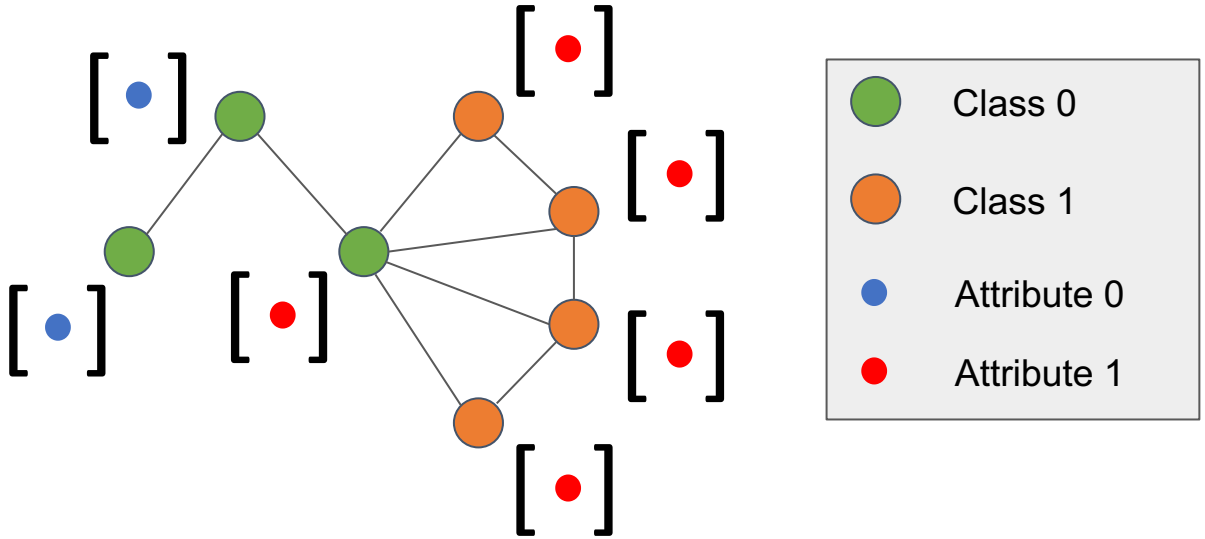
GEMS LAB

Understanding Graph Neural Network Fairness in the Presence of Heterophilic Neighborhoods

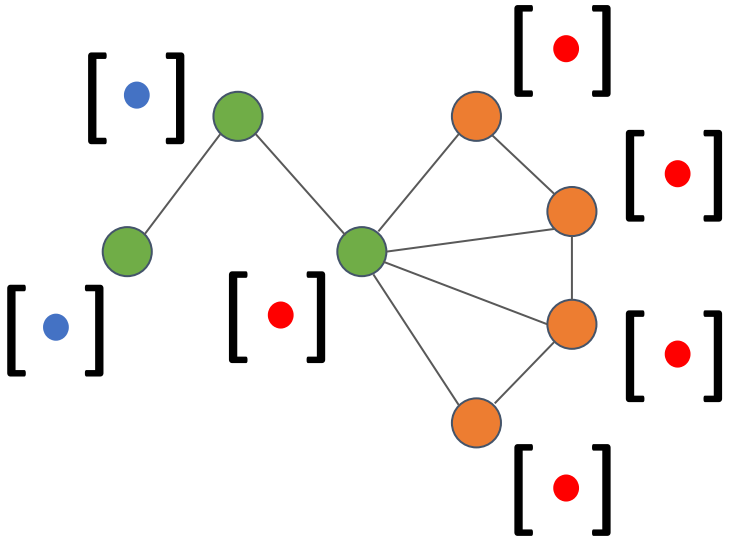
Donald Loveland

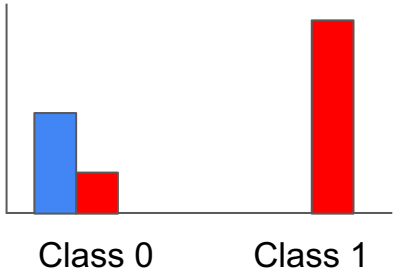
Collaborators: Jiong Zhu, Mark Heimann, Ben Fish, Michael Schaub, Danai Koutra

