Learning Individually Fair Graph Neural Networks

Student: Elly Do

Mentor: Bang An, Furong Huang

REU-CAAR 2021

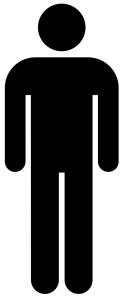
Outline

- Definition of Individual Fairness
- Graph Neural Networks (GNN)
- Why is learning individual fairness a problem for GNN?
- Our project
  - Problem setting and notations
  - Method
  - Conclusion and future work

Individual Fairness

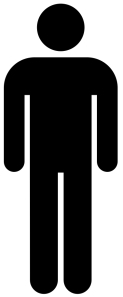
means treating similar individuals similarly

Black person
3.98 GPA
329 GRE



In a grad school admission setting

=



White person
3.98 GPA
329 GRE

"... any two individuals who *are similar with respect to a particular task* should be classified similarly."

(Dwork et al. 2011)

Graph Neural Networks (GNN)

- GNN is a neural network that deals with graph data such as social networks, traffic networks, and molecules.
- It works directly on graph structure where there is a variety of unordered nodes, each of which shares edges with a different number of other nodes.

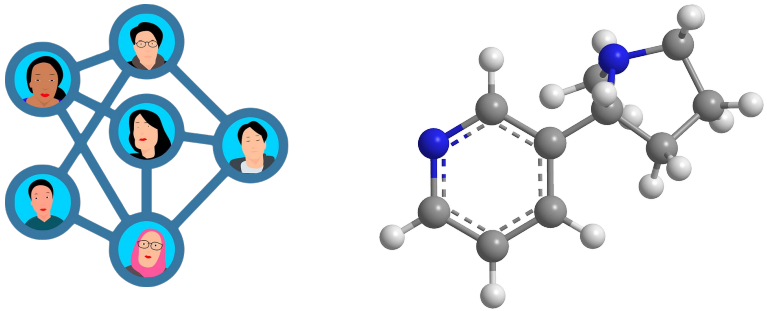


Figure 1: Graph data

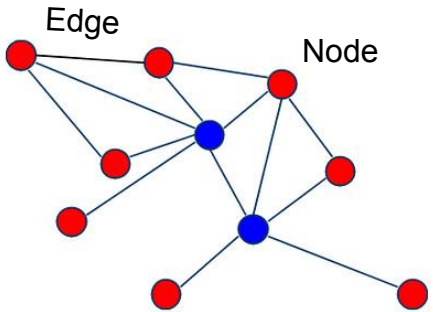
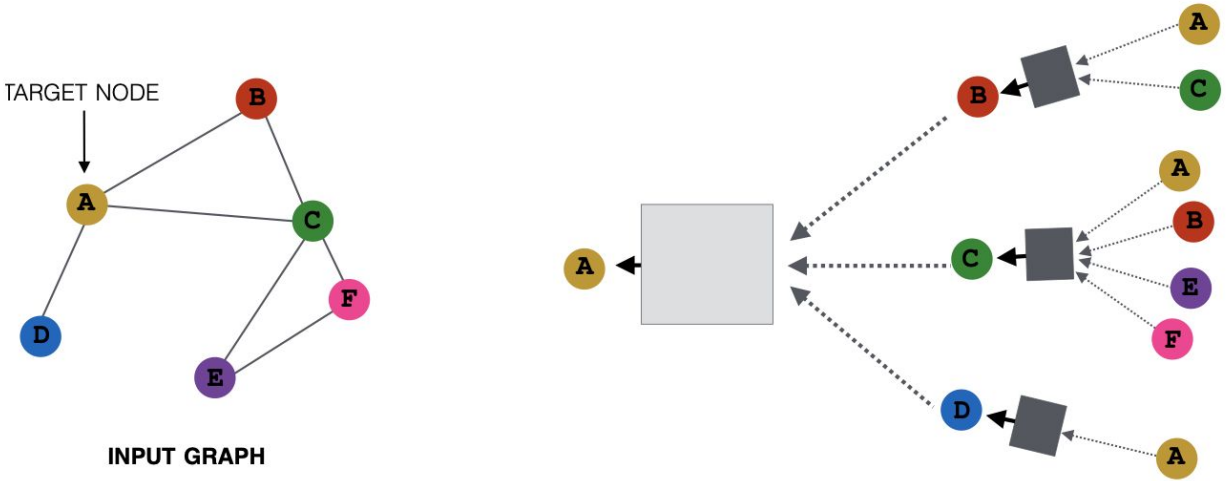


