learning on Graph Data

- Often real-world data is represented as graphs
- Graphs contain structural features which yield additional information
- **Node Classification**
  - Graph of nodes and edges G = (V, E)
  - Each node $u \in V$ has a feature vector x_u and label y_u
  - Nodes split into train, validation and test-set

Graph Neural Networks (GNN)

- Generate representations of nodes that depend on the structure of the graph
- Basic principle



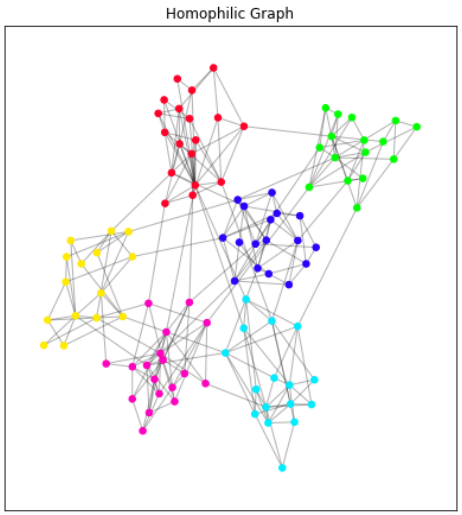
- Most GNNs share similar architecture

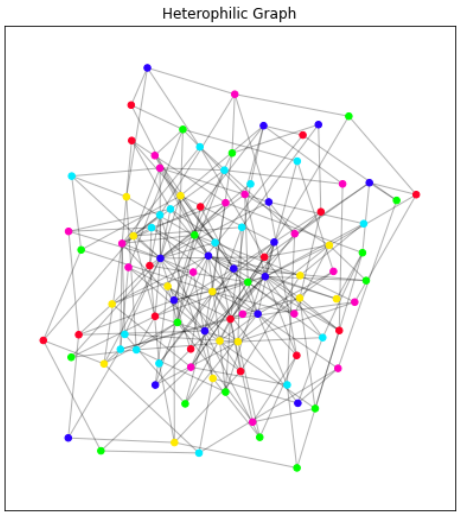
Graph Neural Networks (GNN)

- Evaluated GNNs
  - Graph Convolutional Network (GCN) [1]
  - Graph Attention Network (GAT) [2]
  - Simple Graph Convolution (SGC) [3]
  - Graph Sample and Aggregate (GraphSAGE) [4]
  - H2GCN [5]
- Baseline model (no graph structure)
  - 2-layer MLP

Homophily and Heterophily

- **Homophily** Tendency for nodes to share attributes with their neighbors in the graph
- High homophily $\Rightarrow$ low heterophily and vice versa





Lancichinetti-Fortunato-Radicchi (LFR) Benchmark [6]

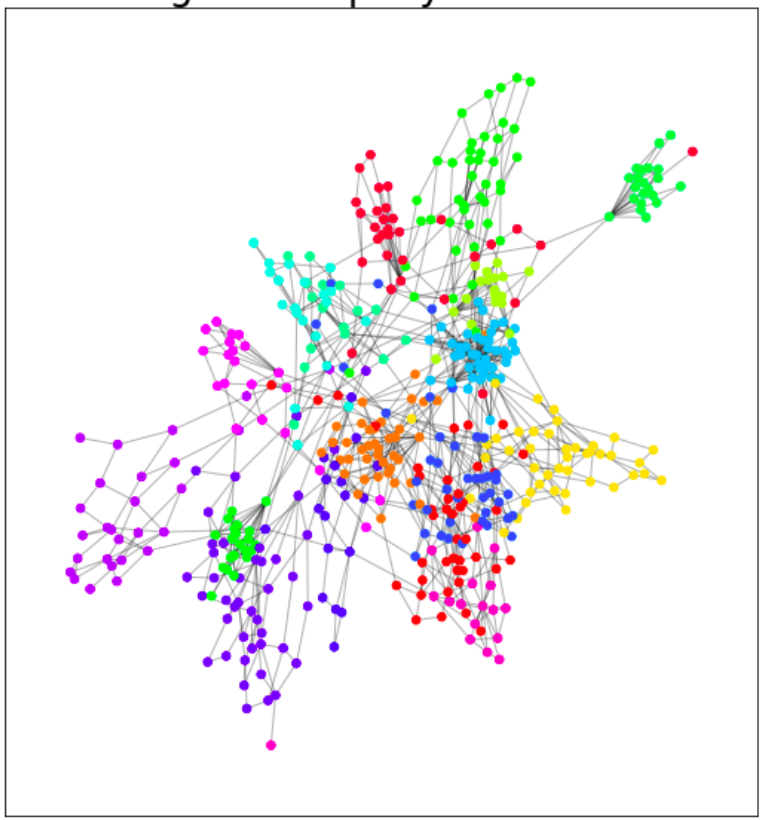
- Algorithm to generate artificial benchmark graphs with realistic node degree and community size distributions
- Nodes are part of defined community
- Controllable Parameters
  - Number of nodes and average degree
  - Edge-Homophily-Ratio
  - Community size and node degree distributions
- Drawback: Not robust for certain parameters