Problem Set Up: Fairness on Graphs



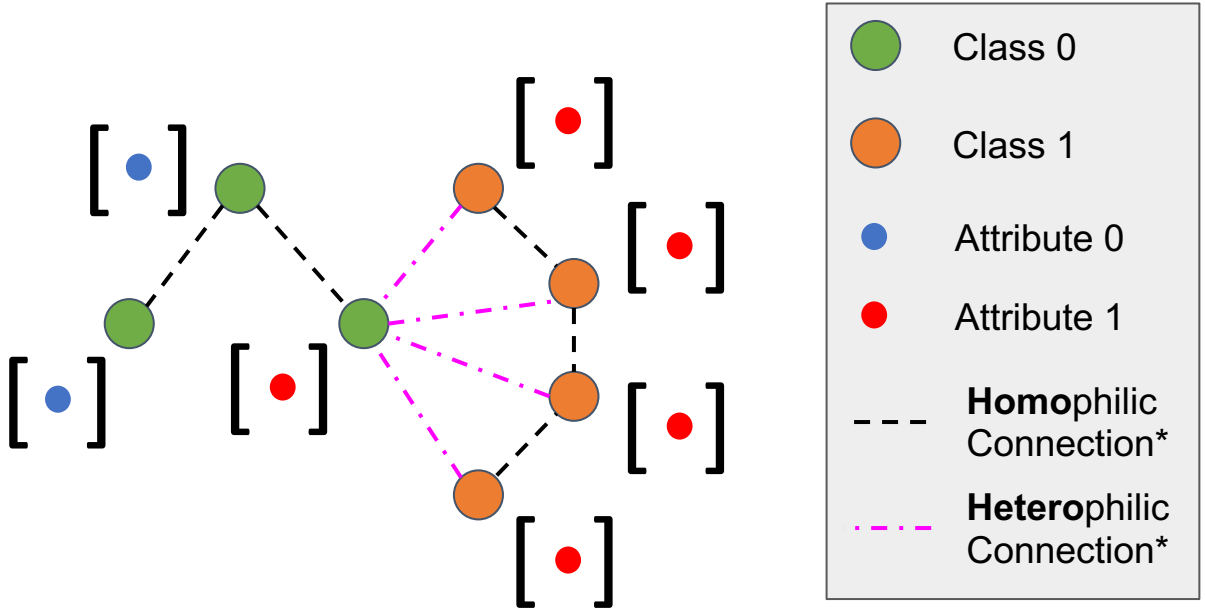
Problem Set Up: Fairness on Graphs





Class-attribute skew

Homophily



Homophily can potentially amplify class-attribute skew depending on the GNN aggregation mechanism.

**In terms of class*

Graph Neural Network Fairness Evaluation

Common approach: report *global* homophily ratio (intra-class edges/total edges)

Implicit assumption: The distribution of *local* homophily ratios (homophily ratio of k-hop neighborhood around a node) is strongly related to the *global* homophily ratio.

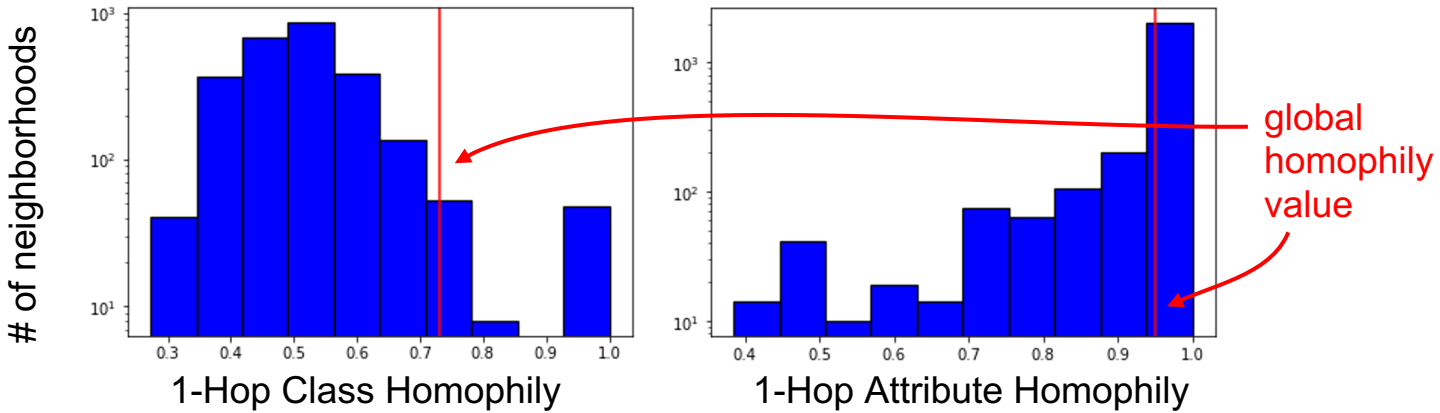
Graph Neural Network Fairness Evaluation

Common approach: report *global* homophily ratio (intra-class edges/total edges)

often not true

Implicit assumption: The distribution of *local* homophily ratios (homophily ratio of k-hop neighborhood around a node) is strongly related to the *global* homophily ratio.

Pokec Dataset:



Our Contributions

Characterize the impacts of *neighborhood patterns* on *fairness* in different GNN architectures.

