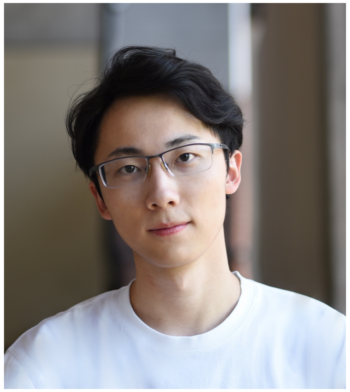
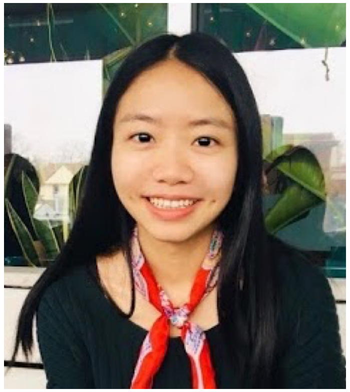
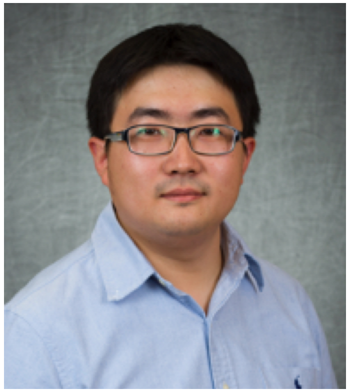
Fairness in Graph Machine Learning: Recent Advances and Future Prospectives



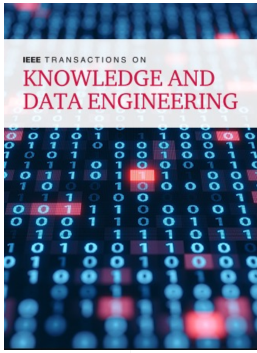



[1]Yushun Dong [2]Oyku Deniz Kose [2]Yanning Shen [1]Jundong Li

[1]University of Virginia

[2]University of California, Irvine

Related Materials of this Tutorial



This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2023.3265598

Fairness in Graph Mining: A Survey

Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li

—Graph mining algorithms have been playing a significant role in myriad fields over the years. However, despite their performance on various graph analytical tasks, most of these algorithms lack fairness considerations. As a consequence, lead to discrimination towards certain populations when exploited in human-centered applications. Recently, algorithmic as been extensively studied in graph-based applications. In contrast to algorithmic fairness on independent and identically (i.i.d.) data, fairness in graph mining has exclusive backgrounds, taxonomies, and fulfilling techniques. In this survey, we comprehensive and up-to-date introduction of existing literature under the context of fair graph mining. Specifically, we novel taxonomy of fairness notions on graphs, which sheds light on their connections and differences. We further present an summary of existing techniques that promote fairness in graph mining. Finally, we discuss current research challenges and stions, aiming at encouraging cross-breeding ideas and further advances.

ms—Algorithmic Fairness, Graph Mining, Debiasing

1 Introduction

Graph-structured data is pervasive in diverse real-world applications, e.g., E-commerce [102], [121], health care [37], [53], traffic forecasting [72], [100], and drug discovery [15], [172]. In recent years, a number of graph mining algorithms have been proposed to gain a deeper understanding of such data. These algorithms have shown promising performance on graph analytical tasks such as node classification [59], [86], [161] and link prediction [4], [103], [109], contributing to great advances in many graph-based applications.

Despite the success of these graph mining algorithms, most of them lack fairness considerations. Consequently, they could yield discriminatory results towards certain populations when such algorithms are exploited in human-centered applications [80]. For example, a social network-based job recommender system may unfavorably recommend fewer job opportunities to individuals of a certain gender [97] or individuals in an underrepresented ethnic group [150]. With the widespread usage of graph mining algorithms, such potential discrimination could also exist in other high-stake applications such as disaster response [159], criminal justice [3], and loan approval [136]. In these applications, critical and life-changing decisions are often made for the individuals involved. Therefore, how to tackle unfairness issues in graph mining algorithms naturally becomes a crucial problem.

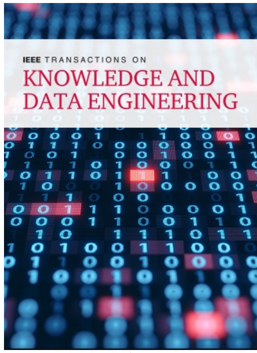
- *Y. Dong is with Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, Virginia, US. E-mail: yd6eb@virginia.edu*
- *J. Ma is with Department of Computer Science, University of Virginia, Charlottesville, Virginia, US. E-mail: jm3mr@virginia.edu*
- *S. Wang is with Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, Virginia, US. E-mail: sw3wv@virginia.edu*
- *C. Chen is with Biocomplexity Institute, University of Virginia, Charlottesville, Virginia, US. E-mail: zrh6du@virginia.edu*
- *J. Li is with Department of Electrical and Computer Engineering, Department of Computer Science, and School of Data Science, University of Virginia, Charlottesville, Virginia, US. E-mail: jundong@virginia.edu*

Compared with achieving fairness in the context of independent and identically distributed (i.i.d.) data, fulfilling fairness in graph mining can be non-trivial due to two main challenges. The first challenge is to formulate proper fairness notions as the criteria to determine the existence of unfairness (i.e., bias). Although a vast amount of traditional algorithmic fairness notions have been proposed centered on i.i.d. data [42], [111], they are unable to reflect the bias exhibited by the relational information (i.e., the topology) in graph data. For example, the same population can be connected with different topologies as in Fig. 1a and 1b, where each node represents an individual, and the color of nodes denotes their demographic subgroup membership, such as different genders. Compared with the graph topology in Fig. 1a, the topology in Fig. 1b has more intra-group edges than inter-group edges. The dominance of intra-group edges in the graph topology is a common type of bias existing in real-world graphs [39], [41], [70], which cannot be captured by traditional algorithmic fairness notions. The second challenge is to prevent the graph mining algorithms from inheriting the bias exhibited in the input relational information [41], [112], [148], [160]. We present a toy example to demonstrate how the information propagation mechanism in Graph Neural Networks (GNNs) [64], [86], [161] induces bias to the output node embeddings from a biased graph topology in Fig. 1c. In the input space, the node features are uniformly distributed. However, when the information propagation is performed on a biased topology as in Fig. 1b, the information received by nodes in different subgroups could be biased [41], leading to a biased embedding distribution in the output space.

There has been emerging research interest in fulfilling algorithmic fairness in graph mining. Nevertheless, the studied fairness notions vary across different works, which can be confusing and impede further progress. Meanwhile, different techniques are developed in achieving various fairness notions. Without a clear understanding of the corresponding mappings, future fair graph mining algorithm design can be difficult. Therefore, a systematic survey of

Our survey paper has been released on arxiv.

Related Materials of this Tutorial



This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2023.3265598

Fairness in Graph Mining: A Survey

Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li

—Graph mining algorithms have been playing a significant role in myriad fields over the years. However, despite their performance on various graph analytical tasks, most of these algorithms lack fairness considerations. As a consequence, lead to discrimination towards certain populations when exploited in human-centered applications. Recently, algorithmic as been extensively studied in graph-based applications. In contrast to algorithmic fairness on independent and identically (i.i.d.) data, fairness in graph mining has exclusive backgrounds, taxonomies, and fulfilling techniques. In this survey, we comprehensive and up-to-date introduction of existing literature under the context of fair graph mining. Specifically, we novel taxonomy of fairness notions on graphs, which sheds light on their connections and differences. We further present an summary of existing techniques that promote fairness in graph mining. Finally, we discuss current research challenges and stions, aiming at encouraging cross-breeding ideas and further advances.

ms—Algorithmic Fairness, Graph Mining, Debiasing

1 Introduction

Graph-structured data is pervasive in diverse real-world applications, e.g., E-commerce [102], [121], health care [37], [53], traffic forecasting [72], [100], and drug discovery [15], [172]. In recent years, a number of graph mining algorithms have been proposed to gain a deeper understanding of such data. These algorithms have shown promising performance on graph analytical tasks such as node classification [59], [86], [161] and link prediction [4], [103], [109], contributing to great advances in many graph-based applications.

Despite the success of these graph mining algorithms, most of them lack fairness considerations. Consequently, they could yield discriminatory results towards certain populations when such algorithms are exploited in human-centered applications [80]. For example, a social network-based job recommender system may unfavorably recommend fewer job opportunities to individuals of a certain gender [97] or individuals in an underrepresented ethnic group [150]. With the widespread usage of graph mining algorithms, such potential discrimination could also exist in other high-stake applications such as disaster response [159], criminal justice [3], and loan approval [136]. In these applications, critical and life-changing decisions are often made for the individuals involved. Therefore, how to tackle unfairness issues in graph mining algorithms naturally becomes a crucial problem.

- *Y. Dong is with Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, Virginia, US. E-mail: yd6eb@virginia.edu*
- *J. Ma is with Department of Computer Science, University of Virginia, Charlottesville, Virginia, US. E-mail: jm3mr@virginia.edu*
- *S. Wang is with Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, Virginia, US. E-mail: sw3wv@virginia.edu*
- *C. Chen is with Biocomplexity Institute, University of Virginia, Charlottesville, Virginia, US. E-mail: zrh6du@virginia.edu*
- *J. Li is with Department of Electrical and Computer Engineering, Department of Computer Science, and School of Data Science, University of Virginia, Charlottesville, Virginia, US. E-mail: jundong@virginia.edu*

Compared with achieving fairness in the context of independent and identically distributed (i.i.d.) data, fulfilling fairness in graph mining can be non-trivial due to two main challenges. The first challenge is to formulate proper fairness notions as the criteria to determine the existence of unfairness (i.e., bias). Although a vast amount of traditional algorithmic fairness notions have been proposed centered on i.i.d. data [42], [111], they are unable to reflect the bias exhibited by the relational information (i.e., the topology) in graph data. For example, the same population can be connected with different topologies as in Fig. 1a and 1b, where each node represents an individual, and the color of nodes denotes their demographic subgroup membership, such as different genders. Compared with the graph topology in Fig. 1a, the topology in Fig. 1b has more intra-group edges than inter-group edges. The dominance of intra-group edges in the graph topology is a common type of bias existing in real-world graphs [39], [41], [70], which cannot be captured by traditional algorithmic fairness notions. The second challenge is to prevent the graph mining algorithms from inheriting the bias exhibited in the input relational information [41], [112], [148], [160]. We present a toy example to demonstrate how the information propagation mechanism in Graph Neural Networks (GNNs) [64], [86], [161] induces bias to the output node embeddings from a biased graph topology in Fig. 1c. In the input space, the node features are uniformly distributed. However, when the information propagation is performed on a biased topology as in Fig. 1b, the information received by nodes in different subgroups could be biased [41], leading to a biased embedding distribution in the output space.

There has been emerging research interest in fulfilling algorithmic fairness in graph mining. Nevertheless, the studied fairness notions vary across different works, which can be confusing and impede further progress. Meanwhile, different techniques are developed in achieving various fairness notions. Without a clear understanding of the corresponding mappings, future fair graph mining algorithm design can be difficult. Therefore, a systematic survey of

Our survey paper has been released on arxiv.

PyGDebias

PyGDebias: 10+ popular algorithms and 20+ graph datasets.

Collected Algorithms

13 different methods in total are implemented in this library. We provide an overview of their characteristics as follows.

Methods	Debiasing Technique	Fairness Notions	Paper & Code
FairGNN [2]	Adversarial Learning	Group Fairness	[Paper] [Code]
EDITS [3]	Edge Rewiring	Group Fairness	[Paper] [Code]
FairWalk [4]	Rebalancing	Group Fairness	[Paper] [Code]
CrossWalk [5]	Rebalancing	Group Fairness	[Paper] [Code]
UGE [6]	Edge Rewiring	Group Fairness	[Paper] [Code]
FairVGNN [7]	Adversarial Learning	Group Fairness	[Paper] [Code]
FairEdit [8]	Edge Rewiring	Group Fairness	[Paper] [Code]
NIFTY [9]	Optimization with Regularization	Group/Counterfactual Fairness	[Paper] [Code]
GEAR [10]	Edge Rewiring	Group/Counterfactual Fairness	[Paper] [Code]
InFoRM [11]	Optimization with Regularization	Individual Fairness	[Paper] [Code]
REDRESS [12]	Optimization with Regularization	Individual Fairness	[Paper] [Code]
GUIDE [13]	Optimization with Regularization	Individual Fairness	[Paper] [Code]
RawlsGCN [14]	Rebalancing	Degree-Related Fairness	[Paper] [Code]

Outline

Outline

Background Introduction

Fairness Notions and Metrics

Theoretical Understanding of Bias

Techniques for Fair Node Embeddings

Real-World Applications

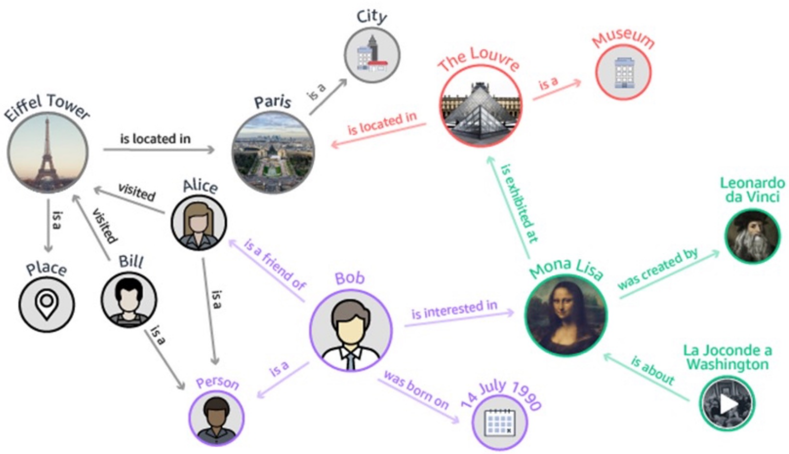
Summary, Challenges, & Future Directions

Graph Machine Learning Algorithms

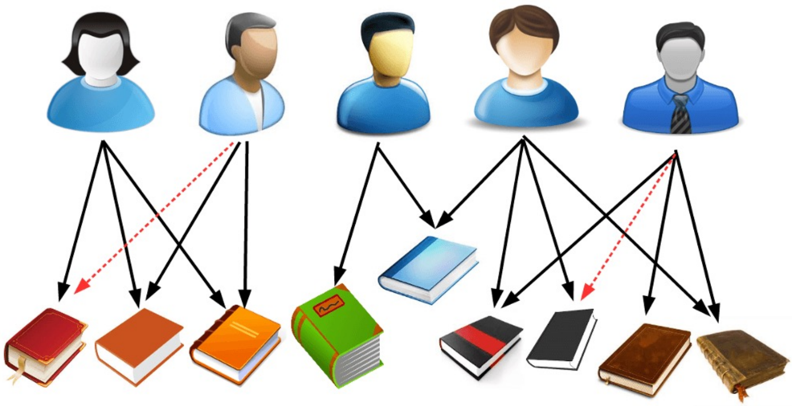
What are graph machine learning (ML) algorithms?

Social networks

Financial Networks



Knowledge Graphs

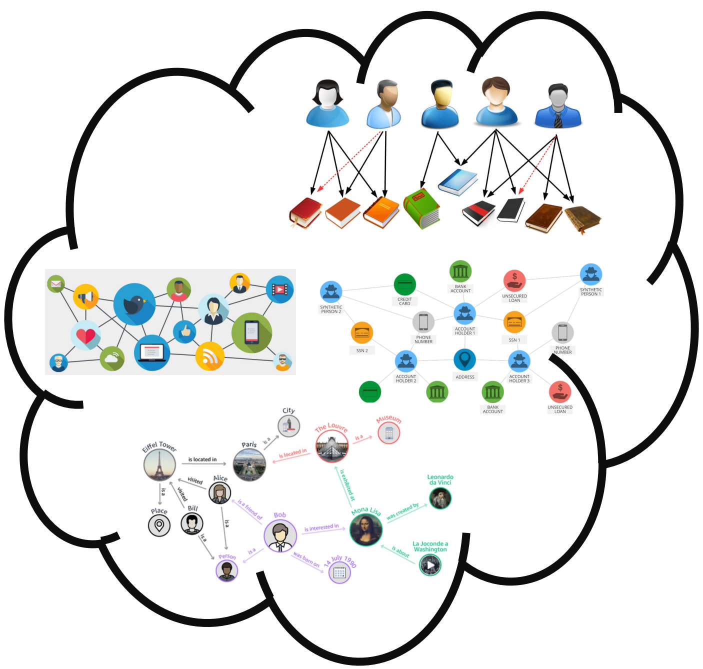


E-Commerce Networks

Graph Machine Learning Algorithms (Cont.)

What are graph machine learning (ML) algorithms?

In general, graph machine learning algorithms **extract information encoded in the graph data** to facilitate our understanding (on these graphs) and gain benefit on various predictive tasks.



Graph Mining Algorithms

- Who are potential friends?
- Which item will this customer buy?
- Which loan applicant is with the lowest risk of debt default?

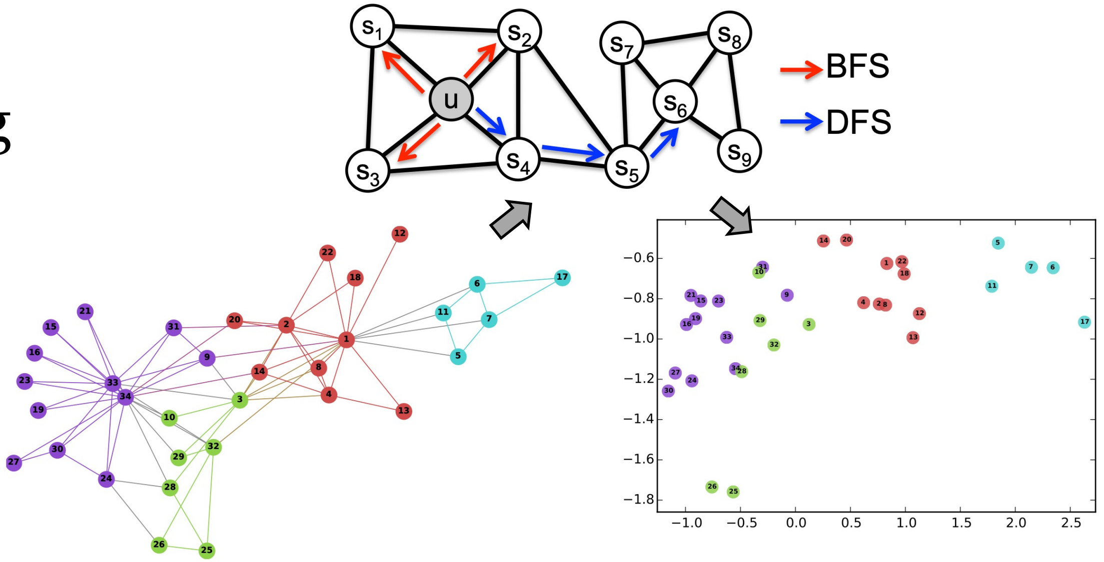
…

Popular Graph ML Algorithms

Shallow graph embedding methods.

Learning node embeddings that preserve the structural proximity.