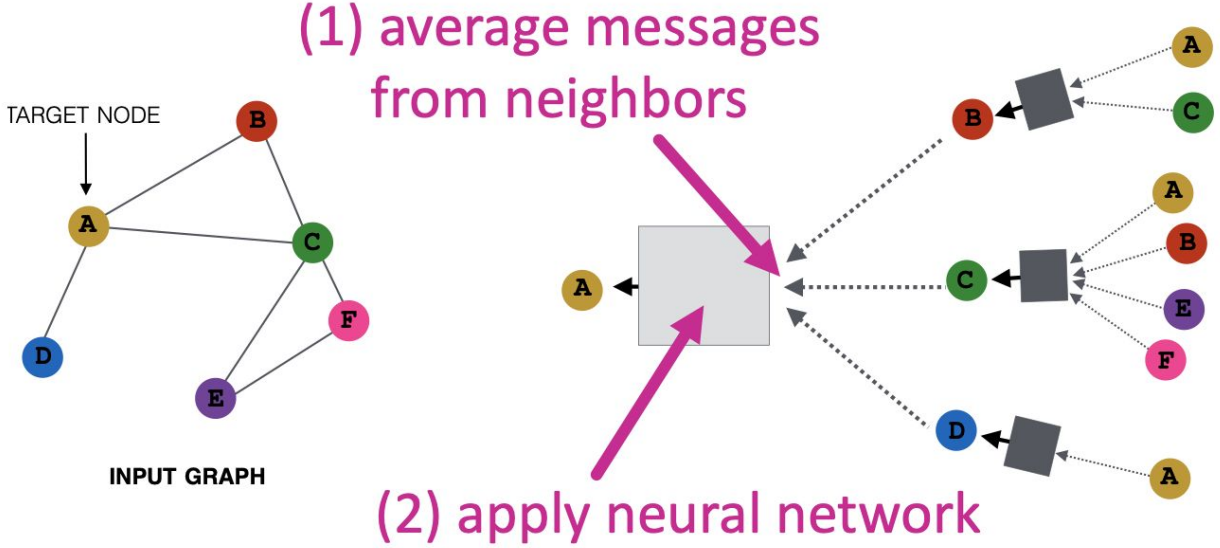
Figure 2: Graph Structure

Neighborhood aggregation (message passing) is the main idea of GNN.
That is to generate node embeddings from a node's neighbors.



Leskovec, J. (2018). *Graph Neural Networks* [PowerPoint slides]. Stanford. https://web.stanford.edu/class/cs224w/slides/08-GNN.pdf.

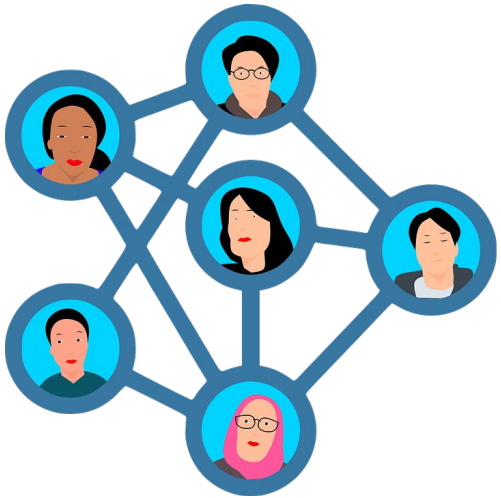
Approach: Average messages from neighbors and then apply neural network.



Leskovec, J. (2018). *Graph Neural Networks* [PowerPoint slides]. Stanford. https://web.stanford.edu/class/cs224w/slides/08-GNN.pdf.

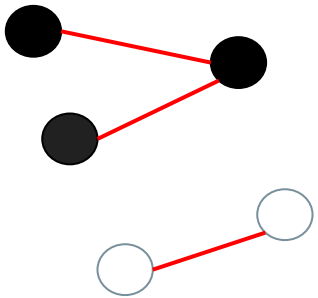
Why is learning individual fairness a problem for GNN?

- Similar nodes in the input have similar embeddings.
  - For instance, imagine we have the social networks:



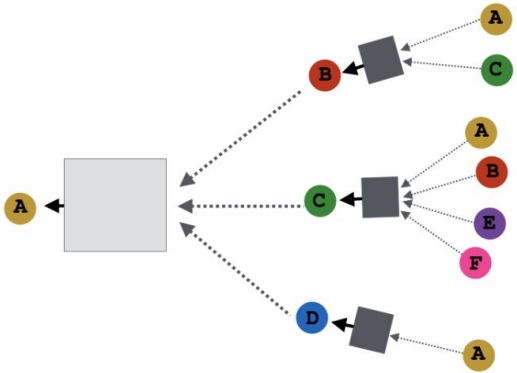
Why is learning individual fairness a problem for GNN?

- Similar nodes in the input have similar embeddings.
  - For instance, imagine we have the social networks:



Black people tend to connect with each other.

The same to white people.



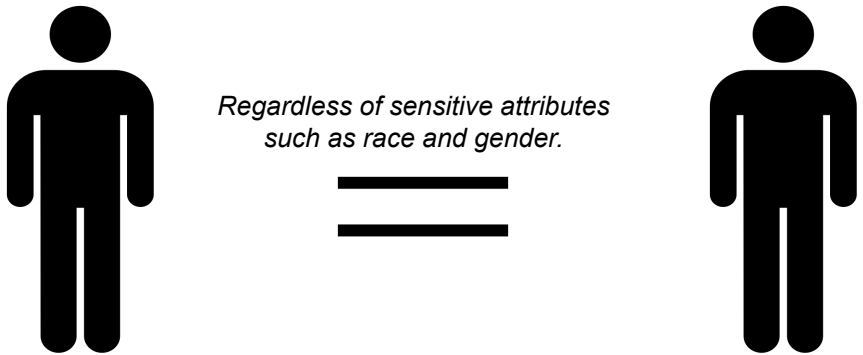
Recall: Node embeddings are based on the target node's neighborhood.

People of the same race are likely to have the same node embedding!

This leads to bias

Why is learning individual fairness a problem for GNN?