Graph Generator

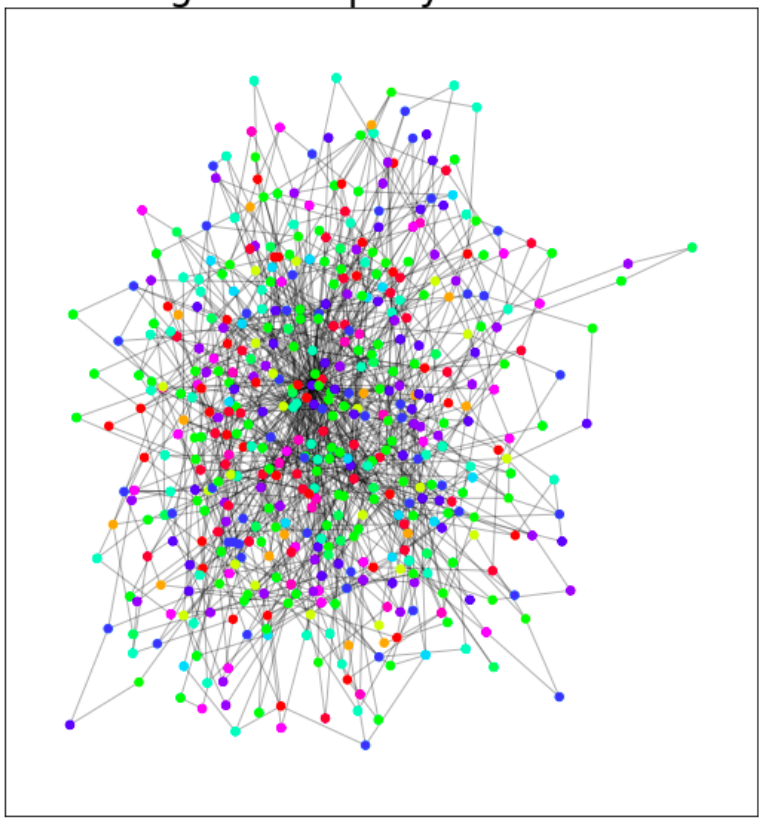
1. Generate graph with LFR Benchmark
   Parameters define structure and homophily of graph
2. Use the LFR community of a node to generate features
   - Features for nodes sampled from Gaussian distributions centered on the corner of a hypercube
   - Each community has a Gaussian centered on a different corner
   - Variance of Gaussians controls the separability