Typical examples: Deepwalk, Node2Vec, etc.,

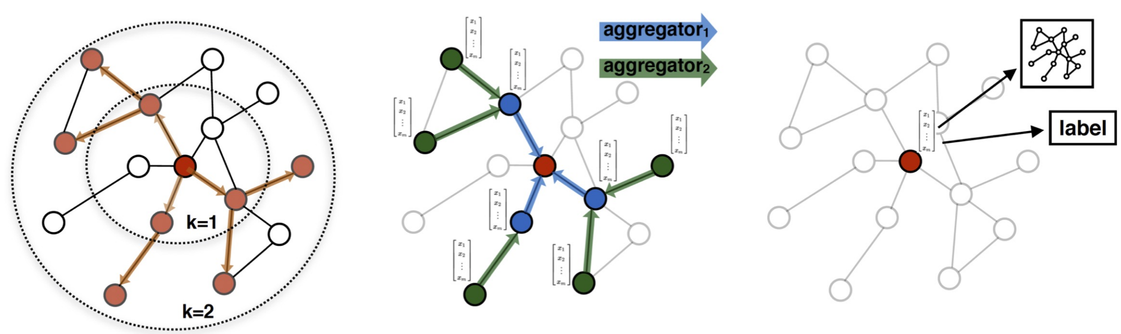


(a) Input: Karate Graph (b) Output: Representation

Graph Neural Networks (GNNs).

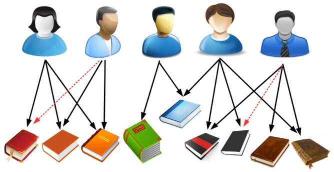
Encoding the node attribute and structure information into the learned node embeddings.

Typical examples: Graph Convolutional Networks (GCNs), GraphSAGE, etc.

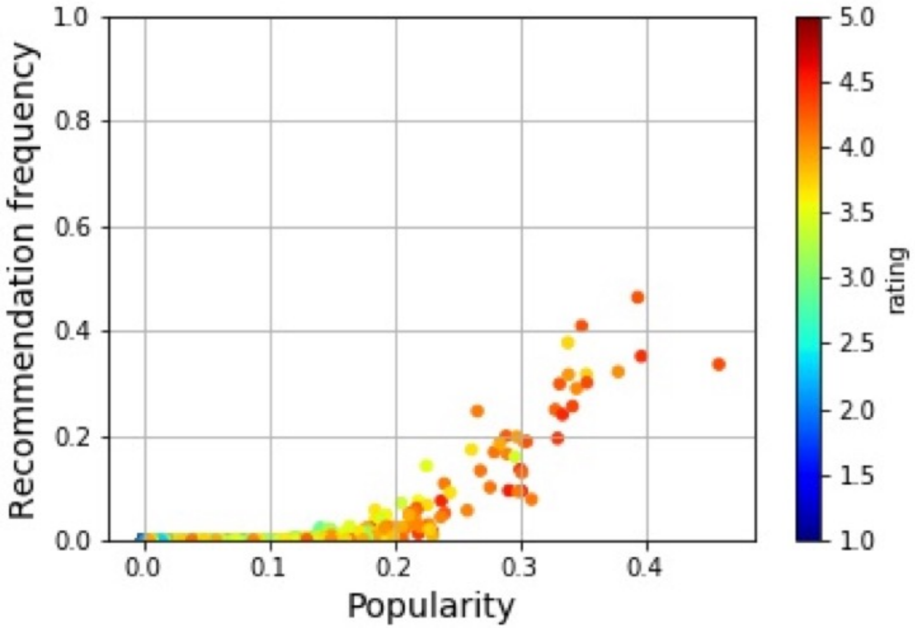


1. Sample neighborhood 2. Aggregate feature information from neighbors 3. Predict graph context and label using aggregated information

The Risk of Bias in Graph ML



Potential discrimination in **recommender systems**.

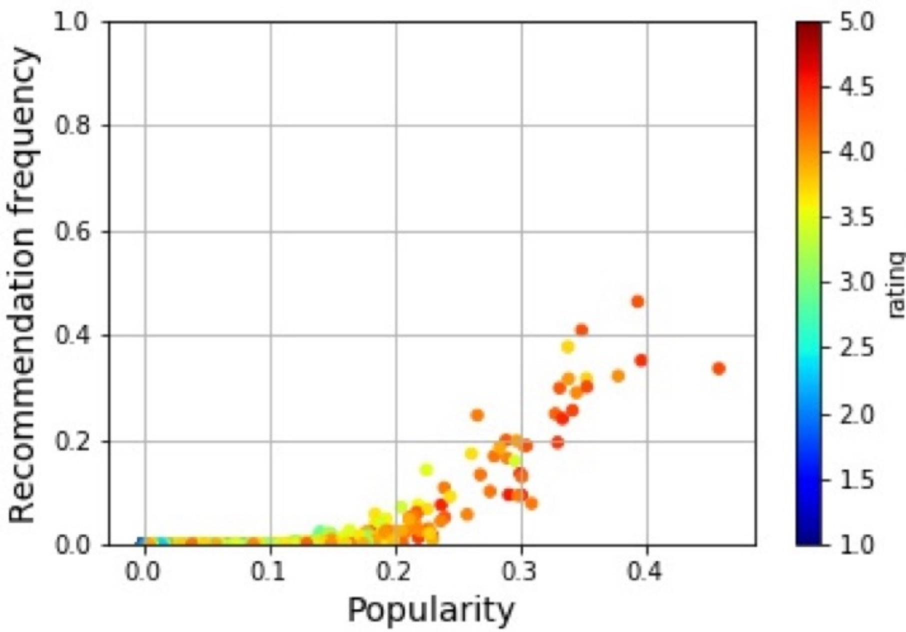


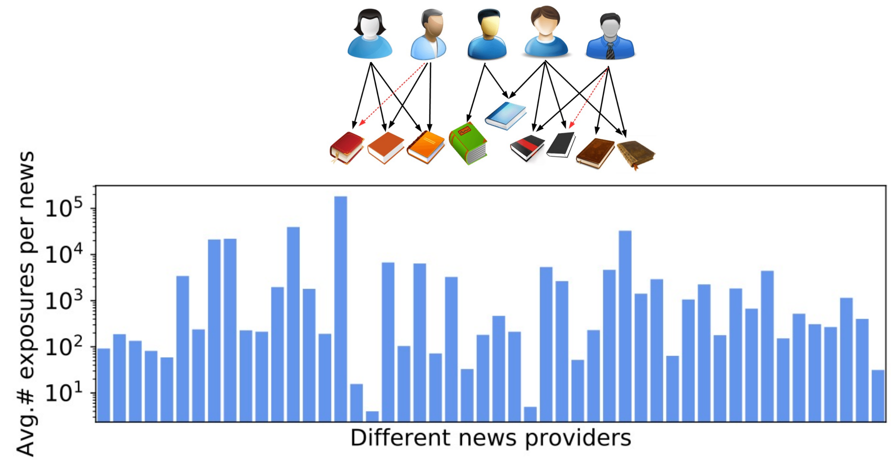
Popular items are often over-emphasized in recommendations, while less popular ones get less exposure [1].

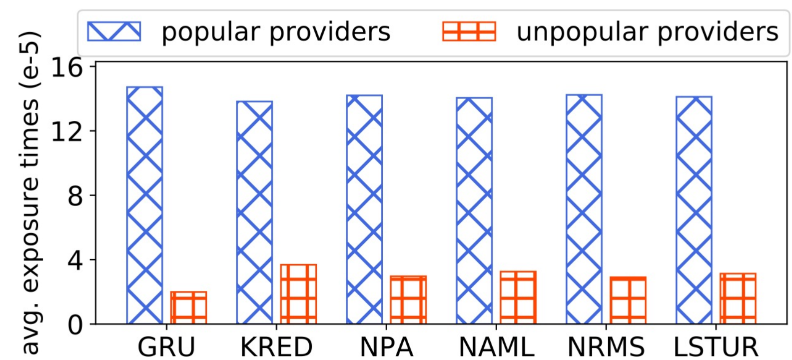
[1] Abdollahpouri, Himan, et al. "The impact of popularity bias on fairness and calibration in recommendation." arXiv preprint arXiv:1910.05755 (2019).

The Risk of Bias in Graph ML

Potential discrimination in **recommender systems**.







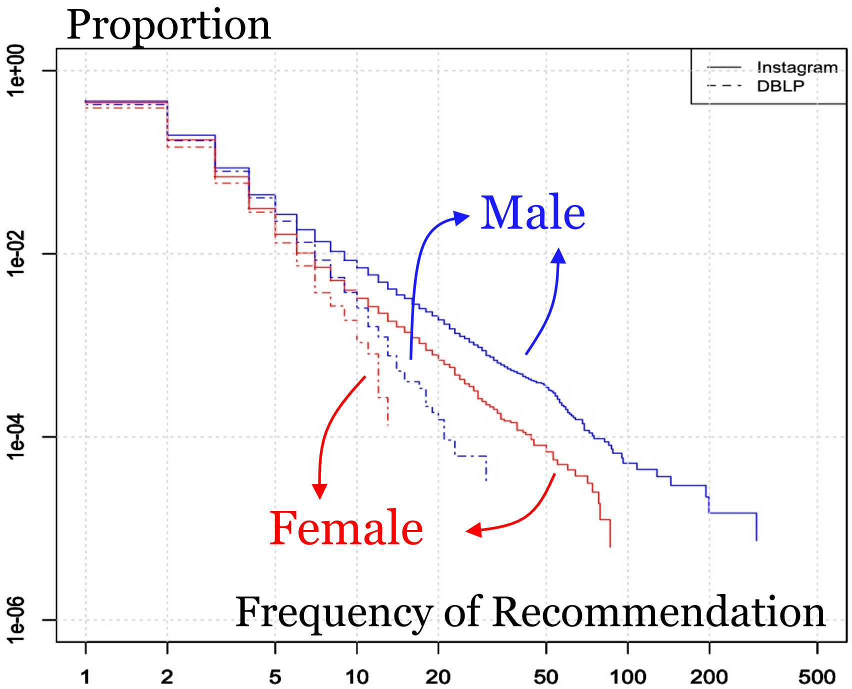
Popular items are often over-emphasized in recommendations, while less popular ones get less exposure [1].

Unpopular providers always bear much less exposure rates across different recommendation models [2].

[1] Abdollahpouri, Himan, et al. "The impact of popularity bias on fairness and calibration in recommendation." arXiv preprint arXiv:1910.05755 (2019).
[2] Qi, Tao, et al. "Profairrec: Provider fairness-aware news recommendation." In SIGIR 2022.

The Risk of Bias in Graph ML (Cont.)

Potential discrimination in **social networks**.

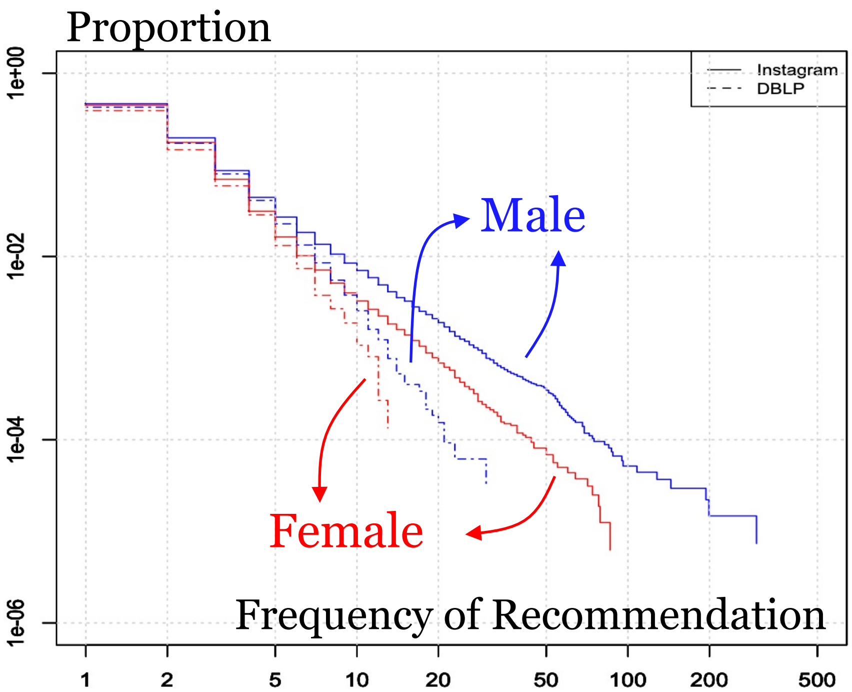


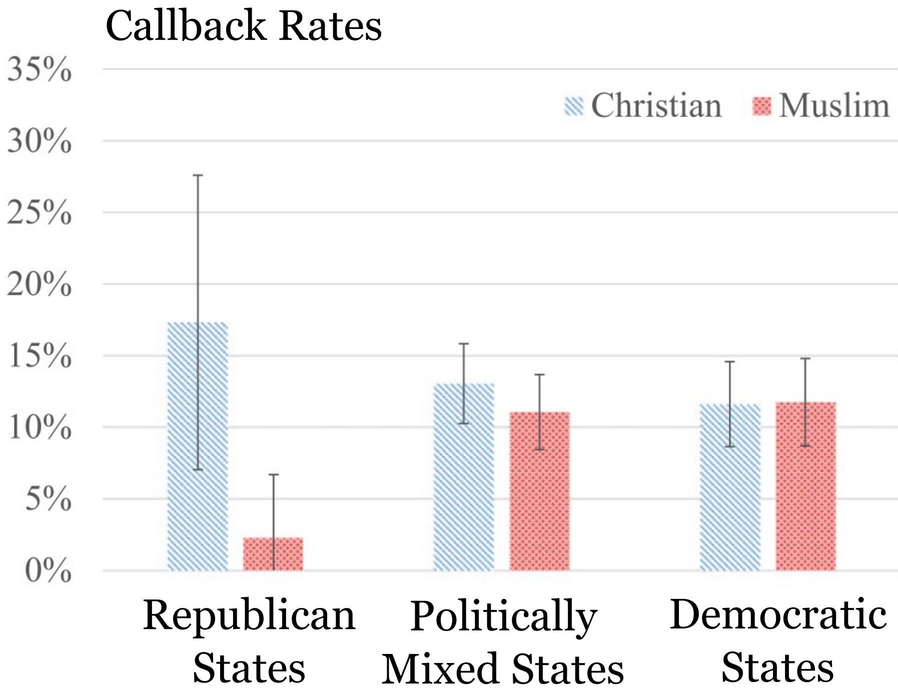
Users who get recommended to be connected exhibit divergence between males and females [1].

[1] Stoica, Ana-Andreea, et al. "Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity." In WWW 2018.

The Risk of Bias in Graph ML (Cont.)

Potential discrimination in **social networks**.





Users who get recommended to be connected exhibit divergence between males and females [1].

Users' religion could also be a source of hiring discrimination in social networks [2].

[1] Stoica, Ana-Andreea, et al. "Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity." In WWW 2018.
[2] Acquisti, Alessandro, et al. "An experiment in hiring discrimination via online social networks." Management Science 66.3 (2020): 1005-1024.

Algorithmic Fairness

Then how to define fairness?

Algorithmic Fairness

Then how to define fairness?

Fairness can be defined in different ways [1]: different real-world applications show biases from various perspectives [2].

[1] Du, Mengnan, et al. "Fairness in deep learning: A computational perspective." IEEE Intelligent Systems 36.4 (2020): 25-34.
[2] Dong, Yushun, et al. "Fairness in graph mining: A survey." IEEE Transactions on Knowledge and Data Engineering (2023).

Algorithmic Fairness

Then how to define fairness?

Fairness can be defined in different ways [1]: different real-world applications show biases from various perspectives [2].

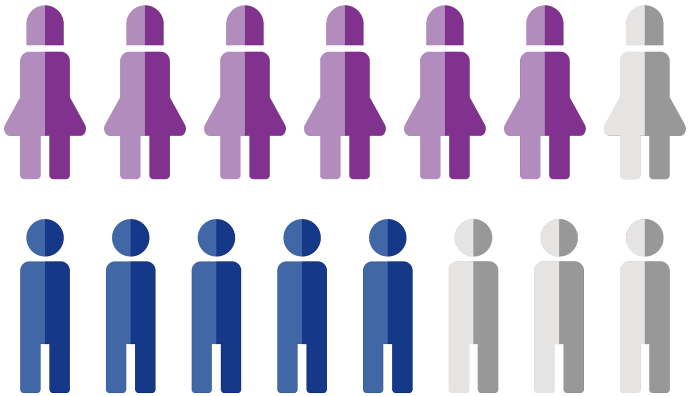
For example, it **depends on the specific studied problem** to determine which case should be considered as fair.

[1] Du, Mengnan, et al. "Fairness in deep learning: A computational perspective." IEEE Intelligent Systems 36.4 (2020): 25-34.
[2] Dong, Yushun, et al. "Fairness in graph mining: A survey." IEEE Transactions on Knowledge and Data Engineering (2023).

Algorithmic Fairness (Cont.)

Then how to define fairness?

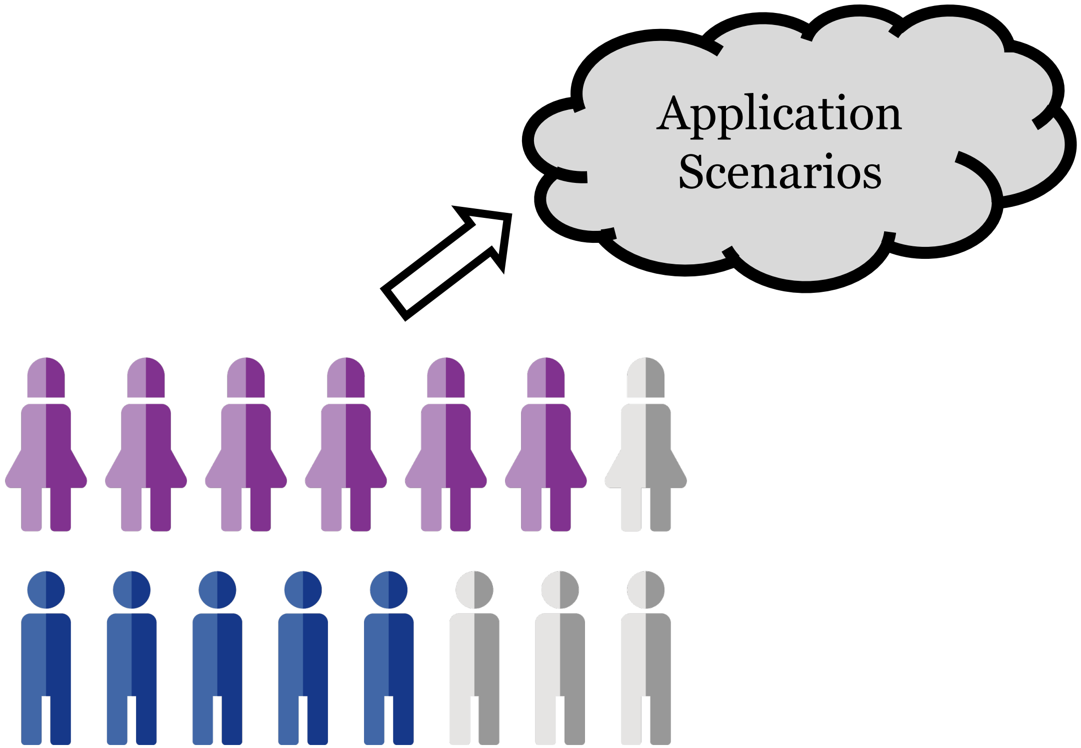
Despite the lack of a **universal criterion** for fairness, we could still study fairness in algorithms: there are **various existing fairness notions** based on people's awareness.