- Based on this neighborhood aggregation mechanism, two similar individuals with respect to a certain task can have very different embeddings, which leads to different outcomes.
- Meanwhile, what we want is:

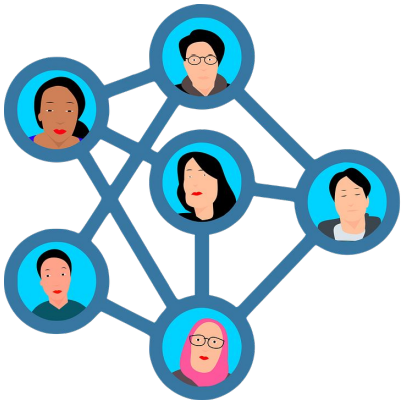


- That is in the node embeddings, nodes that we want to classified similarly should have similar embedding. To achieve that in our project, we push our model to perform well in the link prediction task on fairness graph.

Our Approach

Problem Setting and Notations

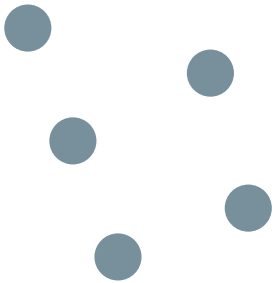
- To illustrate, take an example, we have a dataset of social network users where we know who connects to whom.
- Classification task: To predict whether the income of these users is above median.
- Goal: Accurate classification with minimal bias.



The classification task

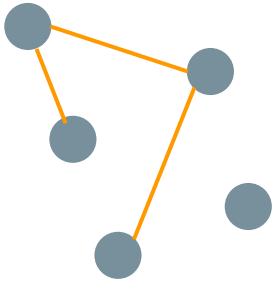
The training data: fairness graph

- A human judgment determines pairwise judgment that decides which individuals should be treated similarly with each other, that is people having similar salary should be classified similarly.
- The fairness graph G captures these pairwise judgment in which similar individuals share an edge.



Human judgment

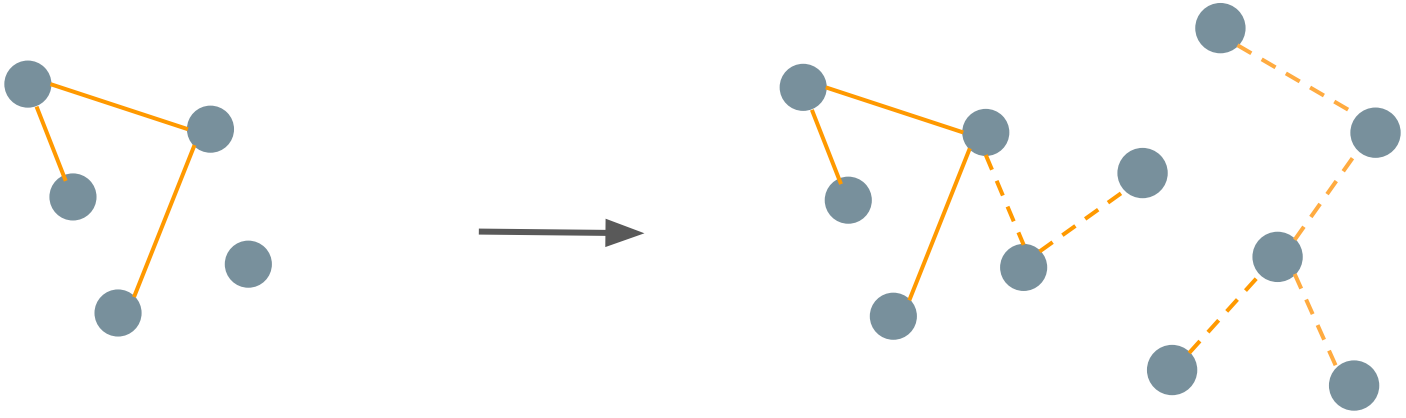
can only make judgment for a subgraph



Lahoti, L., Gummadi, K., & Weikum, G. (2019). Operationalizing individual fairness with pairwise fair representations. *Proceedings of the VLDB Endowment, 13*, 506–518. http://dx.doi.org/10.14778/3372716.3372723.