Homophily-assumed Models

- e.g., GCN[1], SGC[2], GAT[3]

Heterophily-adjusted Models

- e.g., GCN-II[4], FA-GCN[5], GraphSAGE[6], H2-GCN[7]

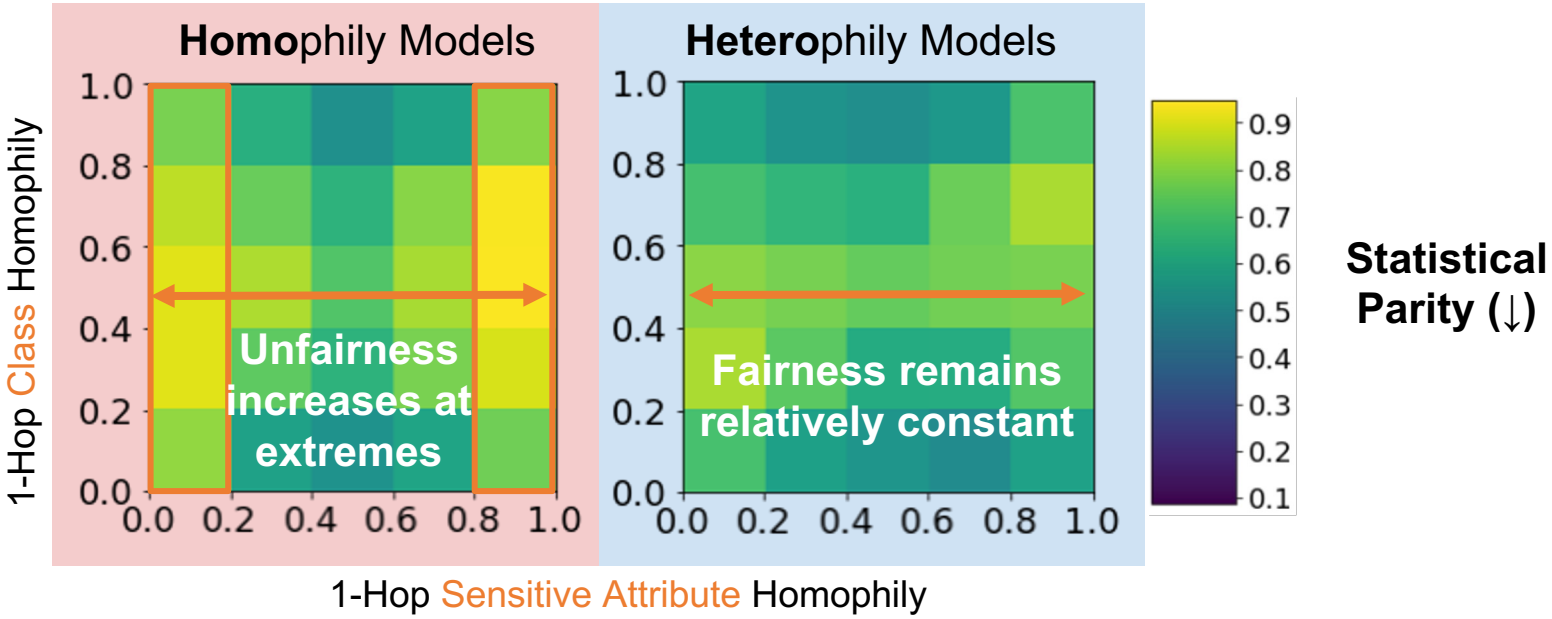
Statistical Parity

$$\mathbb{P}(G(x) = 1 | x_s = 1) - \mathbb{P}(G(x) = 1 | x_s = 0)$$

[[1]Kipf & Welling, ICLR'17; [2]Wu et al. ICML'19; [3]Veličković et al. ICLR'18; [4]Chen et al. ICML'20; [5]Bo et al. AAAI'20; [6]Hamilton et al. NeurIPS'17; [7]Zhu et al. NeurIPS'20]