Algorithmic Fairness (Cont.)

Then how to define fairness?

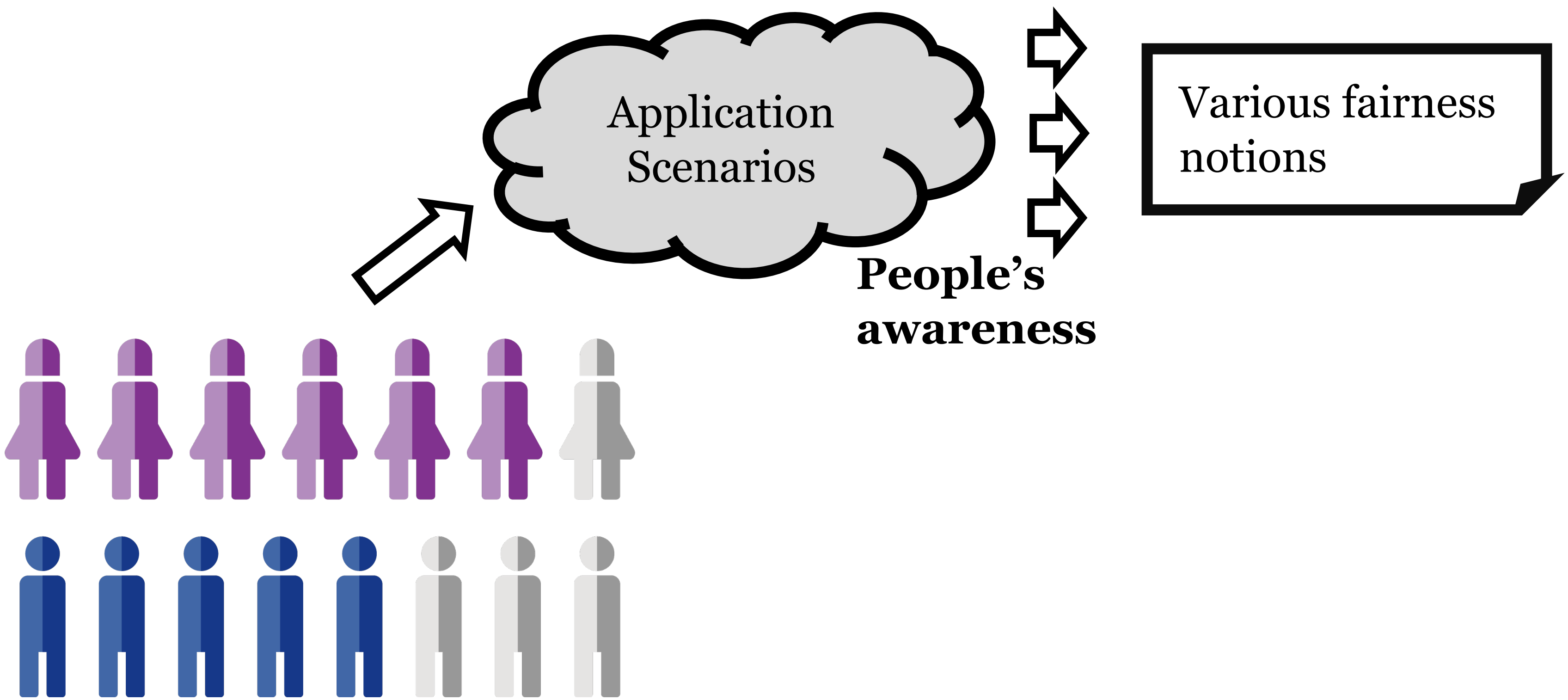
Despite the lack of a **universal criterion** for fairness, we could still study fairness in algorithms: there are **various existing fairness notions** based on people's awareness.



Algorithmic Fairness (Cont.)

Then how to define fairness?

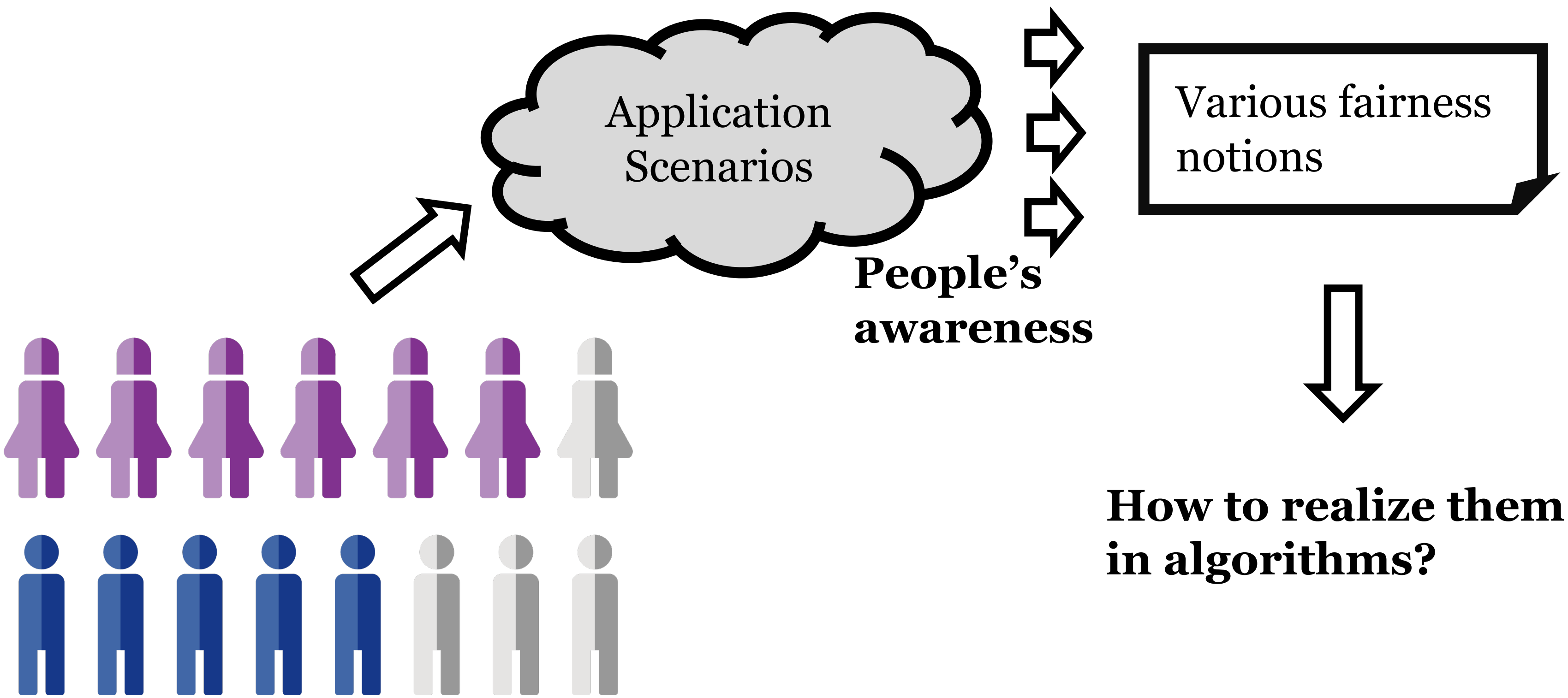
Despite the lack of a **universal criterion** for fairness, we could still study fairness in algorithms: there are **various existing fairness notions** based on people's awareness.



Algorithmic Fairness (Cont.)

Then how to define fairness?

Despite the lack of a **universal criterion** for fairness, we could still study fairness in algorithms: there are **various existing fairness notions** based on people’s awareness.



Fairness in Graph ML Algorithms

Then how to define fairness?

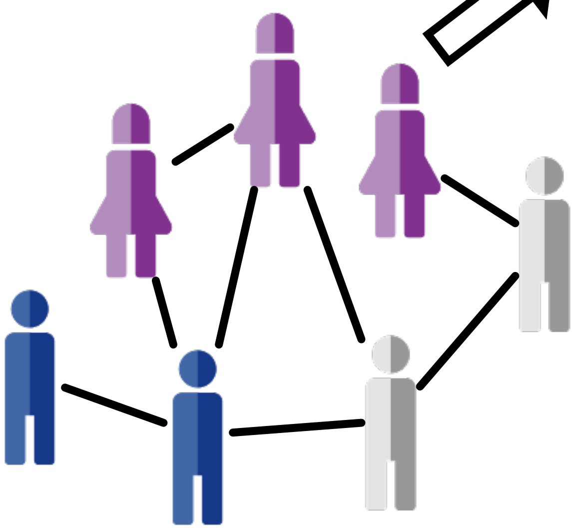
In the realm of **graph machine learning…**

Graph-based application scenarios

People's awareness

★ Fairness notions **on graphs**

★ **How to realize them in graph machine learning algorithms?**



Fulfilling Fairness in Graph ML Algorithms

Unique Challenges of fulfilling fairness in graph ML algorithms.

Fulfilling Fairness in Graph ML Algorithms

Unique Challenges of fulfilling fairness in graph ML algorithms.

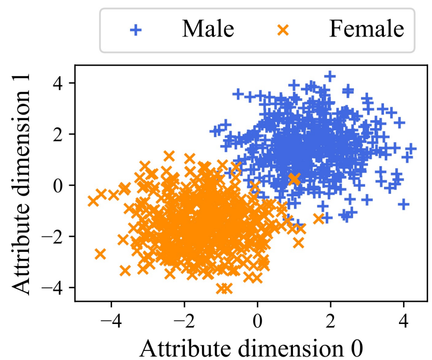
(1) Formulating **proper fairness notions** as the criteria to determine the existence of unfairness (i.e., bias).

Fulfilling Fairness in Graph ML Algorithms

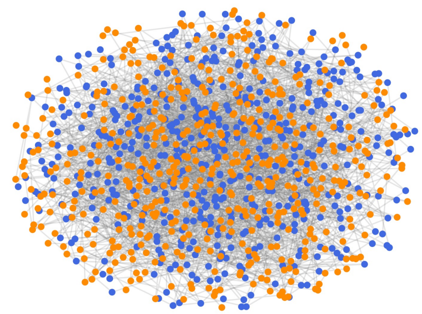
Unique Challenges of fulfilling fairness in graph ML algorithms.

(1) Formulating **proper fairness notions** as the criteria to determine the existence of unfairness (i.e., bias).

Attributed Graph 1

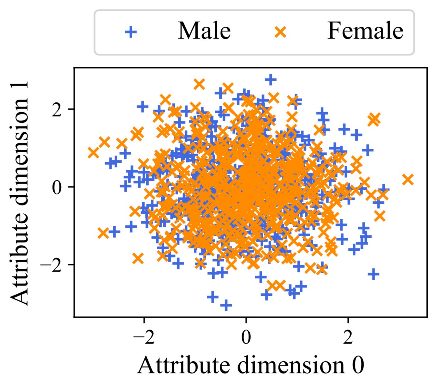


+

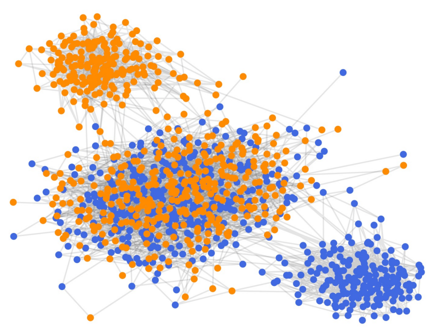


Fair or biased?

Attributed Graph 2



+



Fair or biased?

Fulfilling Fairness in Graph ML Algorithms

Unique Challenges of fulfilling fairness in graph ML algorithms.

(1) Formulating **proper fairness notions** as the criteria to determine the existence of unfairness (i.e., bias).

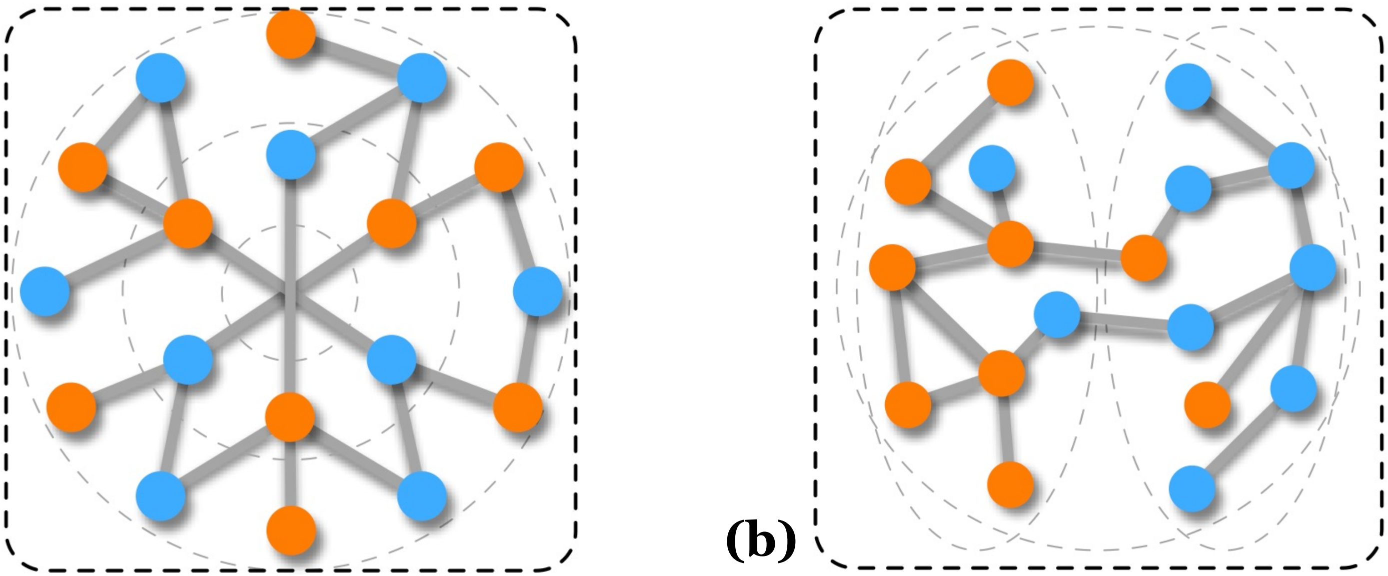
(2) Preventing the graph ML algorithms from **inheriting the bias** exhibited in the input graphs.

Fulfilling Fairness in Graph ML Algorithms

Unique Challenges of fulfilling fairness in graph ML algorithms.

(1) Formulating **proper fairness notions** as the criteria to determine the existence of unfairness (i.e., bias).

(2) Preventing the graph ML algorithms from **inheriting the bias** exhibited in the input graphs.



(a) (b)

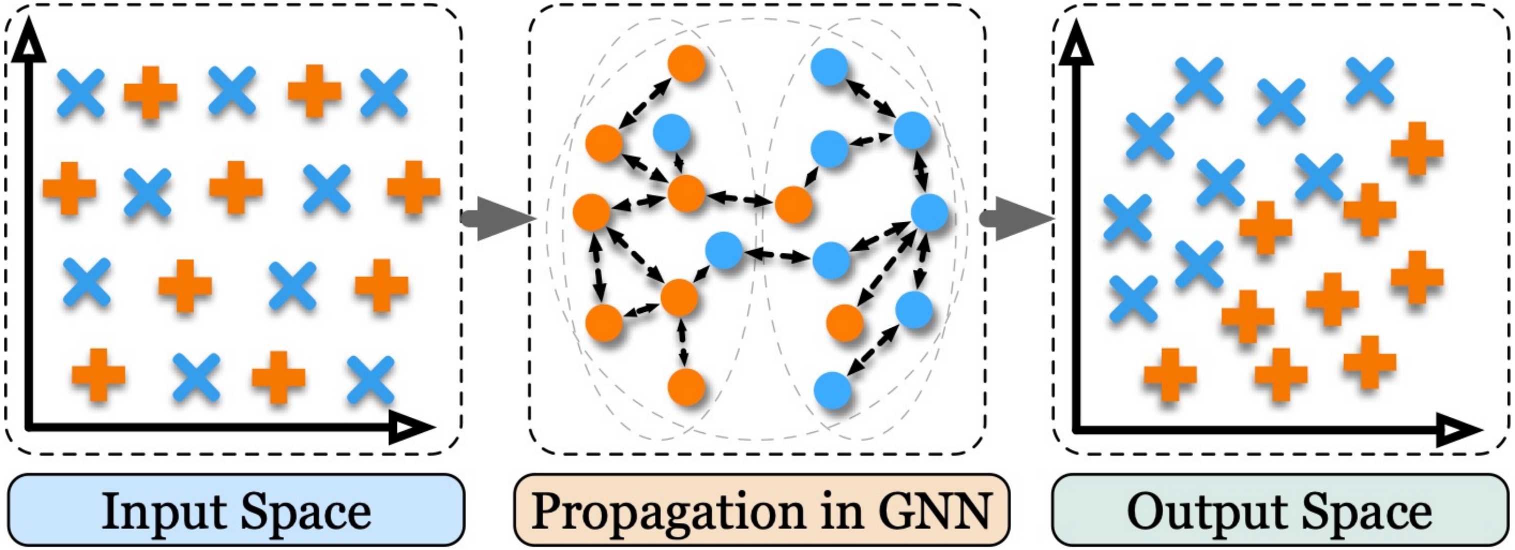
Compared with the structure in (a), the bias in the graph structure of (b) could lead to biased embeddings in Graph Neural Networks (GNNs).

Fulfilling Fairness in Graph ML Algorithms

Unique Challenges of fulfilling fairness in graph ML algorithms.

(1) Formulating **proper fairness notions** as the criteria to determine the existence of unfairness (i.e., bias).

(2) Preventing the graph ML algorithms from **inheriting the bias** exhibited in the input graphs.



An example in Graph Neural Networks (GNNs): the unbalance between intra-group and inter-group edges could easily induce bias in the outcome space [1].