Link prediction based on the fairness graph

Predict accurate links between similar individuals

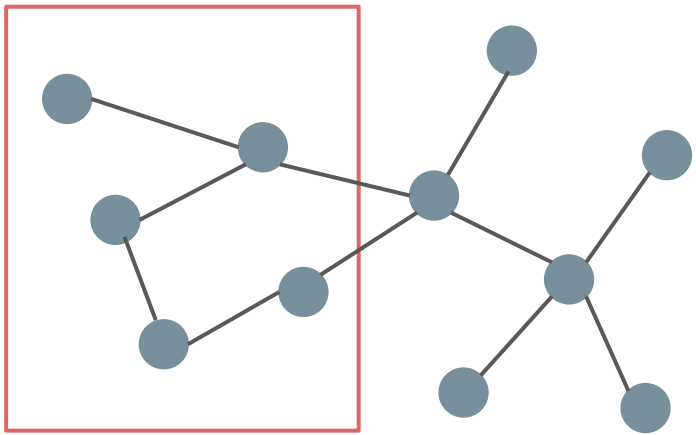


Fairness graph from human judgment

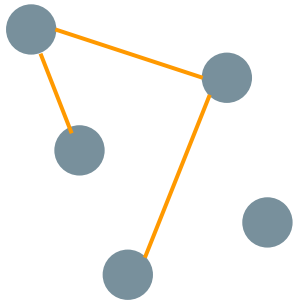
The primary graph with predicted links

We want links between individuals having similar salary

To sum up what we do

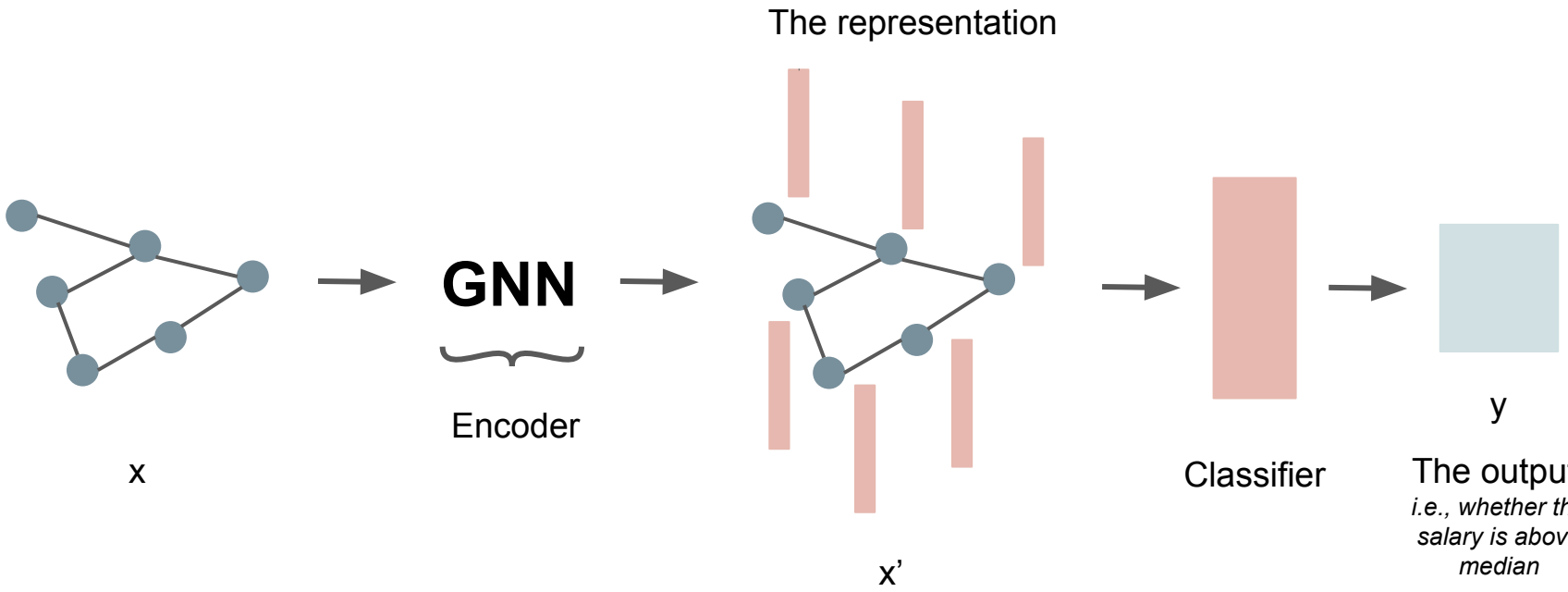


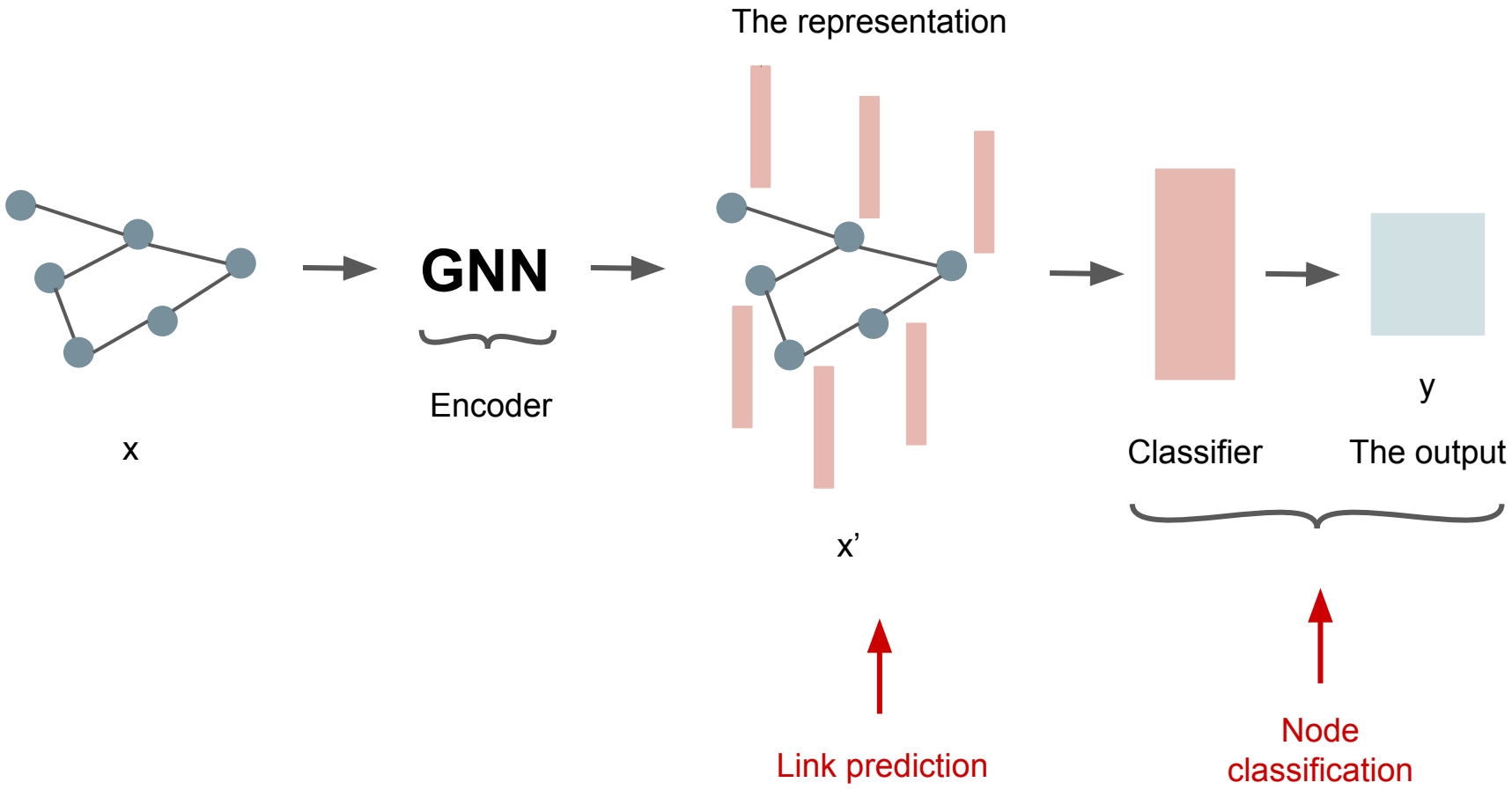
- Goal: to learn an individually fair node classifier.
- Model: GNN.
- Primary task: Node classification.
- Auxiliary task: Link prediction.