Example Graphs with 500 nodes

Edge-Homophily-Ratio: 0.9

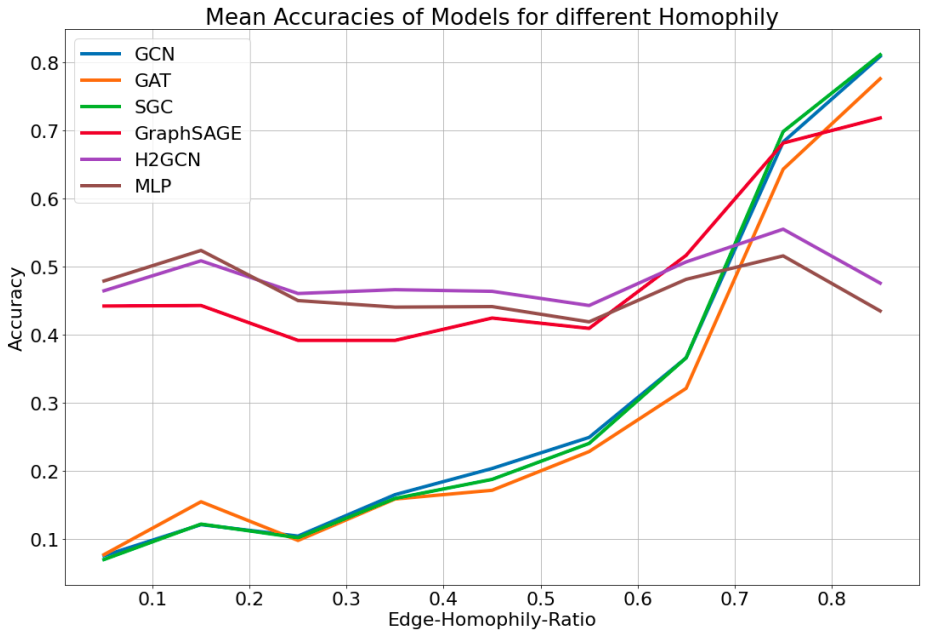


Edge-Homophily-Ratio: 0.1



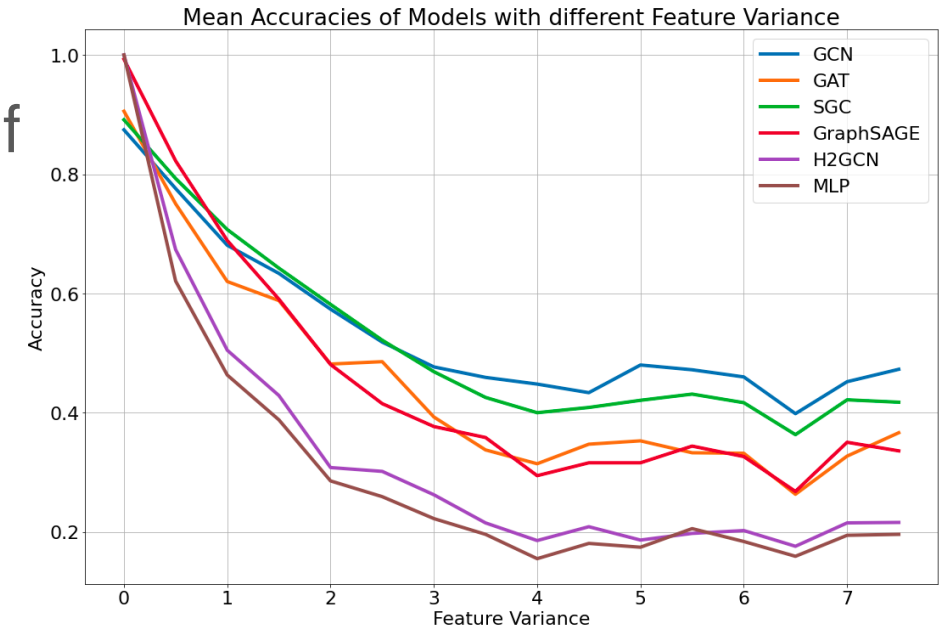
Controlling Homophily

- Edge-Homophily-Ratio controls accuracy of GNN models
- Higher homophily $\Rightarrow$ increased GNN accuracy
- MLP model accuracy stays the same



Controlling Community Separability

- Variance of Gaussians defines the separation of communities by the features
- Higher variance $\Rightarrow$ decreased GNN and MLP accuracy



Generate Graphs similar to popular Graph Datasets

	Original Cora	LFR Cora
Edges	5355	5667
Nodes	2708	2708
Node Degree Power-Law	1.91	1.98
Homophily Edge-Ratio	0.8	0.8
Homophily 1-hop	0.81	0.82
Homophily 2-hop	0.69	0.6
Uncertainty Coefficient [7]	0.63	0.36
Louvaine Communities	32.75	22
Avg. Clustering Coefficient	0.24	0.17

Sensitive Feature Idea

- Measure influence of a sensitive feature (could be gender, race etc.) on GNNs using the graph generator
- Hypothesis: GNNs utilize a sensitive feature based on communities better than MLP, because it gives the GNNs additional structural information

Method(1)

1. Generate Graph. The communities of LFR now called Sub-Communities
2. Add binary sensitive feature (**Controlled Feature**) to each node. Each Sub-Community has a certain probability for the Controlled Feature to be 1.
3. Sub-Communities are combined to a reduced number of Label-Communities using modularity optimization. These labels are then used for prediction.
4. As a baseline, the same is done with a randomly sampled binary feature (**Zodiac-Sign Feature**).