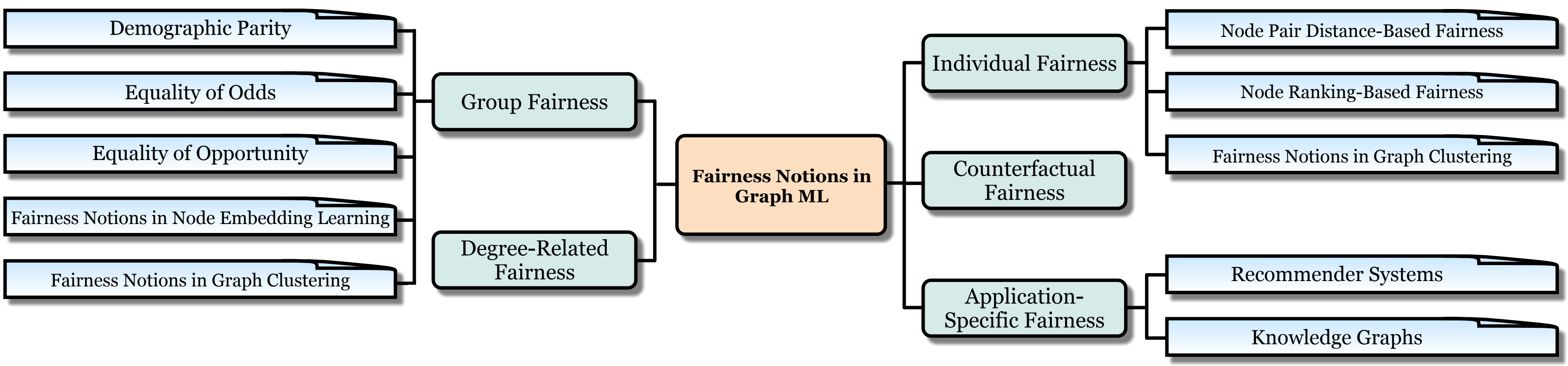
[1] Dong, Yushun, et al. "Fairness in graph mining: A survey." IEEE Transactions on Knowledge and Data Engineering (2023).

Outline

- Background Introduction
- Fairness Notions and Metrics
- Theoretical Understanding of Bias
- Techniques for Fair Node Embeddings
- Real-World Applications
- Summary, Challenges, & Future Directions

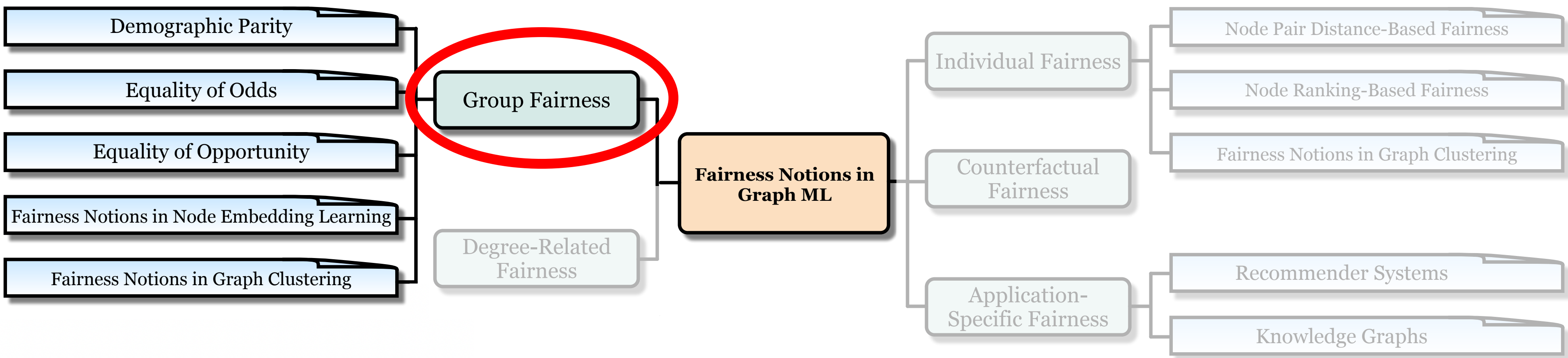
Taxonomy of Fairness Notions

A taxonomy of commonly used algorithmic fairness notions in graph ML.



Taxonomy of Fairness Notions (Cont.)

A taxonomy of commonly used algorithmic fairness notions in graph ML.



A general idea of group fairness: categorical **sensitive attributes** (e.g., gender, race) divide the whole population into different sensitive subgroups, and each group should gain **their fair share of interest** [1].

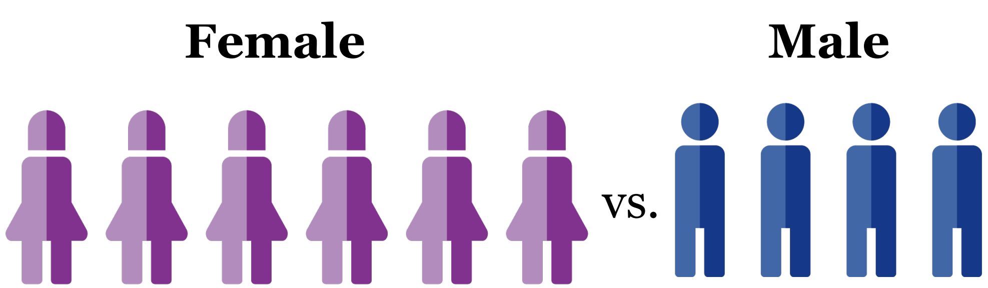
[1] Dwork, Cynthia, et al. "Fairness through awareness." In ITCS 2012.

Demographic Parity

Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.



[1] Dwork, Cynthia, et al. "Fairness through awareness." In ITCS 2012.

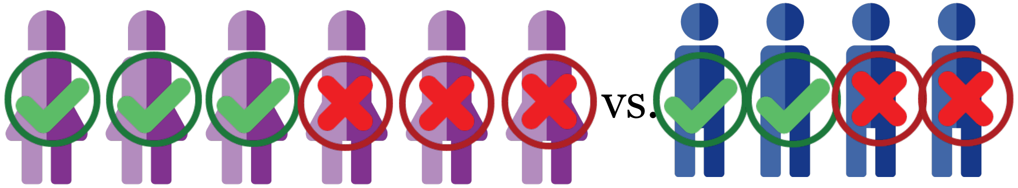
Demographic Parity

Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.

Female **Male**



[1] Dwork, Cynthia, et al. "Fairness through awareness." In ITCS 2012.

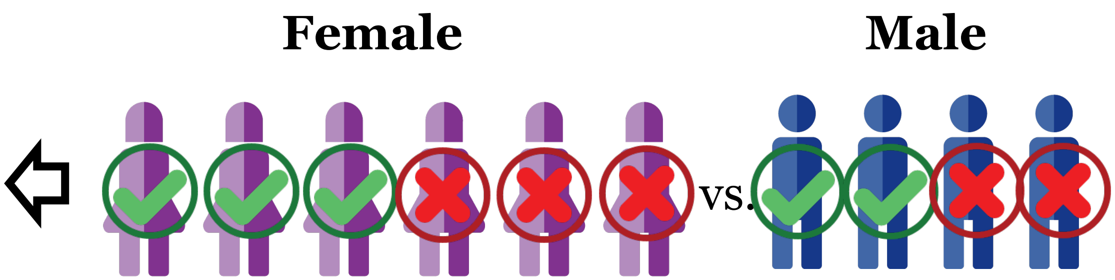
Demographic Parity

Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.

Fair in perspective of Demographic Parity.



[1] Dwork, Cynthia, et al. "Fairness through awareness." In ITCS 2012.

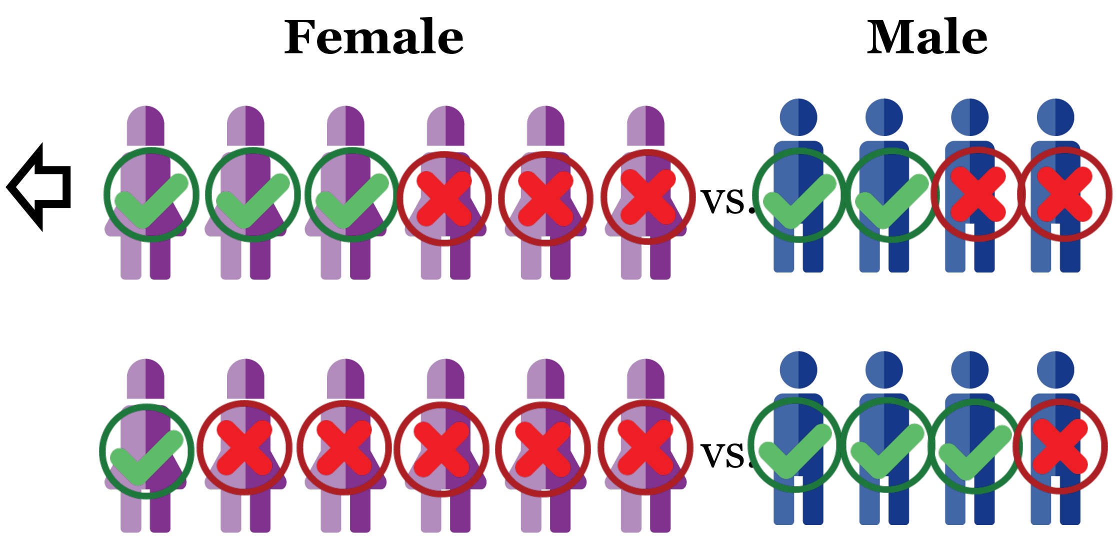
Demographic Parity

Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.

Fair in perspective of Demographic Parity.



[1] Dwork, Cynthia, et al. "Fairness through awareness." In ITCS 2012.

Demographic Parity

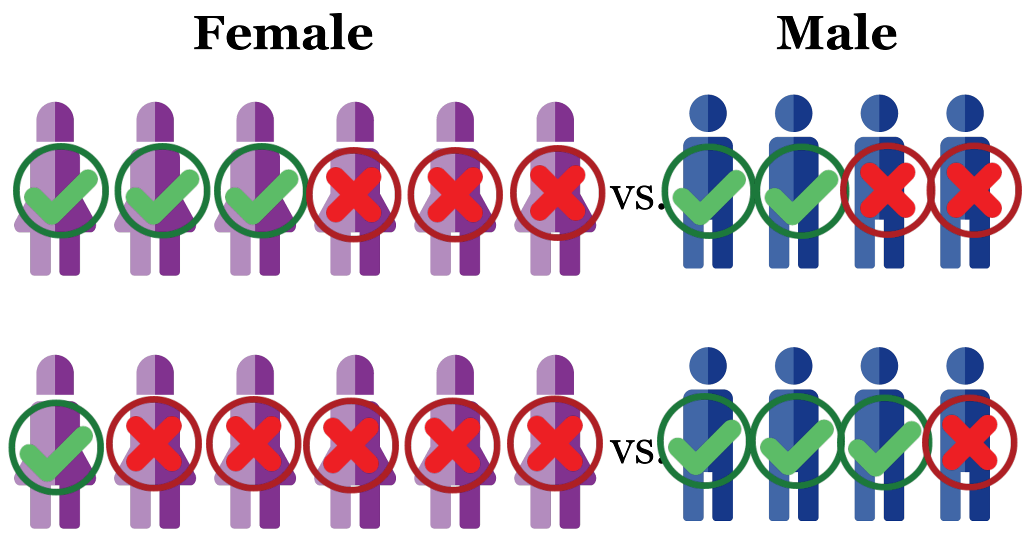
Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.

Fair in perspective of Demographic Parity.

Unfair in perspective of Demographic Parity.



[1] Dwork, Cynthia, et al. "Fairness through awareness." In ITCS 2012.

Demographic Parity (Cont.)

Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.

Criterion: $$\mathrm{P}\left(\hat{Y}=1 \middle| S=0\right)= \mathrm{P}\left(\hat{Y}=1 \middle| S=1\right)$$

Metric: $$\Delta_{DP}=\left|\mathrm{P}\left(\hat{Y}=1 \middle| S=0\right)- \mathrm{P}\left(\hat{Y}=1 \middle| S=1\right)\right|$$

[1] Dwork, Cynthia, et al. "Fairness through awareness." In ITCS 2012.

Demographic Parity (Cont.)

Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.

Criterion: $$\mathrm{P}(\hat{Y} = 1 | S = 0) = \mathrm{P}(\hat{Y} = 1 | S = 1)$$

Metric: $$\Delta_{DP} = |\mathrm{P}(\hat{Y} = 1 | S = 0) - \mathrm{P}(\hat{Y} = 1 | S = 1)|$$

Recent works on fairness for graph ML algorithms have **extended this notion to other settings**, including link prediction [2, 3] and scenarios with continuous sensitive feature(s) values [4];

[1] Dwork, Cynthia, et al. "Fairness through awareness." In ITCS 2012.
[2] Acquisti, Alessandro, et al. "An experiment in hiring discrimination via online social networks." Management Science 66.3 (2020): 1005-1024.
[3] Du, Mengnan, et al. "Fairness in deep learning: A computational perspective." IEEE Intelligent Systems 36.4 (2020): 25-34.
[4] Dong, Yushun, et al. "Fairness in graph mining: A survey." IEEE Transactions on Knowledge and Data Engineering (2023).

Equality of Odds/Opportunity

Group Fairness:

Equality of Odds [1] vs. **Equality of Opportunity** [1]

[1] Hardt, Moritz, et al. "Equality of opportunity in supervised learning." In NeurIPS, 2016.

Equality of Odds/Opportunity

Group Fairness:

Equality of Odds [1] vs. **Equality of Opportunity** [1]

Equality of Odds: the **positive rates** are enforced to be the same between sensitive subgroups conditional on the **ground truth class labels**.

[1] Hardt, Moritz, et al. "Equality of opportunity in supervised learning." In NeurIPS, 2016.

Equality of Odds/Opportunity

Group Fairness:

Equality of Odds [1] vs. **Equality of Opportunity** [1]

Equality of Odds: the **positive rates** are enforced to be the same between sensitive subgroups conditional on the **ground truth class labels**.

Criterion: $\mathrm{P}(\hat{Y} = 1 | S = 0, Y = y) = \mathrm{P}(\hat{Y} = 1 | S = 1, Y = y)$

Metric: $\Delta_{EOD} = |P(\hat{Y} = 1 | S = 0, Y = 1) - P(\hat{Y} = 1 | S = 1, Y = 1)| + |P(\hat{Y} = 1 | S = 0, Y = 0) - P(\hat{Y} = 1 | S = 1, Y = 0)|$

[1] Hardt, Moritz, et al. "Equality of opportunity in supervised learning." In NeurIPS, 2016.

Equality of Odds/Opportunity

Group Fairness:

Equality of Odds [1] vs. **Equality of Opportunity** [1]

The intuition of Equality of Odds: to enforce the true positive rate (right and positive results) and false positive rate (wrong but positive results) to be the same across sensitive subgroups;

[1] Hardt, Moritz, et al. "Equality of opportunity in supervised learning." In NeurIPS, 2016.

Equality of Odds/Opportunity

Group Fairness:

Equality of Odds [1] vs. **Equality of Opportunity** [1]

The intuition of Equality of Odds: to enforce the true positive rate (right and positive results) and false positive rate (wrong but positive results) to be the same across sensitive subgroups;

Equality of Opportunity: the **positive rates** are enforced to be the same between sensitive subgroups conditional on the **positive ground truth class labels**.

[1] Hardt, Moritz, et al. "Equality of opportunity in supervised learning." In NeurIPS, 2016.

Equality of Odds/Opportunity

Group Fairness:

Equality of Odds [1] vs. **Equality of Opportunity** [1]

The intuition of Equality of Odds: to enforce the true positive rate (right and positive results) and false positive rate (wrong but positive results) to be the same across sensitive subgroups;

Equality of Opportunity: the **positive rates** are enforced to be the same between sensitive subgroups conditional on the **positive ground truth class labels**.

Criterion: $\mathrm{P}\left(\hat{Y}=1 \middle| S=0, Y=1\right) = \mathrm{P}\left(\hat{Y}=1 \middle| S=1, Y=1\right)$

Metric: $\Delta_{EO} = |P\left(\hat{Y}=1 \middle| S=0, Y=1\right) - P(\hat{Y}=1|S=1, Y=1)|$

[1] Hardt, Moritz, et al. "Equality of opportunity in supervised learning." In NeurIPS, 2016.

Equality of Odds/Opportunity

Group Fairness:

Equality of Odds [1] vs. **Equality of Opportunity** [1]

The intuition of Equality of Odds: to enforce the true positive rate (right and positive results) and false positive rate (wrong but positive results) to be the same across sensitive subgroups;

The intuition of Equality of Opportunity: to enforce the true positive rate (right and positive results) to be the same across sensitive subgroups;

[1] Hardt, Moritz, et al. "Equality of opportunity in supervised learning." In NeurIPS, 2016.

Equality of Odds/Opportunity

Group Fairness:

Equality of Odds [1] vs. **Equality of Opportunity** [1]

The intuition of Equality of Odds: to enforce the true positive rate (right and positive results) and false positive rate (wrong but positive results) to be the same across sensitive subgroups;

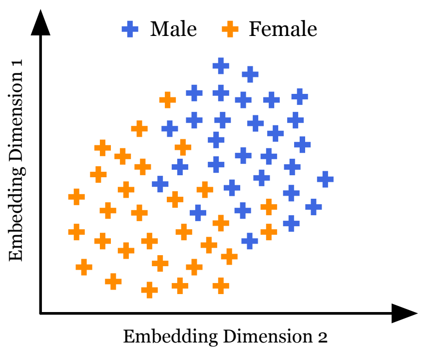
The intuition of Equality of Opportunity: to enforce the true positive rate (right and positive results) to be the same across sensitive subgroups;

Extension to tasks other than node classification, e.g., link prediction [1, 2].

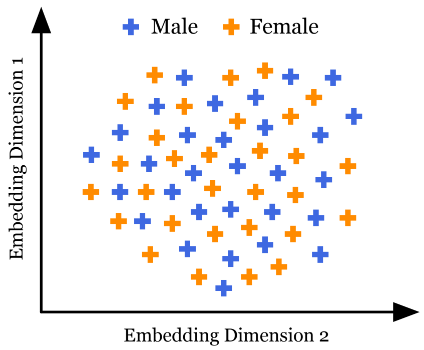
[1] Hardt, Moritz, et al. "Equality of opportunity in supervised learning." In NeurIPS, 2016.
[2] Acquisti, Alessandro, et al. "An experiment in hiring discrimination via online social networks." Management Science 66.3 (2020): 1005-1024.

Fairness in Node Embedding Learning

(1) Distribution-Based Fairness.



Unfair node embeddings



Fair node embeddings

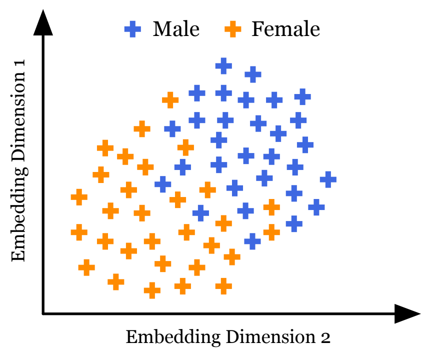
Criterion: Learned node embedding distributions across sensitive subgroups should be **similar**.

Metric: Measures of distance between distributions, e.g., Wasserstein distance [1, 2].

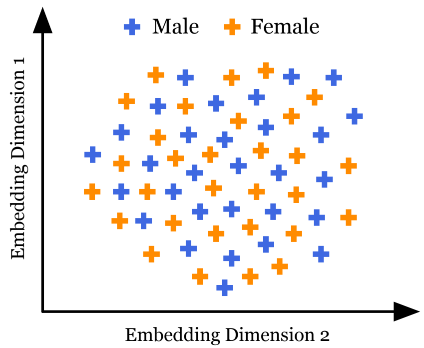
[1] Dong, Yushun, et al. "Edits: Modeling and mitigating data bias for graph neural networks." In WWW 2022.
[2] Fan, Wei, et al. "Fair graph auto-encoder for unbiased graph representations with Wasserstein distance." In ICDM 2021.