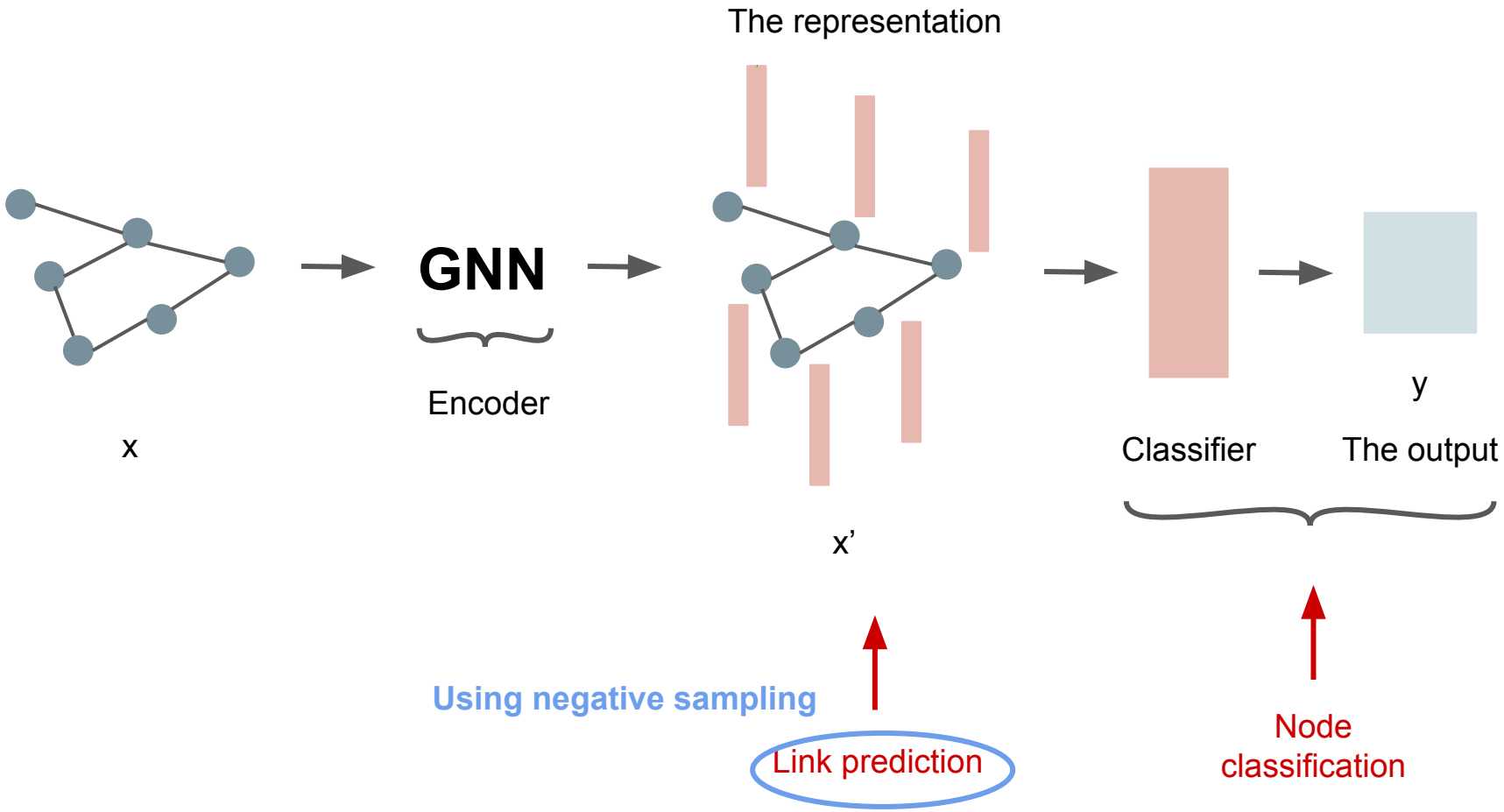


Fairness graph is the training data.