Method(2)

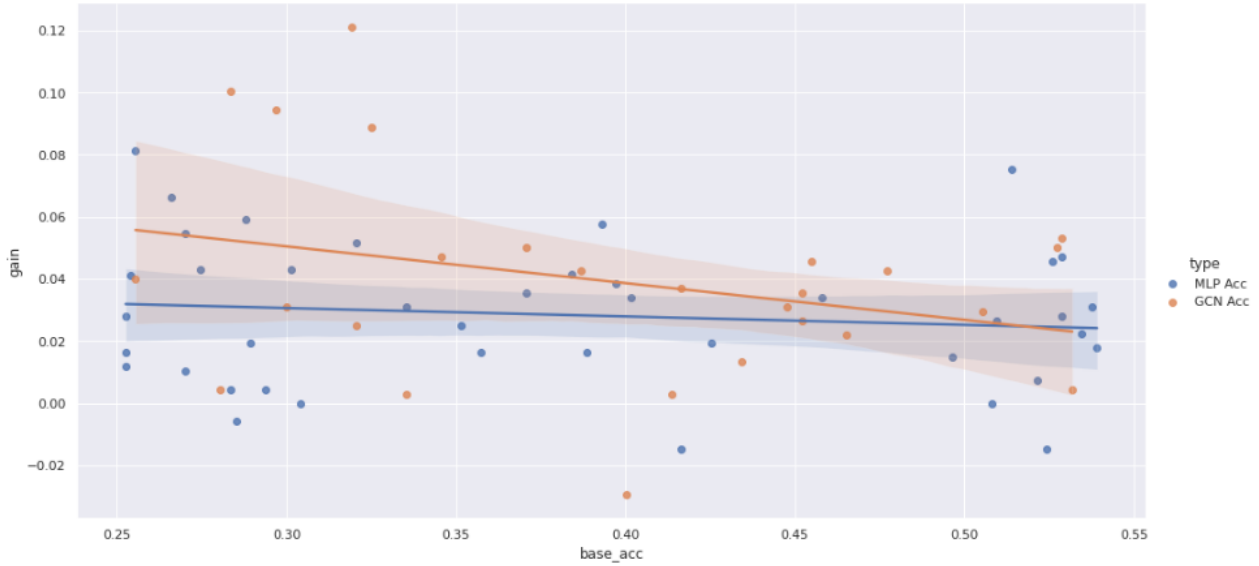
- Controlled Feature
  - Different distribution in Sub-Communities
  - Random distribution in Label-Communities
- Zodiac-Sign Feature
  - Random distribution in Sub-Communities
  - Random distribution in Label-Communities

Method(3)

5. Evaluate the 2 graphs and compute the accuracy gain of Controlled Feature Graph over the Zodiac-Sign Graph for GCN and MLP model
6. Repeat process for different feature variances

Results

- ≈25 Sub-Communities and 7 Label-Communities
- x-axis: Base accuracy with the Zodiac-Sign Feature
- y-axis: Accuracy gain with Controlled Feature



Discussion

- MLP performance very close to GCN performance
- Problem: Information Leakage
  Some information found between the Controlled Feature and the Label-Communities
- Further experiments necessary for conclusive results

Thank you for your Attention!

[1] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. 2017. arXiv: 1609.02907 [cs.LG].

[2] Petar Veličković et al. Graph Attention Networks. 2018. arXiv: 1710.10903 [stat.ML].

[3] Luca Pasa et al. Simple Graph Convolutional Networks. 2021. arXiv: 2106.05809 [cs.LG].

[4] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. 2018. arXiv: 1706.02216 [cs.SI].

[5] Jiong Zhu et al. Beyond Homophily in Graph Neural Networks: Current Limitations and Effective Designs. 2020. arXiv: 2006.11468 [cs.LG].

[6] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. "Benchmark graphs for testing community detection algorithms". In: Physical Review E 78.4 (Oct. 2008). ISSN: 1550-2376. DOI: 10.1103/physreve.78.046110.

[7] Hussain Hussain et al. "On the Impact of Communities on Semi-supervised Classification Using Graph Neural Networks". In: Studies in Computational Intelligence (2021), pp. 15–26. ISSN: 1860-9503. DOI: 10.1007/978-3-030-65351-4_2.