Fairness in Node Embedding Learning

(1) Distribution-Based Fairness.



Unfair node embeddings



Fair node embeddings

Criterion: Learned node embedding distributions across sensitive subgroups should be **similar**.

Metric: Measures of distance between distributions, e.g., Wasserstein distance [1, 2].

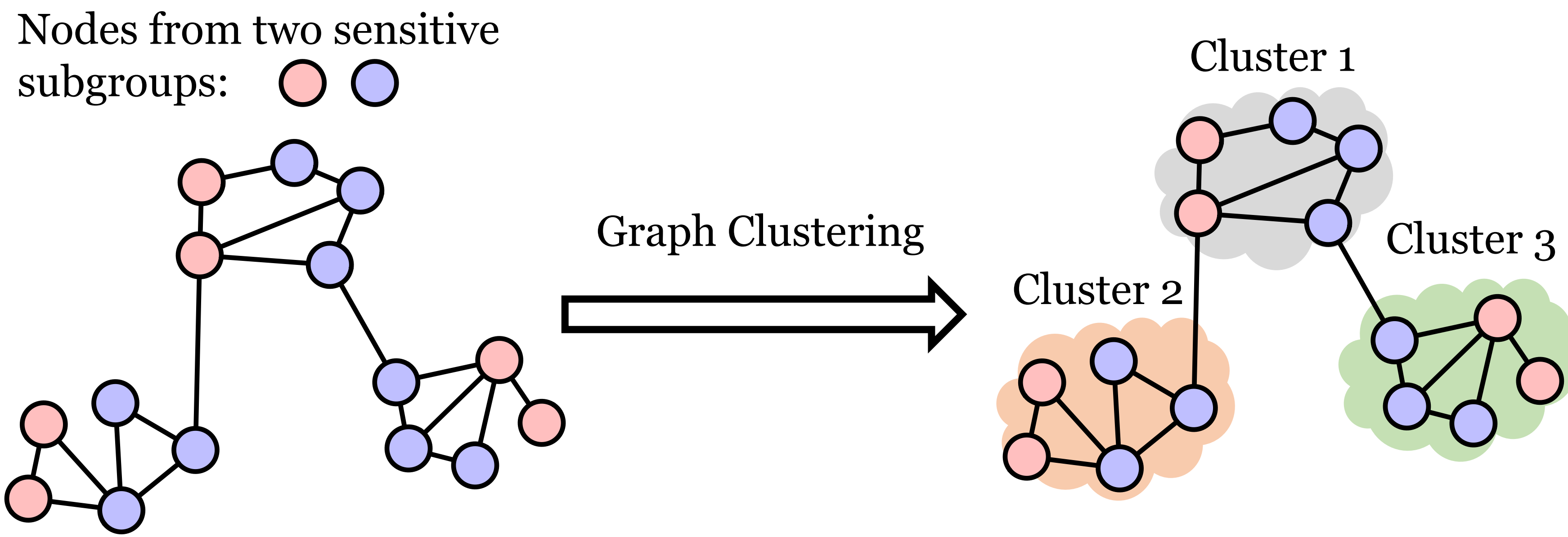
(2) Model-Based Fairness.

Criterion: There should be no information about sensitive attributes encoded in the learned node embeddings.

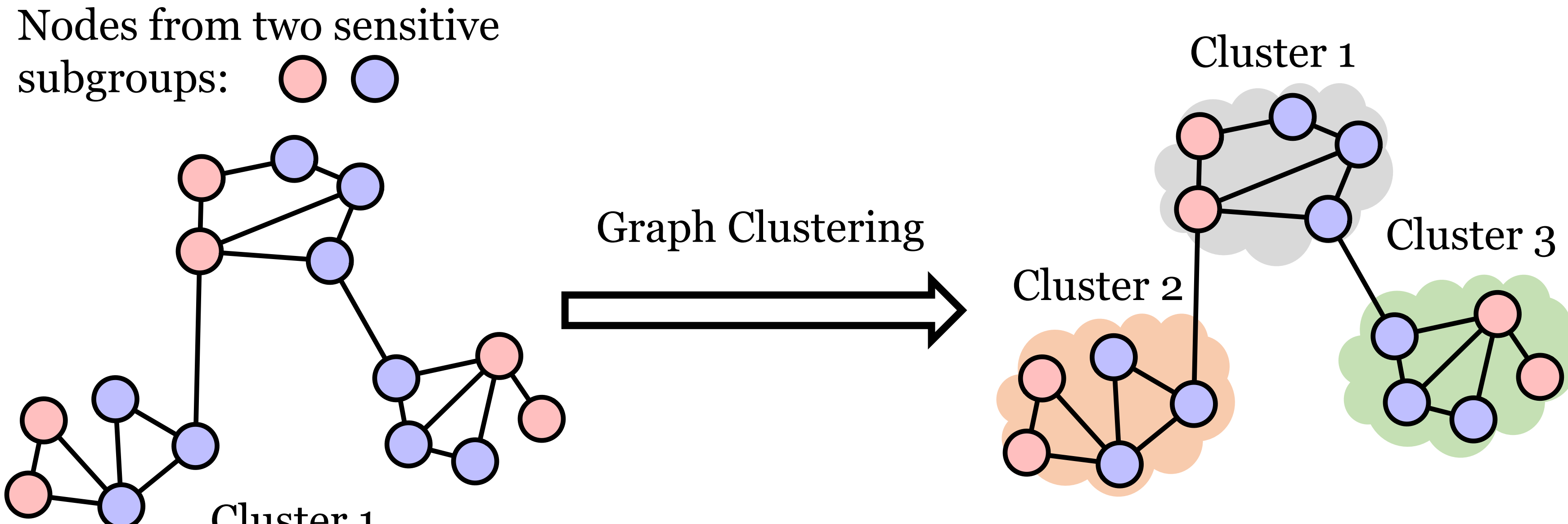
Metric: **Prediction accuracy** on the sensitive attributes with a predictive model (the lower, the better) [3].

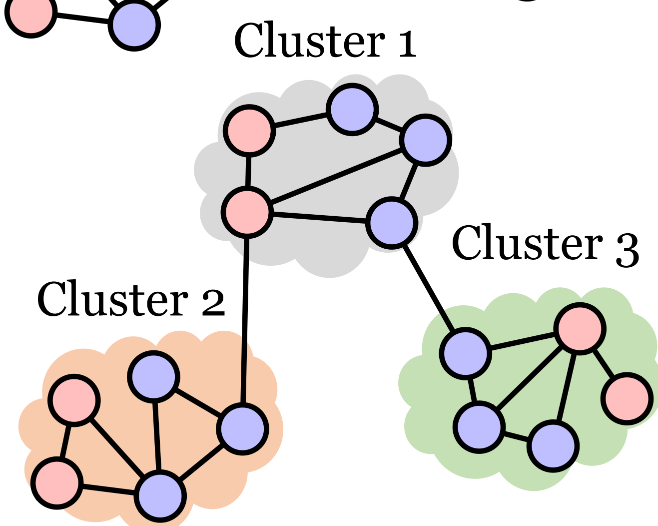
[1] Dong, Yushun, et al. "Edits: Modeling and mitigating data bias for graph neural networks." In WWW 2022.
[2] Fan, Wei, et al. "Fair graph auto-encoder for unbiased graph representations with Wasserstein distance." In ICDM 2021.
[3] Wu, Le, et al. "Learning fair representations for recommendation: A graph-based perspective." In WWW 2021.

Fairness in Graph Clustering



Fairness in Graph Clustering





Criterion: Nodes from different sensitive subgroups should be **proportionally represented** in each cluster [1].

Metric: *Balance score*, which measures fairness with the minimum ratio of node number [1].

$$Balance(C_k) = \min_{i \neq i', i, i' \in \{1, \ldots, H\}} \frac{|V_i \cap C_k|}{|V_{i'} \cap C_k|}$$

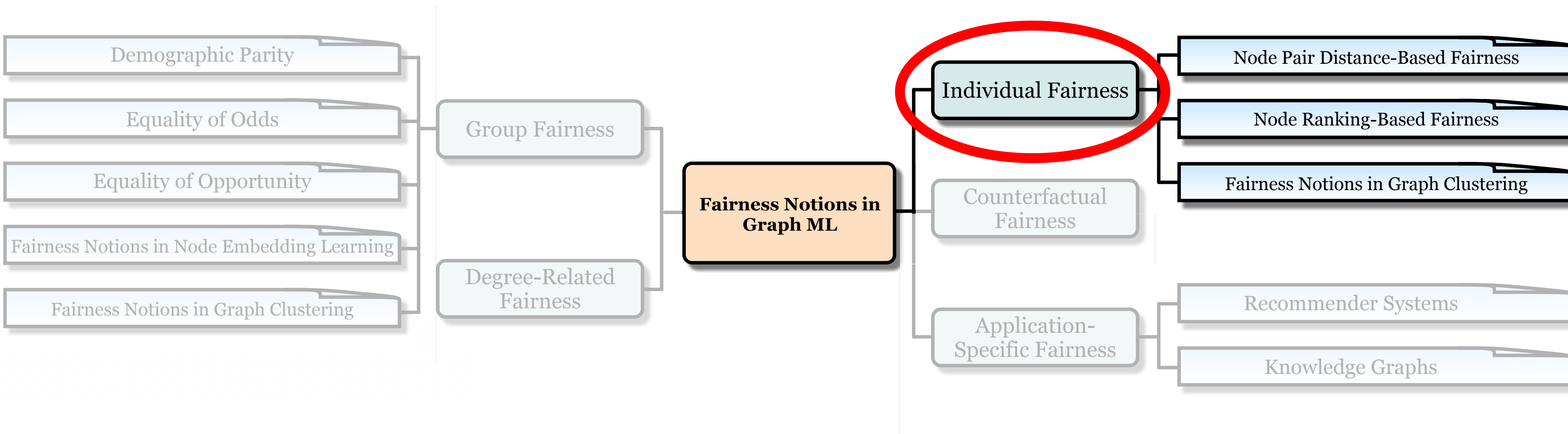
V_i: node set of sensitive subgroup i;

C_l: node set of cluster l;

[1] Kleindessner, Matthäus, et al. "Guarantees for spectral clustering with fairness constraints." In ICML 2019.

Taxonomy of Fairness Notions

Another critical fairness notion in graph ML: Individual Fairness.

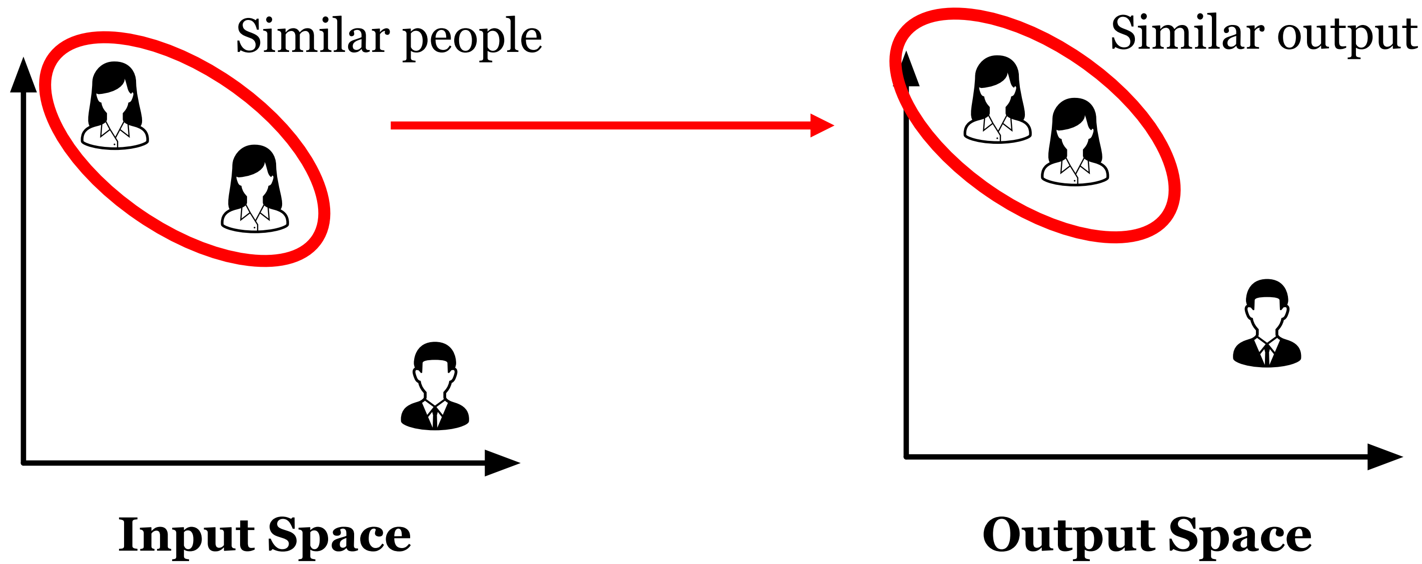


A general idea of individual fairness: **similar individuals should receive similar outputs** from the graph ML algorithms [1].

[1] Zeng, Ziqian, et al. Fair representation learning for heterogeneous information networks. In AAAI, 2021.

Node Pair Distance-Based Fairness

For any pair of node, this fairness notion enforces **the output distance to be smaller than a scaled input distance** - which is consistent with the general idea of "similar individual should receive similar output" [1].



[1] Kang, Jian, et al. Inform: Individual fairness on graph mining. In SIGKDD, 2020.

Node Pair Distance-Based Fairness

For any pair of node, this fairness notion enforces **the output distance to be smaller than a scaled input distance** - which is consistent with the general idea of "similar individual should receive similar output" [1].

Mathematically, we have

$$\underbrace{D_1(f(x), f(y))}_{\text{Output distance}} \leq L\underbrace{D_2(x, y)}_{\text{Input distance}} \quad \forall\,(x, y)$$

L: Lipschitz Constant

In practice, individual fairness enforces the following inequality

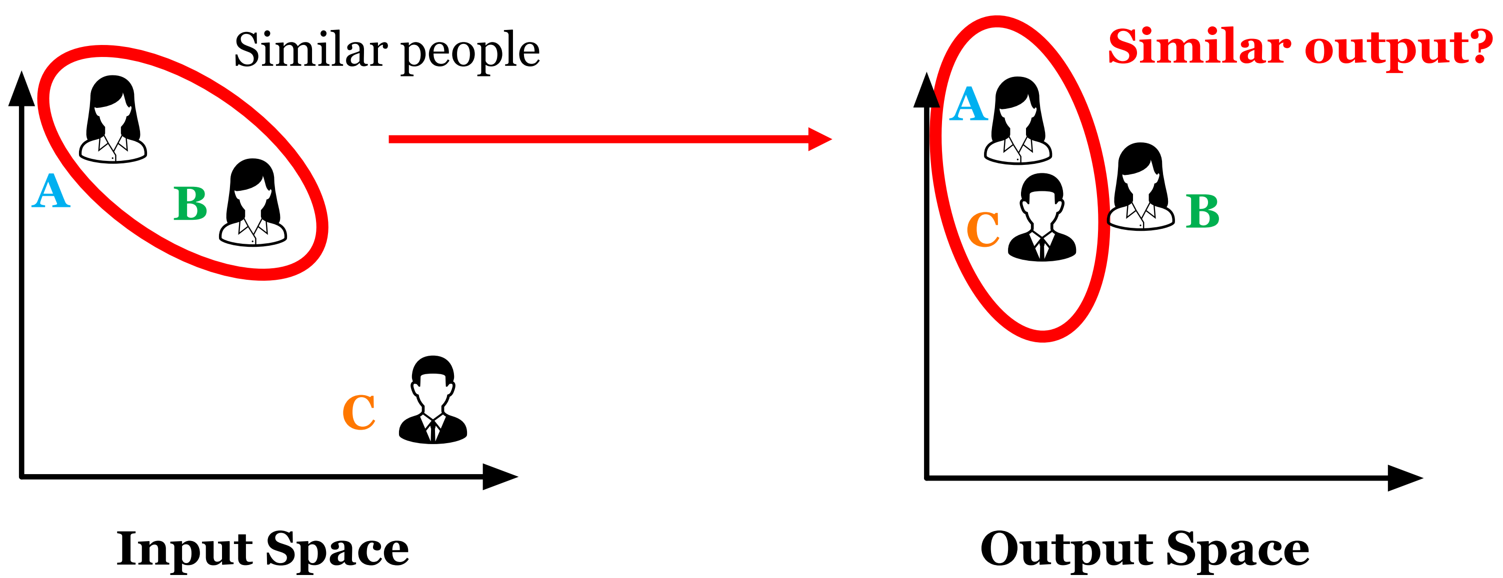
$$\|\mathbf{Y}[i,:] - \mathbf{Y}[j,:]\|_F^2 \leq \frac{\epsilon}{\mathbf{S}[i,j]} \quad \forall\, i, j = 1, \dots, n$$

Y: Output matrix to compute D_1; **S**: Similarity matrix according to $D_2(x, y)$

[1] Kang, Jian, et al. Inform: Individual fairness on graph mining. In SIGKDD, 2020.

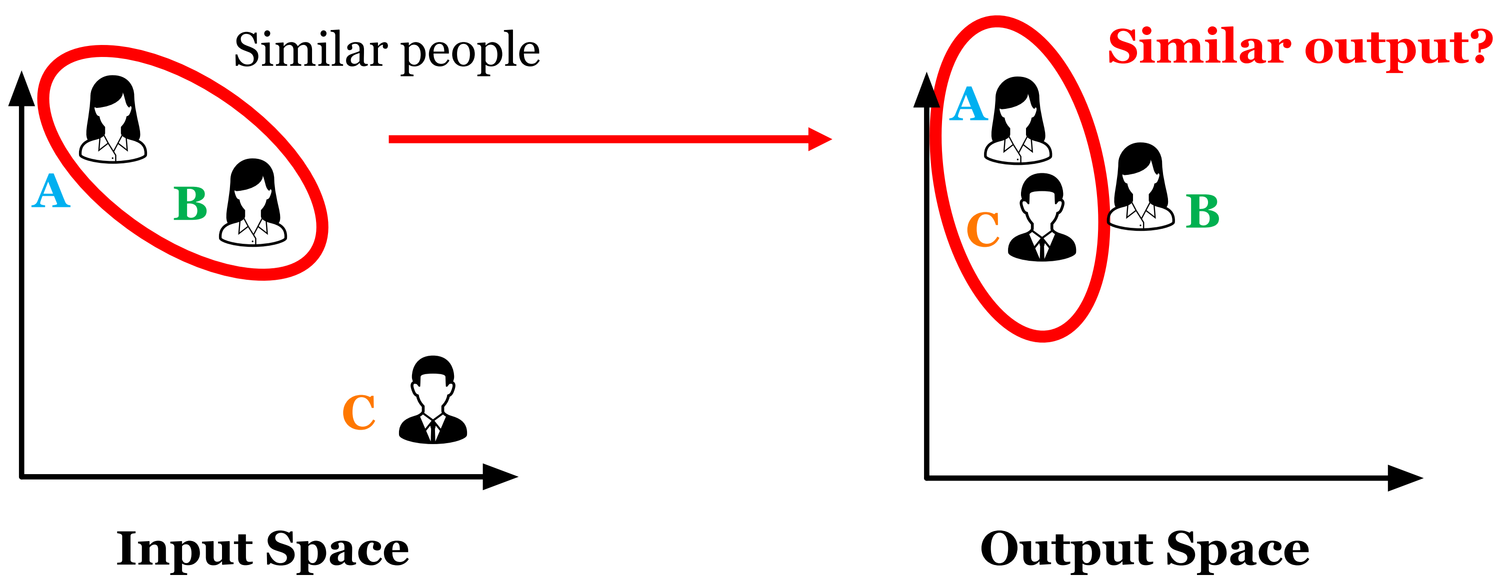
Node Ranking-Based Fairness

Node Pair Distance-Based Fairness can lead to unfairness in a relative perspective: **B is closer to A compared with C in the input space, but A and C is closer in the output space.**



Node Ranking-Based Fairness

Node Pair Distance-Based Fairness can lead to unfairness in a relative perspective: **B is closer to A compared with C in the input space, but A and C is closer in the output space.**



This could lead to a **sense of unfairness** for involved individuals.