The method


The representation
GNN
Encoder
x
x'
Classifier
y
The output,
i.e., whether the salary is above median

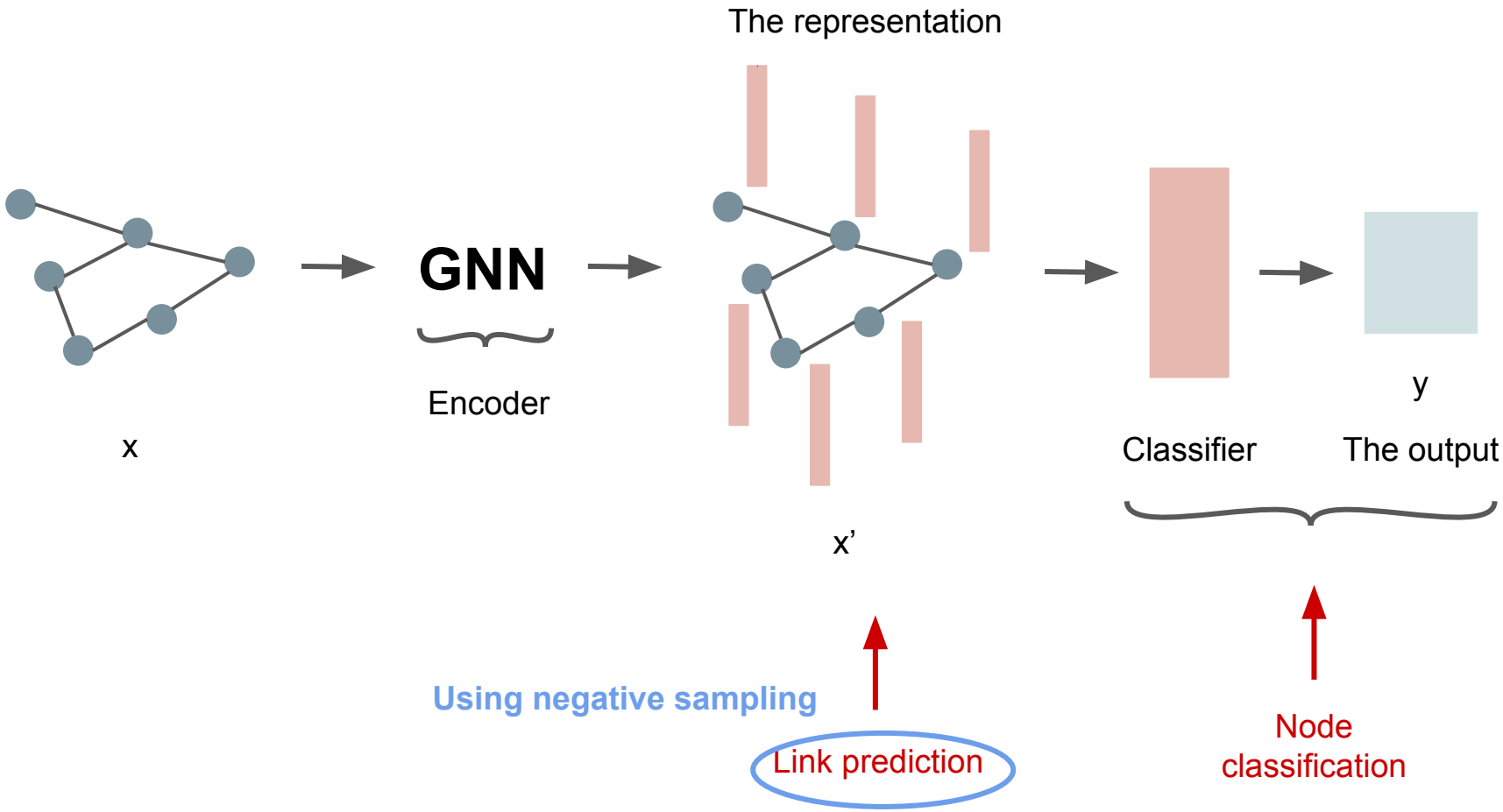

The representation
GNN
Encoder
x
x'
Classifier
The output
y
Link prediction
Node classification