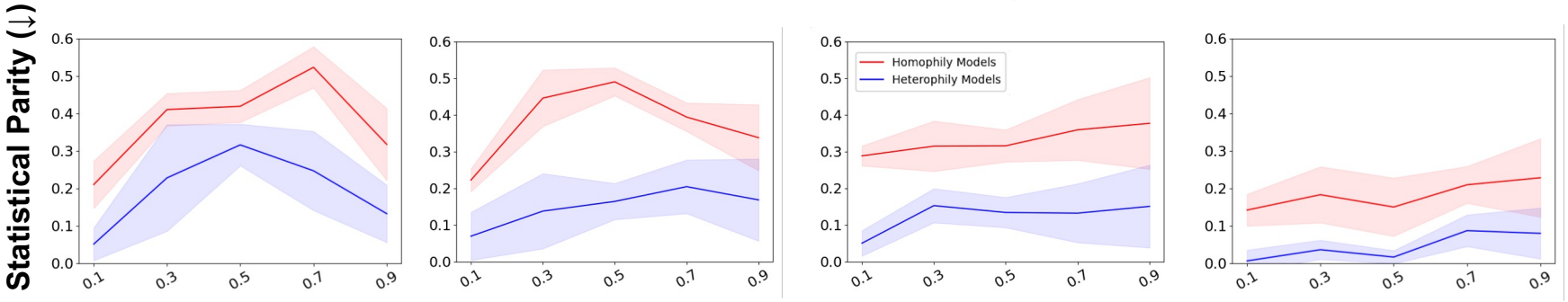
Major Takeaways: Synthetic Data

Heterophily-adjusted models are able to **improve fairness** by upwards of 20% in settings with both high heterophily and high homophily; less likely to overgeneralize than homophily-assumed models.



Major Takeaways: Synthetic Data

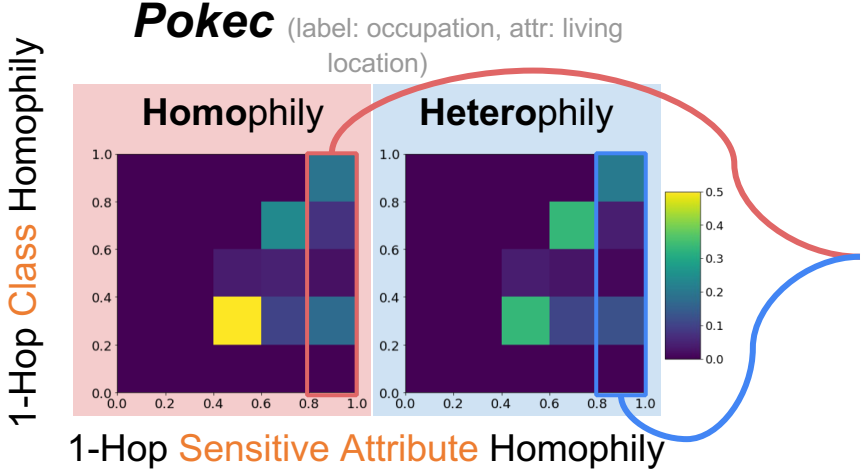
Homophily models produce **high unfairness** even with minimal access to the sensitive attribute due to a non-linear trend in bias amplification.

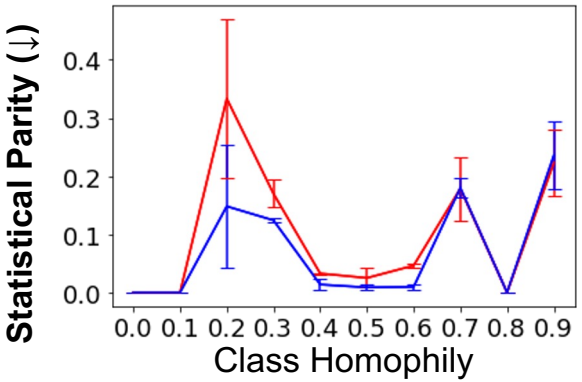


Major Takeaways: Real-world Data

Dataset	Global Class Homophily	Global Sensitive Attribute Homophily	Model Type	F1	Statistical Parity
Pokec	0.75	0.96	Homophily	0.63±0.02	0.064±0.042
			Heterophily	0.72±0.00	0.011±0.006
NBA	0.40	0.73	Homophily	0.66±0.05	0.084±0.055
			Heterophily	0.76±0.01	0.041±0.032

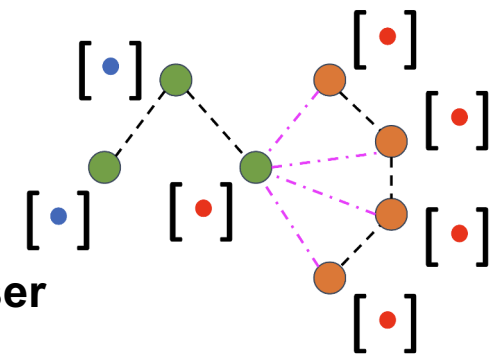
Zooming in to unfair regions, **hetero**phily models are **more fair**, particularly in high heterophily settings.





Understanding Graph Neural Network Fairness in the Presence of Heterophilic Neighborhoods

- **Hetero**phillic models can **improve fairness** (at no loss of predictive performance) by up to
  - [synthetic data] 20% in high homophilic and heterophilic neighborhoods.
  - [real data] 18% in class heterophilic neighborhoods.
- **Hetero**phillic models **amplify class-attribute skew to a lesser degree** than homophilic models despite strong homophily.



CSE

Donald Loveland
(dlovelan@umich.edu)

GEMS LAB