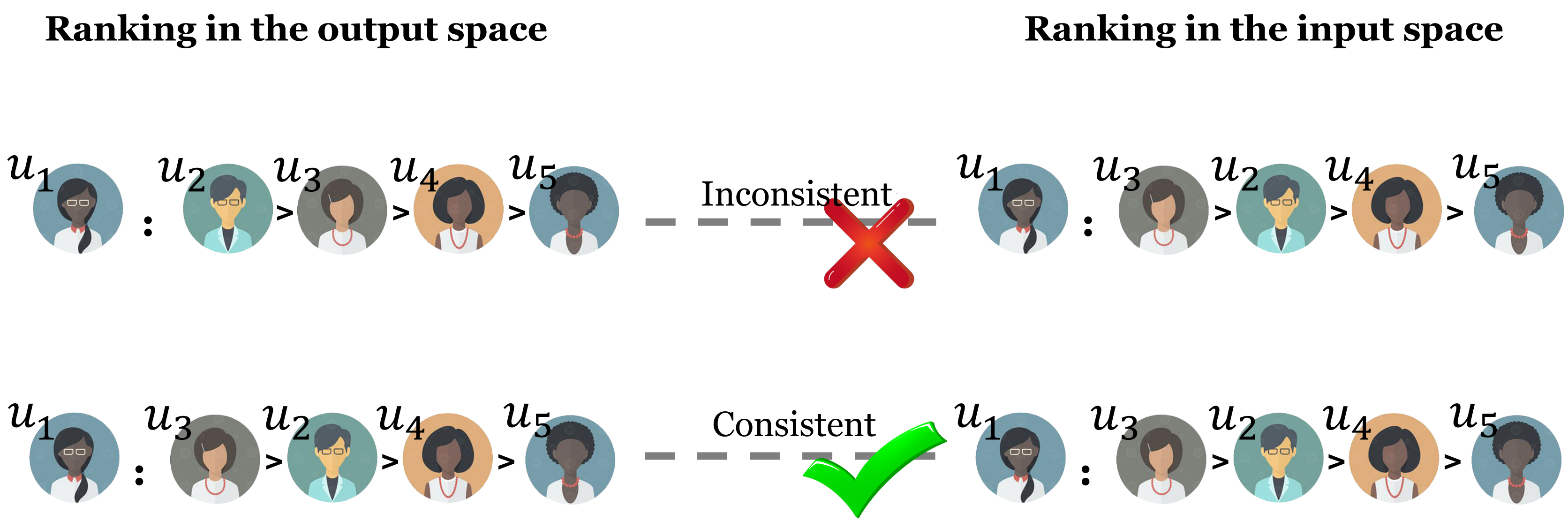
Node Ranking-Based Fairness

Criterion: for each individual, its similarity **rankings** (between itself and all other people) in both input and output space should be the **same** [1].

[1] Dong, Yushun, et al. "Individual fairness for graph neural networks: A ranking based approach." In SIGKDD, 2021.

Node Ranking-Based Fairness

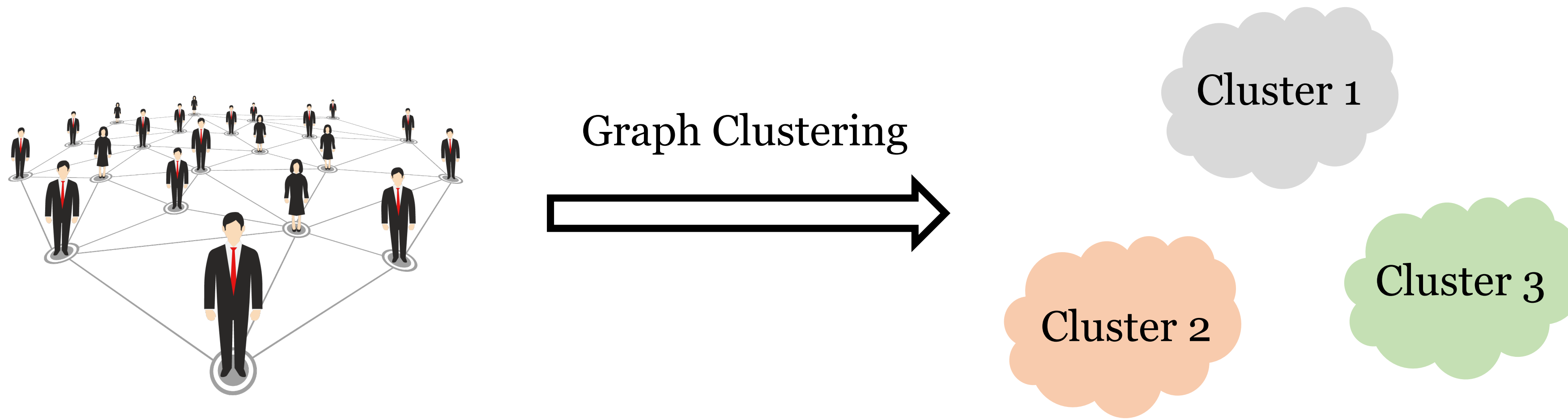
Criterion: for each individual, its similarity **rankings** (between itself and all other people) in both input and output space should be the **same** [1].



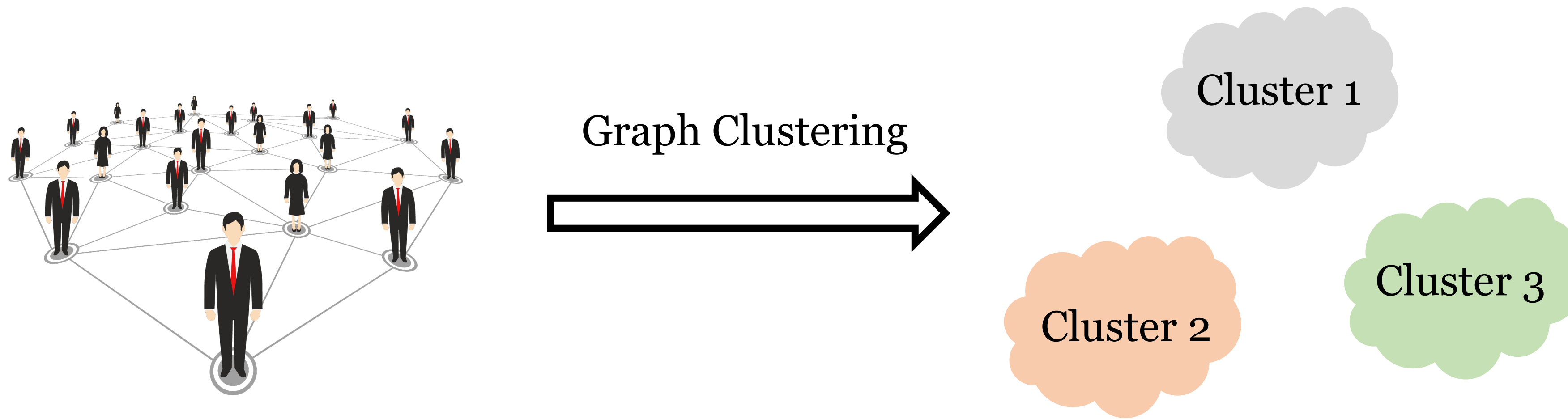
Metrics: average ranking similarity across all individuals, e.g., average NDCG@k [2].

[1] Dong, Yushun, et al. "Individual fairness for graph neural networks: A ranking based approach." In SIGKDD, 2021.
[2] Kleindessner, Matthäus, et al. "Guarantees for spectral clustering with fairness constraints." In ICML, 2019.

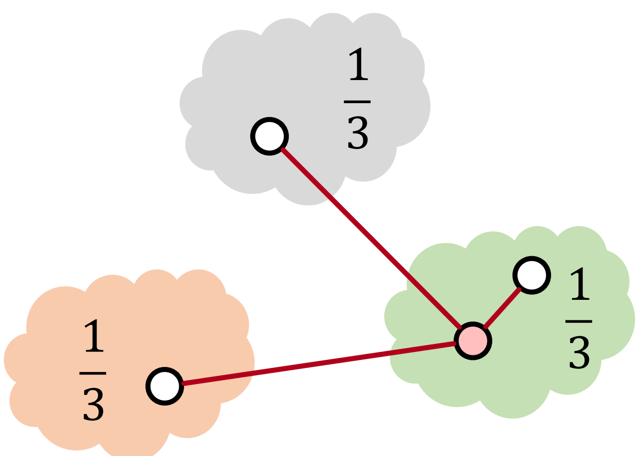
Individual Fairness in Graph Clustering



Individual Fairness in Graph Clustering

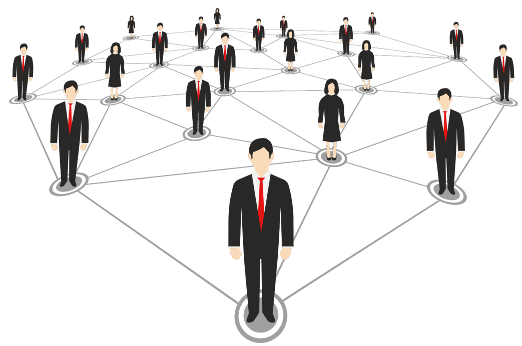


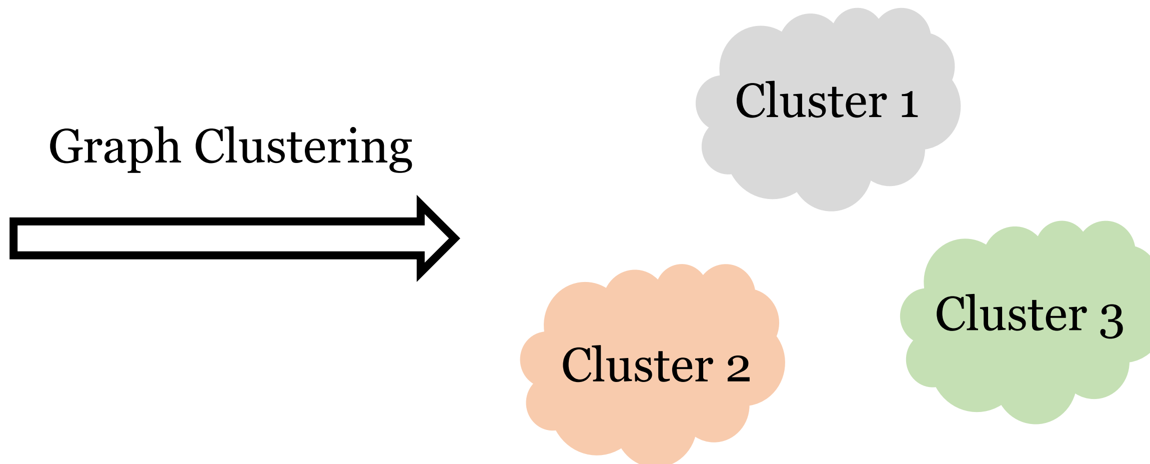
Criterion: For every node ○ , its neighbors should be proportionally represented by each cluster [1].

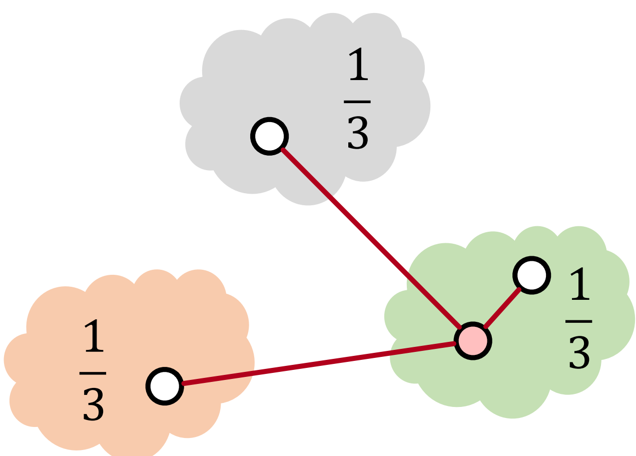


[1] Gupta, Shubham, et al. “Protecting individual interests across clusters: Spectral clustering with guarantees.” arXiv preprint arXiv:2105.03714, 2021.

Individual Fairness in Graph Clustering







Criterion: For every node , its neighbors should be proportionally represented by each cluster [1].

Metric: how disproportionately neighbors of a node are assigned in different clusters (node-level) [1].

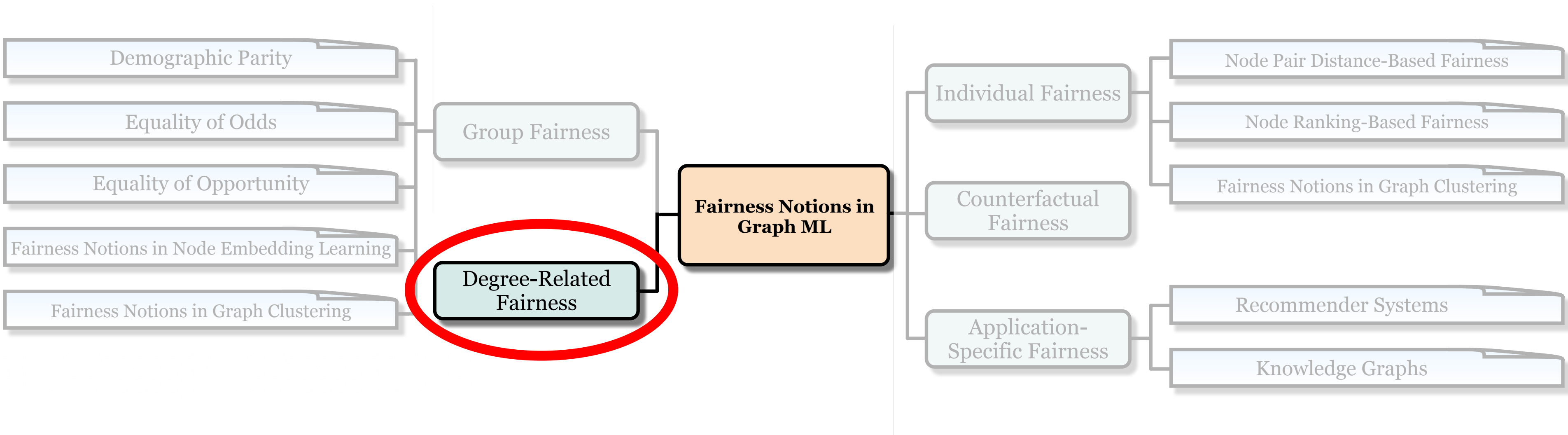
$$\rho_i = \min_{k,l \in \{1,\ldots,K\}} \frac{|C_k \cap N_{v_i}|}{|C_l \cap N_{v_i}|}$$

C_k: node set of cluster k;
C_l: node set in cluster l;
N_{v_i}: Neighbor set of node v_i;

[1] Gupta, Shubham, et al. "Protecting individual interests across clusters: Spectral clustering with guarantees." arXiv preprint arXiv:2105.03714, 2021.

Taxonomy of Fairness Notions

- A fairness notion **tailored with graph structure**: Degree-Related Fairness.

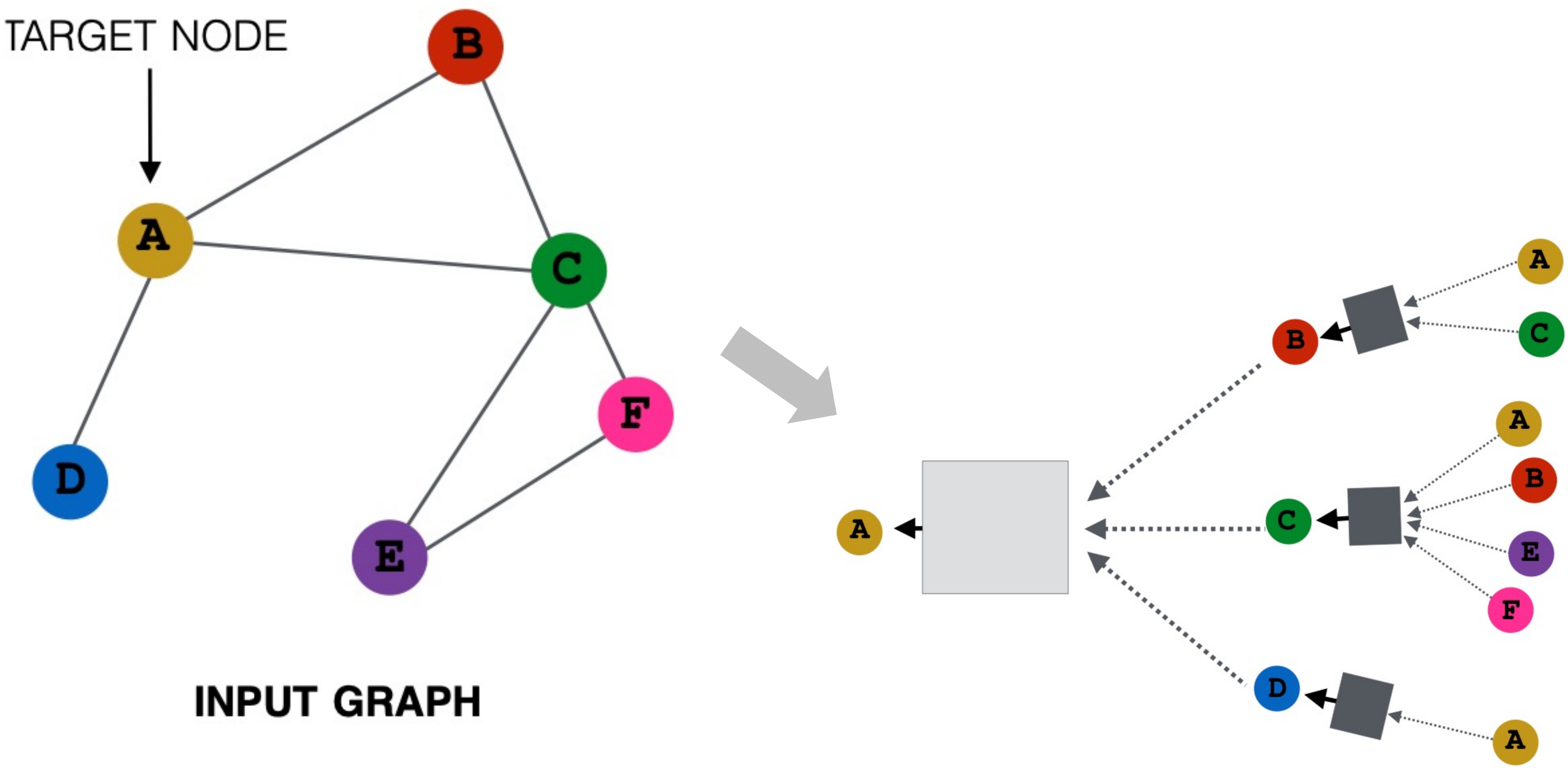


A general idea of degree-related fairness: the degree of nodes should be independent from the quality of their corresponding predictions [1, 2, 3].