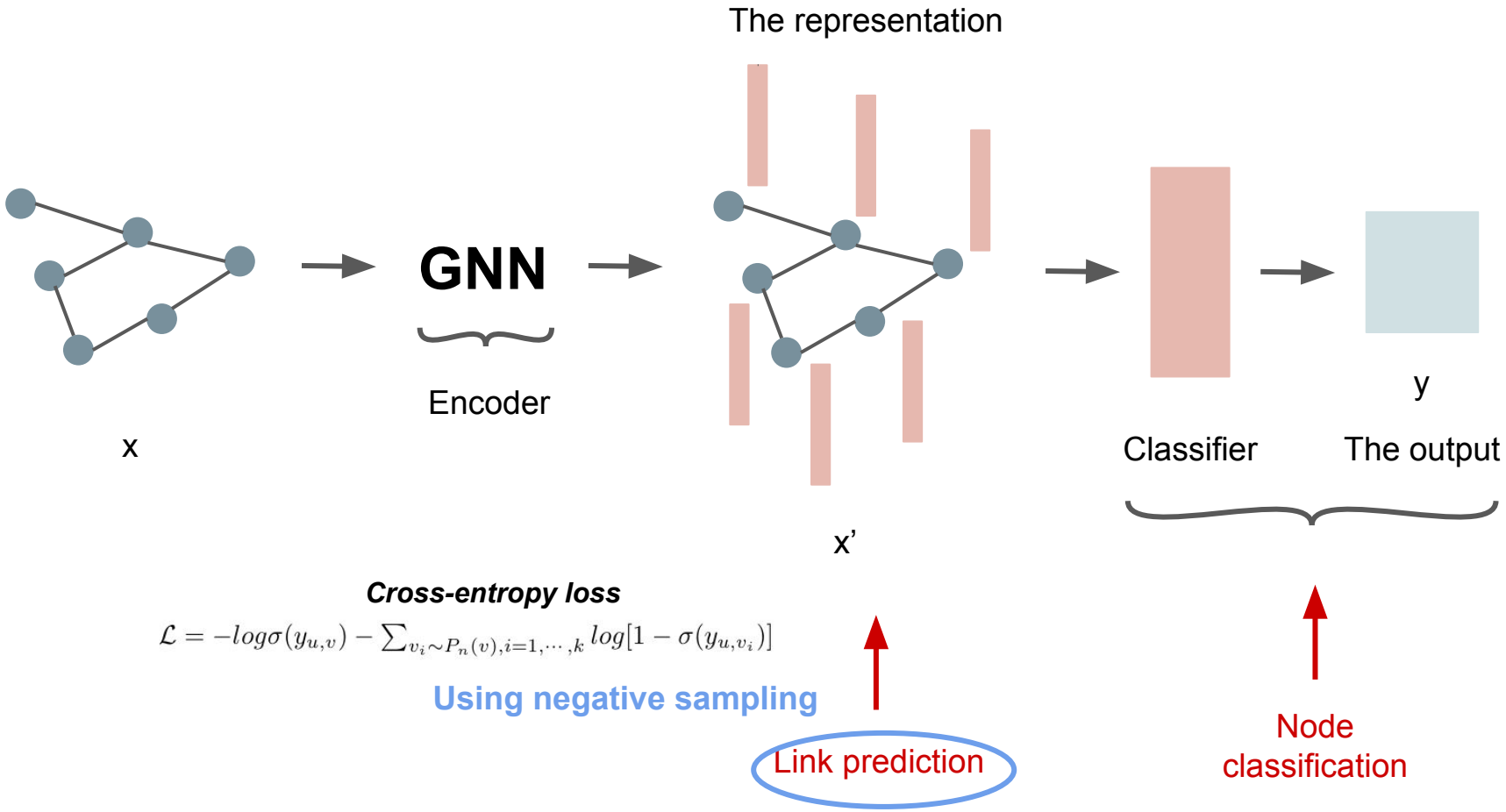

The representation
GNN
Encoder
x
x'
y
Classifier
The output
Using negative sampling
Link prediction
Node classification

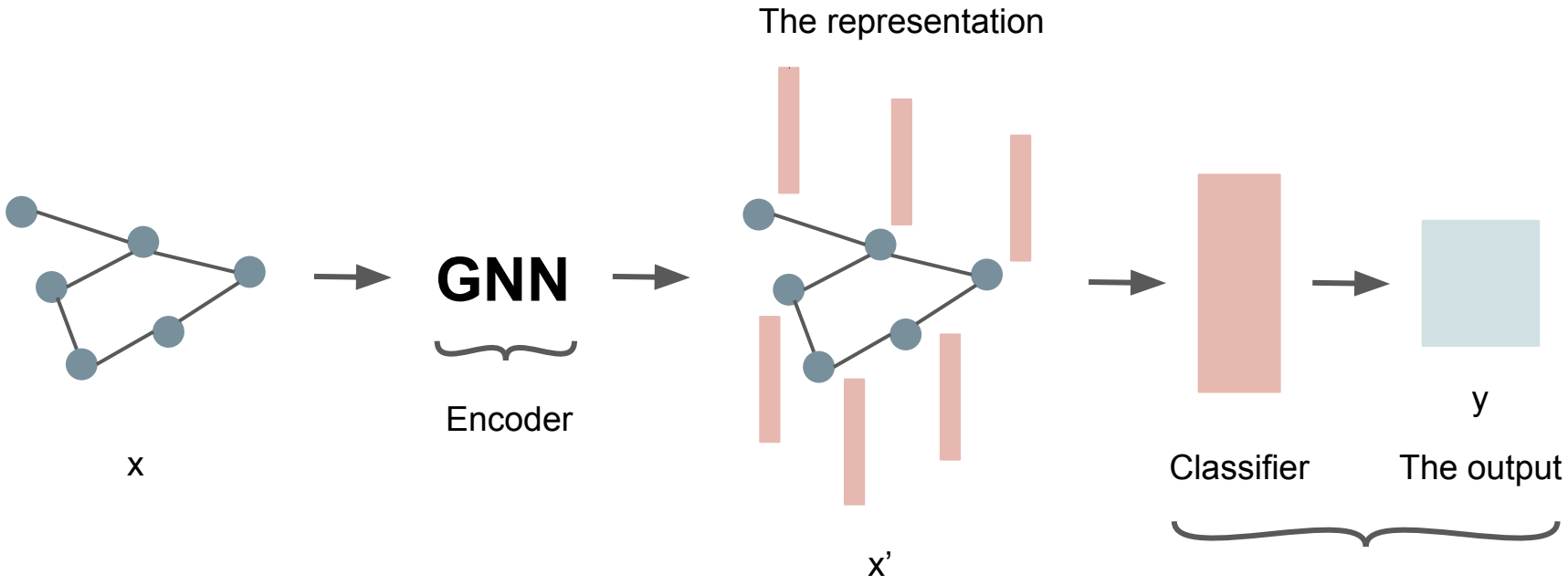
Negative Sampling

- In a GNN model, let *u* and *v* be two similar nodes whose node representations are $h_u^{(L)}$ and $h_v^{(L)}$ respectively.
- Negative sampling means:
  - Compare the scores between nodes sharing an edge against the scores between nodes that are not neighbors.
  - For instance, we want to enforce the score between node *u* and *v* to be higher than the score between node *u* and node *v'* from an arbitrary noise distribution $v' \sim P_n v$.