[1] Tang, Xianfeng, et al. "Investigating and mitigating degree-related biases in graph convoltuional networks." In CIKM, 2020
[2] Kang, Jian, et al. "Rawlsgcn: Towards Rawlsian difference principle on graph convolutional network." In WWW, 2022.
[3] Liu, Zemin, et al. "On Generalized Degree Fairness in Graph Neural Networks." arXiv preprint arXiv:2302.03881 (2023).

Degree-Related Fairness

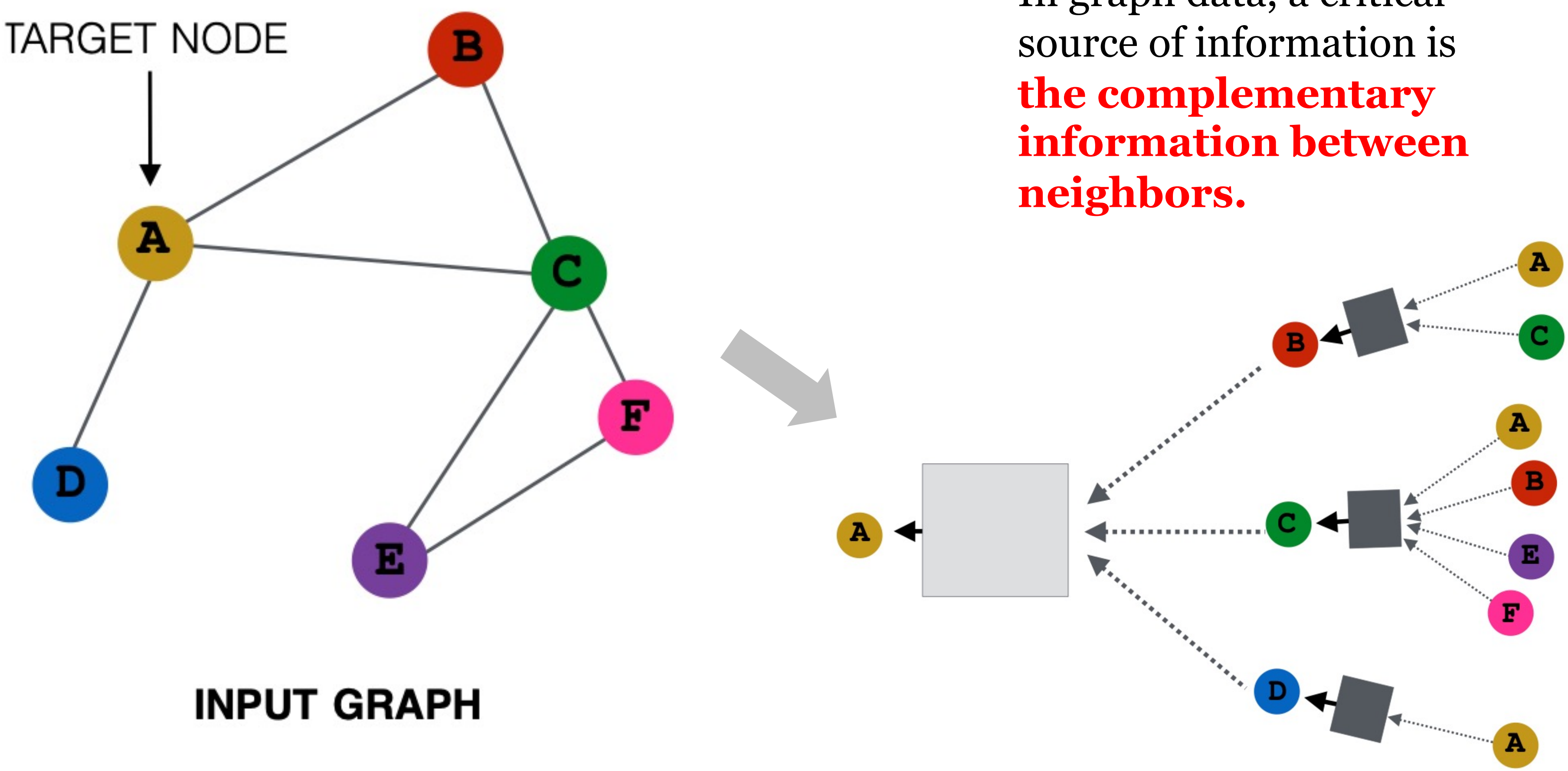
A typical **information aggregation** in Graph Neural Networks:



Degree-Related Fairness

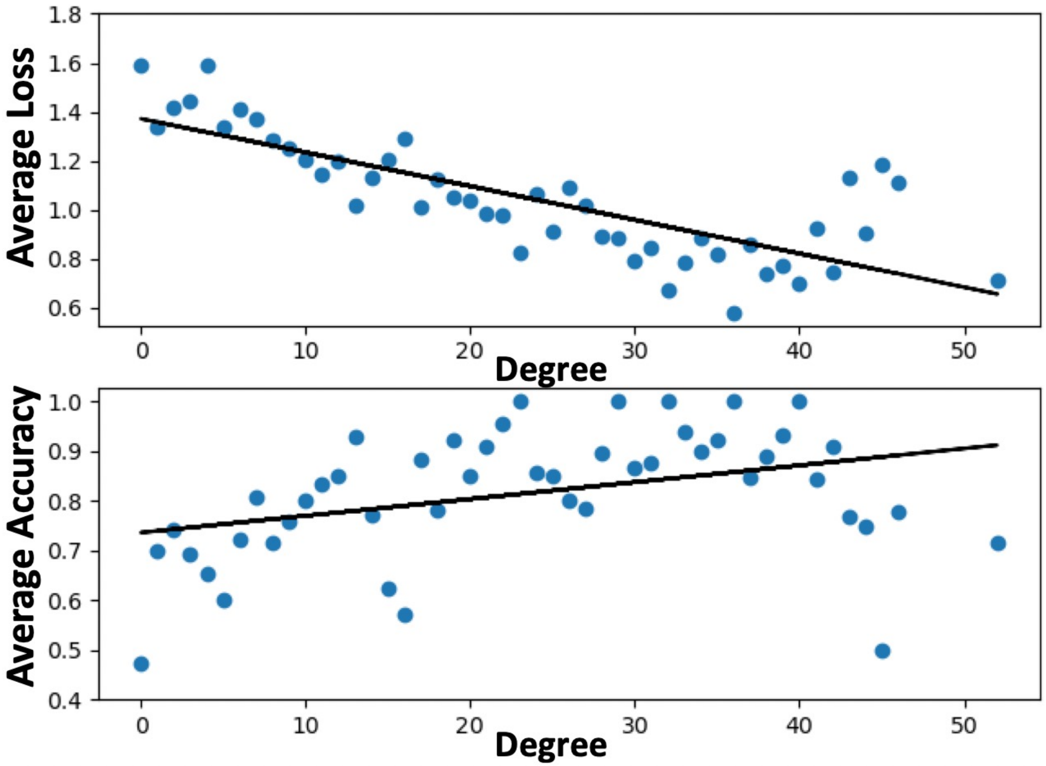
A typical **information aggregation** in Graph Neural Networks:

In graph data, a critical source of information is **the complementary information between neighbors.**



Degree-Related Fairness (Cont.)

A typical **average loss distribution** across node degrees in Graph Neural Networks [1]:



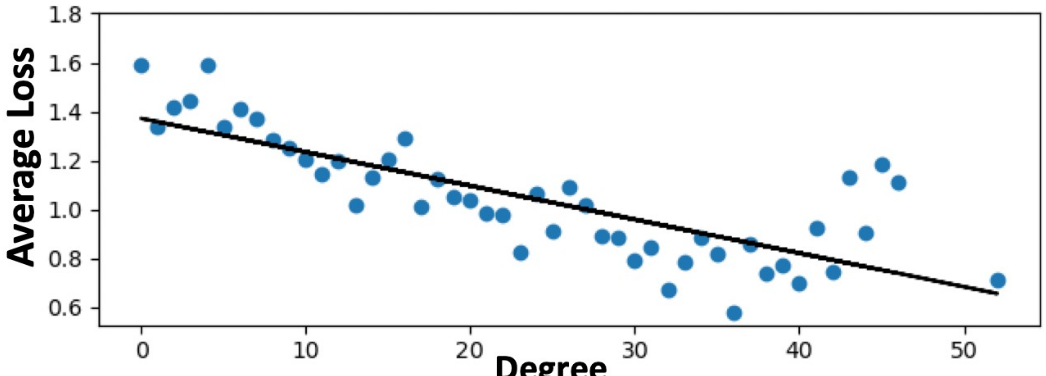
In graph data, a critical source of information is **the complementary information between neighbors.**

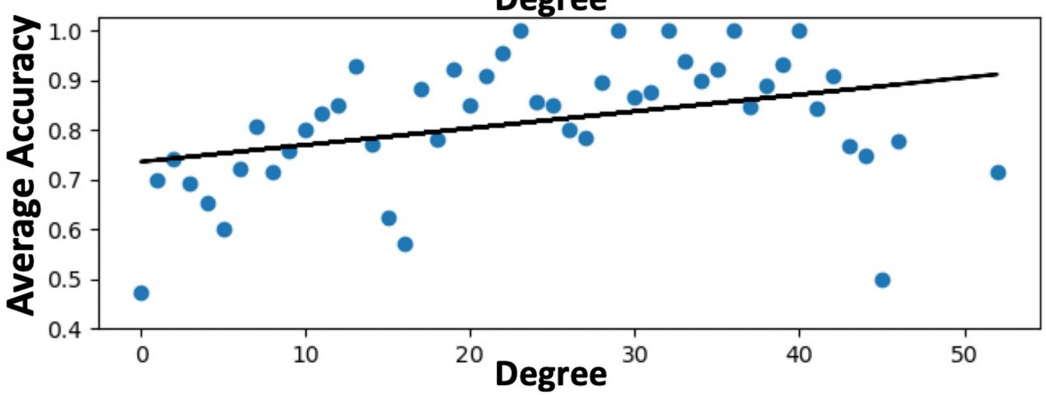
However, graph mining algorithms relying on such information tend to yield predictions with **much worse quality** for low-degree nodes, as they have **fewer neighbors**.

[1] Jian, Kang, et al. "Rawlsgcn: Towards Rawlsian difference principle on graph convolutional network." In TheWebConf, 2020.

Degree-Related Fairness (Cont.)

A typical **average loss distribution** across node degrees in Graph Neural Networks:





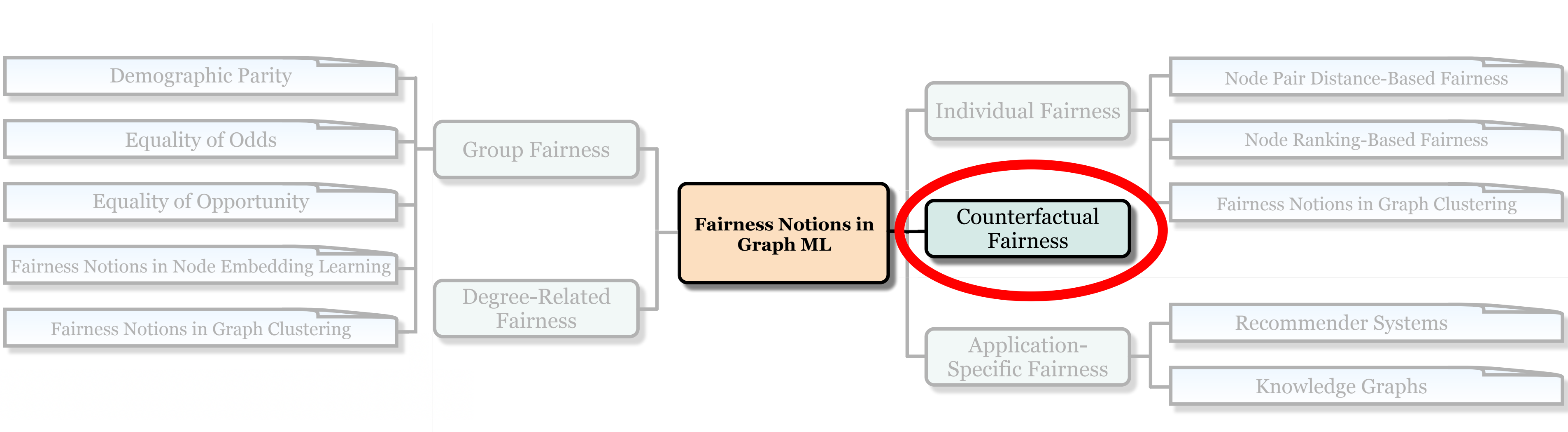
In graph data, a critical source of information is **the complementary information between neighbors.**

However, graph mining algorithms relying on such information tend to yield predictions with **much worse quality** for low-degree nodes, as they have **fewer neighbors**.

Degree-Related Fairness requires that nodes should bear similar utility (e.g., node classification accuracy) in the graph mining algorithms **regardless of their degrees**.

Taxonomy of Fairness Notions

A fairness notion **from the causal perspective**: counterfactual fairness.



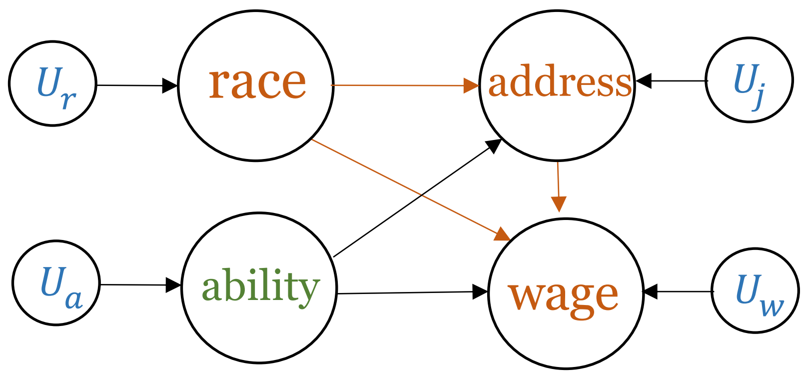
A general idea of counterfactual fairness: the sensitive information of any individual **should not causally influence** the corresponding output [1].

[1] Kusner, Matt J., et al. "Counterfactual fairness." In NeurIPS, 2017.

Background: Causal Model

Structural causal model [1]

- Independent exogenous variables (U)
- Endogenous variables
- Causal graph (a Directed Acyclic Graph) & structural equations (functions which describe the relations between variables)



Biased information

[1] Pearl, Judea. Causality. Cambridge university press, 2009.

Counterfactual Fairness

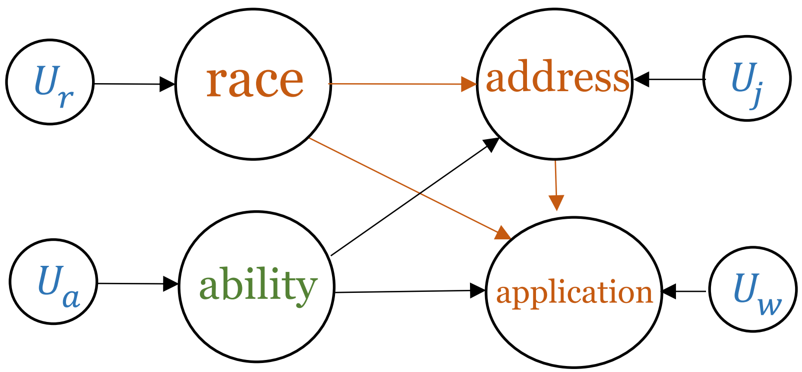
Prediction $\hat{Y}$ is **counterfactually fair** if under any features $X = x$ and sensitive attribute $S = s$:

$$P(\boxed{\widehat{Y}_{S\leftarrow s}} = y|\boxed{X = x}, \boxed{S = s}) = P(\boxed{\widehat{Y}_{S\leftarrow s'}} = y|X = x, S = s)$$

The value of the prediction if S had been set to s (s')
Notice: other features may change correspondingly.