Link Prediction. DGL. Retrieved March 18, 2020, from https://docs.dgl.ai/en/0.6.x/guide/training-link.html.


The representation
GNN
Encoder
x
x'
y
Classifier
The output
Using negative sampling
Link prediction
Node classification


The representation
GNN
Encoder
x
x'
y
Classifier
The output
Cross-entropy loss
$\mathcal{L} = -log\sigma(y_{u,v}) - \sum_{v_i \sim P_n(v), i=1,\cdots,k} log[1 - \sigma(y_{u,v_i})]$
Using negative sampling
Link prediction
Node classification



Cross-entropy loss

$$\mathcal{L} = -log\sigma(y_{u,v}) - \sum_{v_i \sim P_n(v), i=1,\cdots,k} log[1 - \sigma(y_{u,v_i})]$$

Using negative sampling

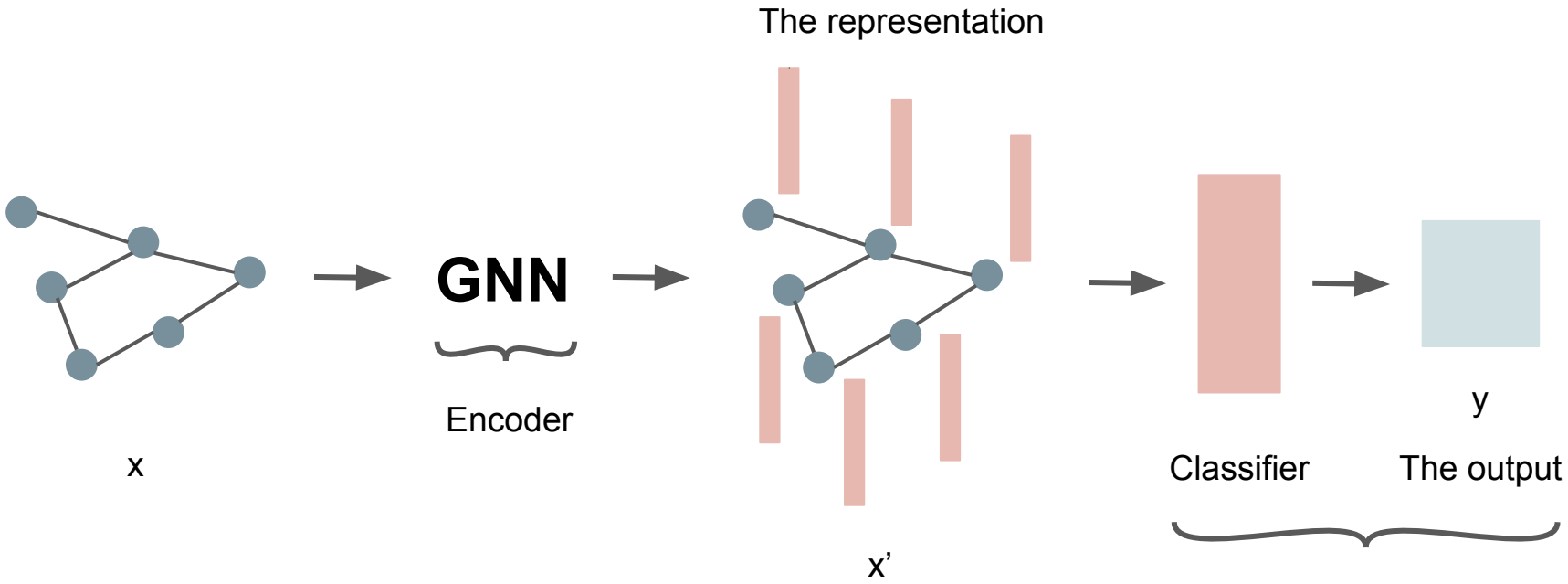
Link prediction

Node classification

Loss function:

$$\mathcal{L} = -\sum_{l \in \mathcal{Y}_L} \sum_{f=1}^{F} Y_{lf} \ln Z_{lf} ,$$

Kipf, T. & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. *arXiv*. https://arxiv.org/abs/1609.02907.



Cross-entropy loss

$$\mathcal{L} = -log\sigma(y_{u,v}) - \sum_{v_i \sim P_n(v), i=1,\cdots,k} log[1 - \sigma(y_{u,v_i})]$$

Using negative sampling

Trade-off?

Link prediction

Node classification

Loss function:

$$\mathcal{L} = -\sum_{l \in \mathcal{Y}_L} \sum_{f=1}^{F} Y_{lf} \ln Z_{lf} ,$$

Kipf, T. & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. *arXiv*. https://arxiv.org/abs/1609.02907.

Conclusion and Future Work

- We study individual fairness for GNN by propose a heuristic method using the link prediction for the fairness graph.
- Work on the trade-off between two loss functions.
- Experiments to show that this method works.
- Combine with group fairness (another type of fairness) to see if we can achieve both at the same time.

Thank you!

Questions?