Features

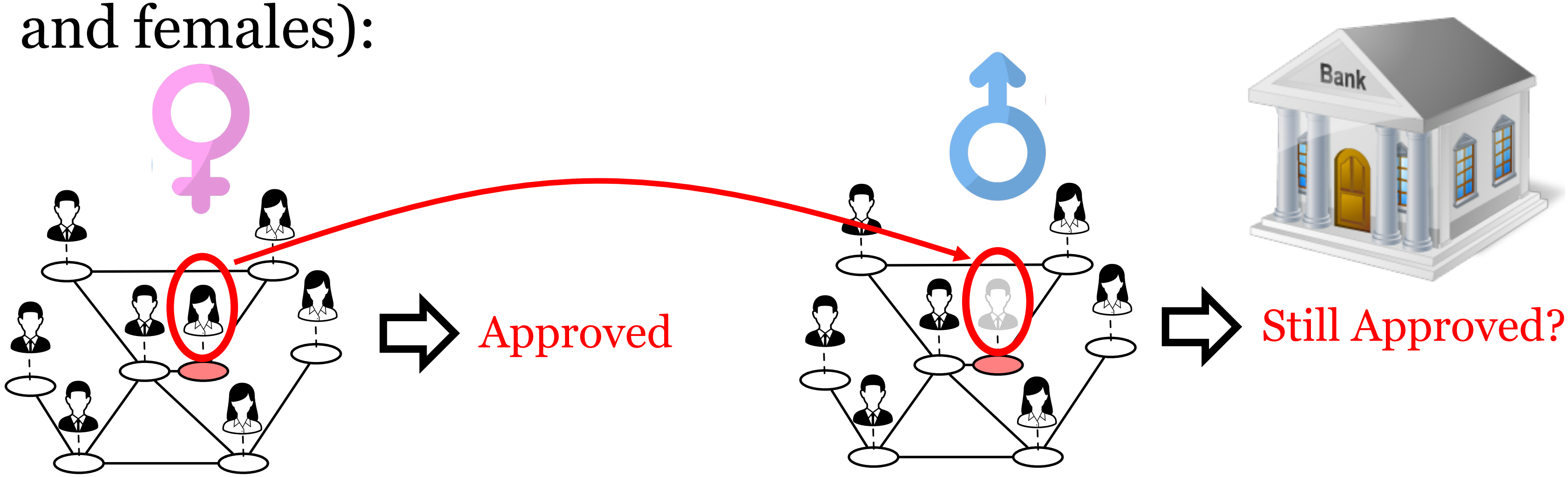
Sensitive attribute



Descendants of the sensitive attribute will be also changed after intervention

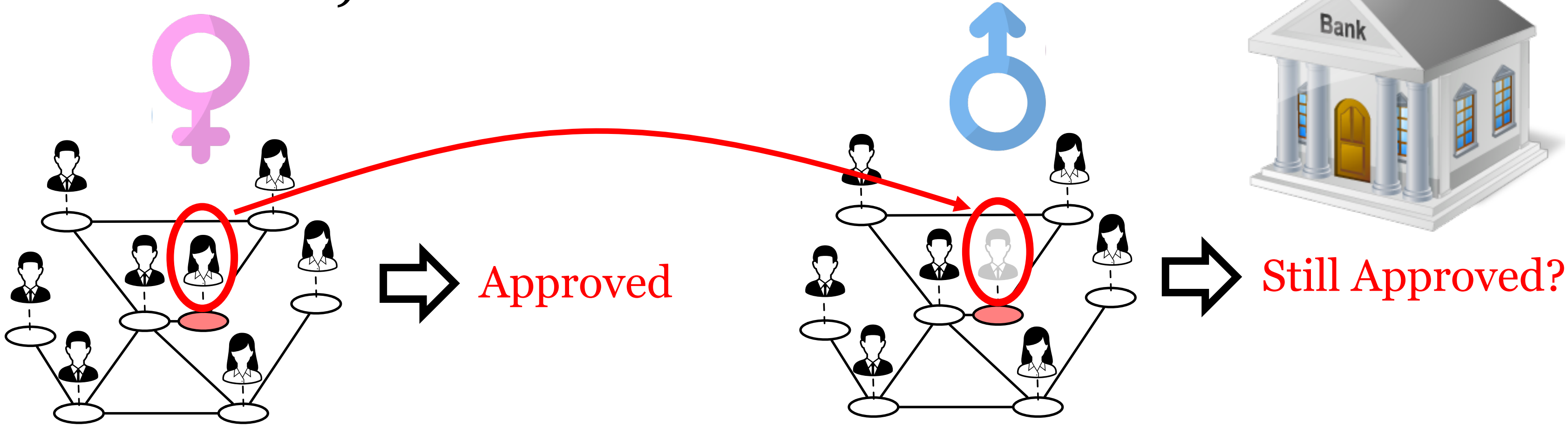
Counterfactual Fairness on Graphs

Consider a network of loan applicants (including males and females):



Counterfactual Fairness on Graphs

Consider a network of loan applicants (including males and females):



Criterion: If the sensitive feature of an individual is changed into a different value (e.g., from s to s'), the output should still be maintained the same [1].

$$P\left(\hat{Y}_{S\leftarrow s} = y | X = x, S = s\right) = P\left(\hat{Y}_{S\leftarrow s'} = y | X = x, S = s\right)$$

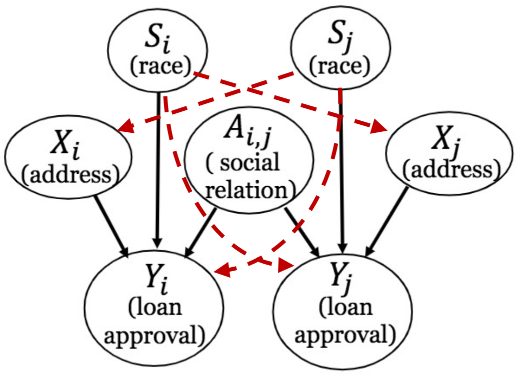
Metric: the percentage of nodes whose predicted label changes when their sensitive feature values are changed.

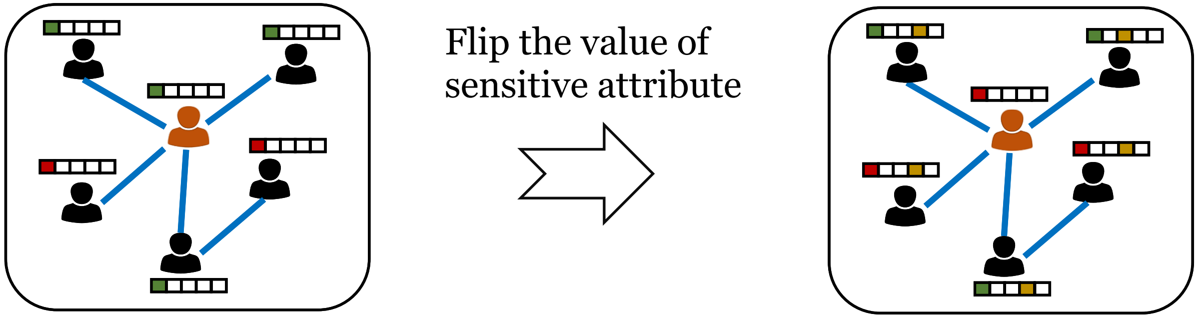
[1] Kusner, Matt J., et al. "Counterfactual fairness." In NeurIPS, 2017.

Counterfactual Fairness on Graphs

Limitations of the above fairness notion:

(1) The sensitive attributes of each node's neighbors may causally affect the prediction w.r.t. this node (red dashed edges);

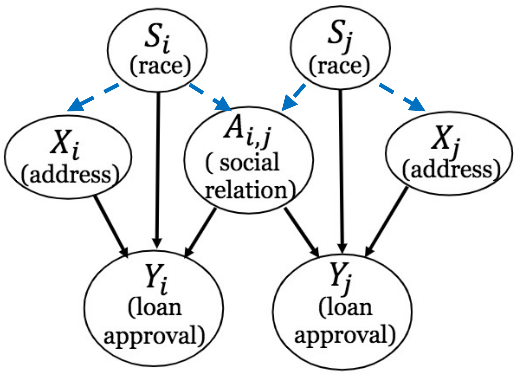


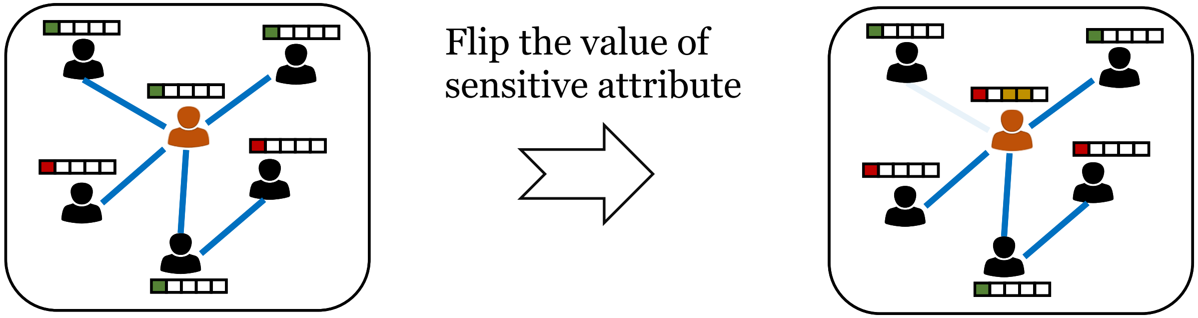


Counterfactual Fairness on Graphs

Limitations of the above fairness notion:

(2) The sensitive attributes may causally affect other features and the graph structure (blue dashed edges).





Graph Counterfactual Fairness

- **Graph counterfactual fairness** [1]: An encoder $Z_i = (\Phi(X, A))_i$ satisfies graph counterfactual fairness if for any node i:

$$P(\boxed{(Z_i)_{S \leftarrow \mathbf{s}'}}|X = \mathbf{X}, A = \mathbf{A}) = P(\boxed{(Z_i)_{S \leftarrow \mathbf{s}''}}|\boxed{X = \mathbf{X},} \boxed{A = \mathbf{A}}),$$

The node representation of i when the values of the sensitive attributes of all nodes on the graph are set to s' (s'')

Node features (including sensitive attribute)

Graph structure

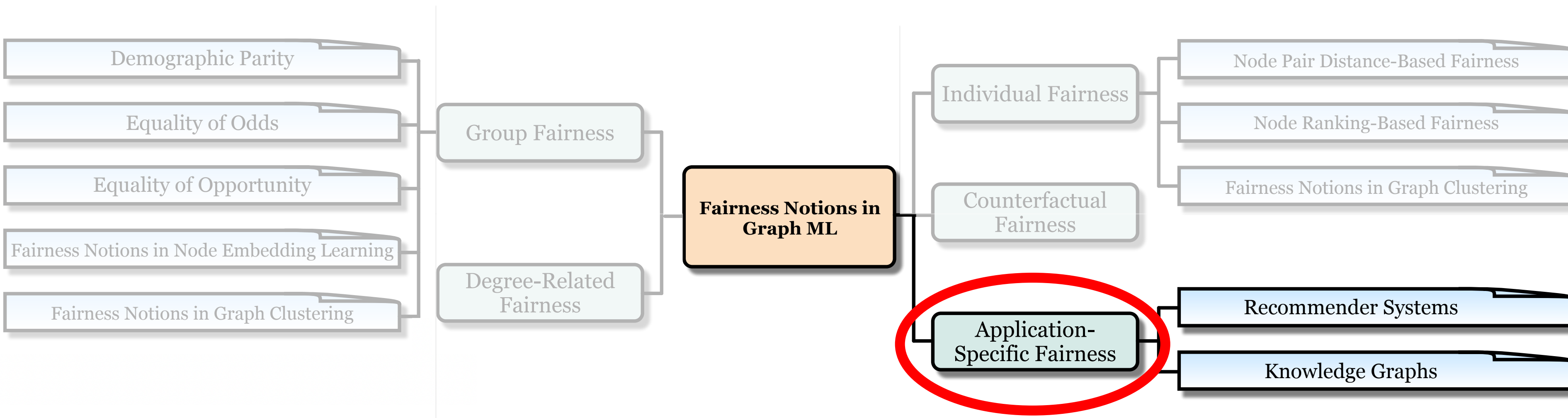
s' (s'') : an n-dimensional vector for an n-node graph

- Example: the prediction for one's loan application being approved should be the same regardless of this applicant's <u>and his/her friends' (connected in a social network)</u> sensitive information.

[1] Ma, Jing, et al. "Learning fair node representations with graph counterfactual fairness." In WSDM, 2022.

Taxonomy of Fairness Notions

Fairness notions **in real-world applications:** application-specific fairness.



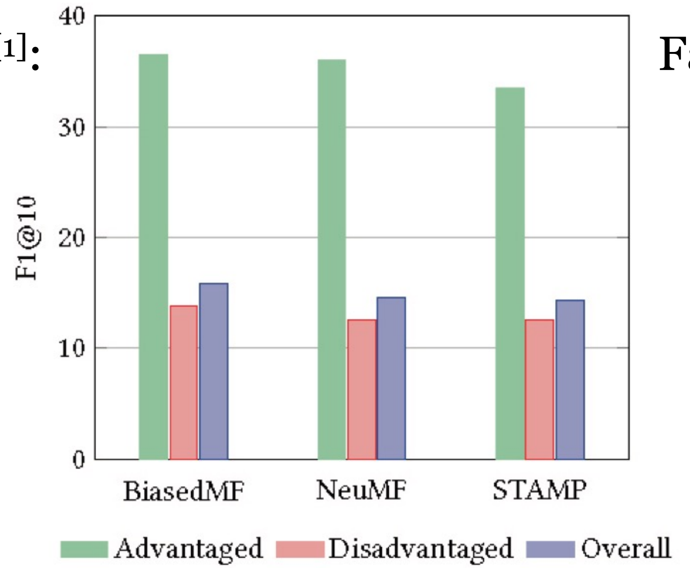
In real-world applications, certain scenarios could bring a sense of unfairness, which requires defining **application-specific fairness** to depict if there is any exhibited bias.

User Fairness in Recommendation

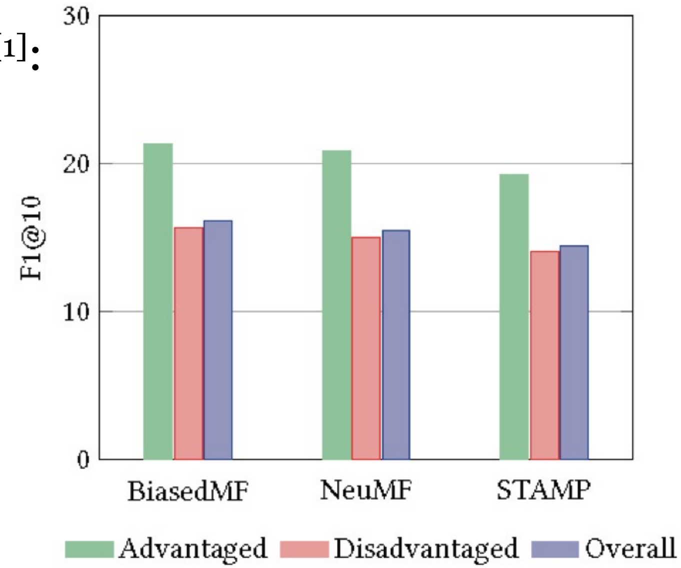
Application-specific fairness in recommender systems.

(1) User Fairness. **Quantitative recommendation utility** for different groups.

More biased [1]:



Fairer [1]:



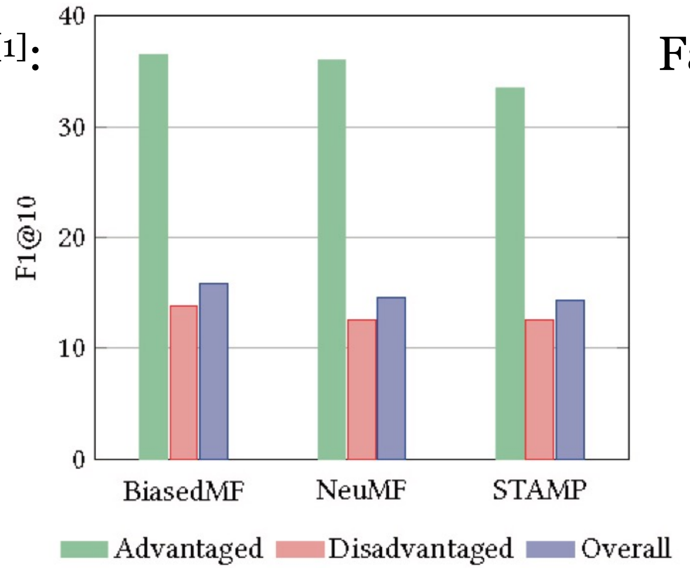
[1] Li, Yunqi, et al. "User-oriented fairness in recommendation." In WWW, 2021.

User Fairness in Recommendation

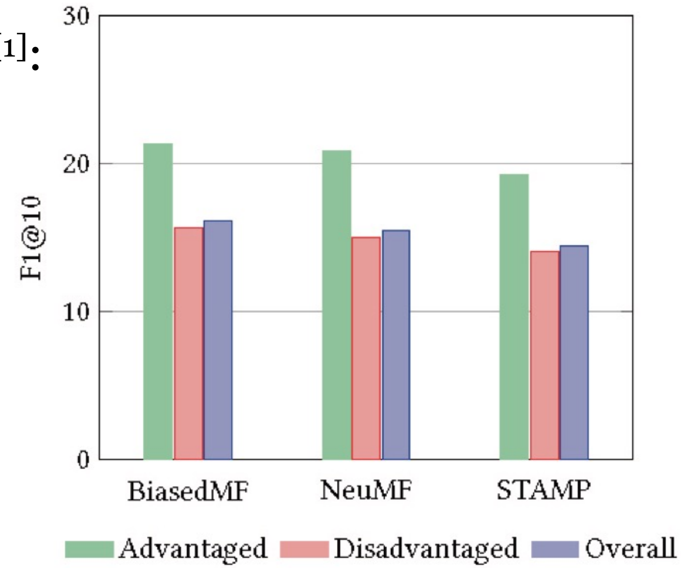
Application-specific fairness in recommender systems.

(1) User Fairness. **Quantitative recommendation utility** for different groups.

More biased [1]:



Fairer [1]:



Criterion: User fairness requires that the **recommendation quality** for different users should be similar [1, 2].

Metric: Measured with the recommendation quality discrepancy between different groups of users (e.g., active users vs. inactive users) [1, 2].

[1] Li, Yunqi, et al. "User-oriented fairness in recommendation." In WWW, 2021.
[2] Fu, Zuohui, et al. "Fairness-aware explainable recommendation over knowledge graphs." In SIGIR, 2020.

Popularity Fairness in Recommendation

Application-specific fairness in recommender systems.

(2) Popularity Fairness.

The filter bubble phenomenon: sometimes users are isolated from less popular items or information.

Popularity Fairness in Recommendation

Application-specific fairness in recommender systems.

(2) Popularity Fairness.

The filter bubble phenomenon: sometimes users are isolated from less popular items or information.

Criterion: Popular instances **should not be over-emphasized** compared with other instances [1].

Metric: Measured with the average recommendation rate of less popular instances.

[1] Fisher, Joseph, et al. "Measuring social bias in knowledge graph embeddings." In workshop of AKBC, 2020.

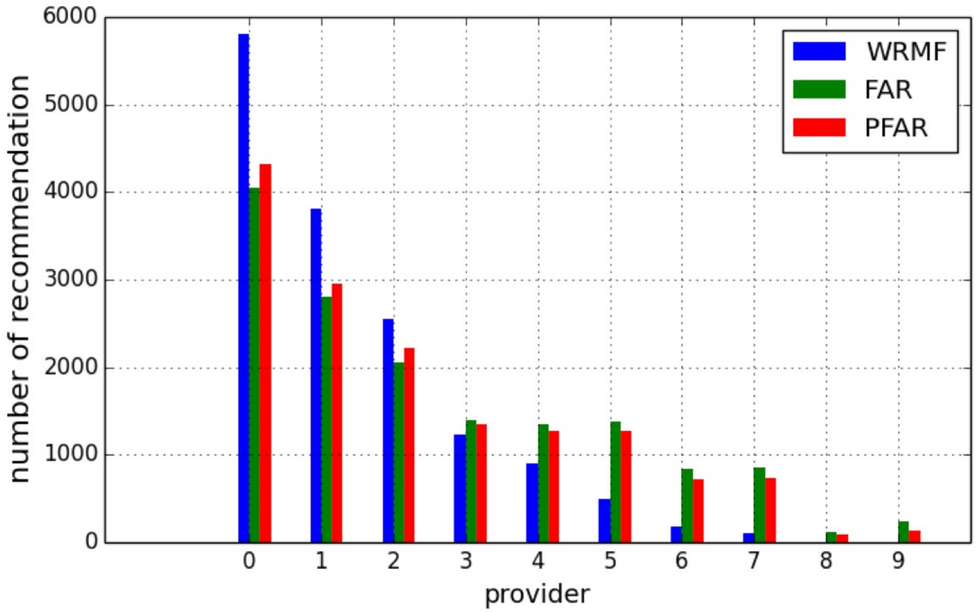
Provider Fairness in Recommendation

Application-specific fairness in recommender systems.

(3) Provider Fairness.

In a recommender system:

there could be significant differences in **the exposure rate** of items from different providers in a recommendation system [1].



[1] Liu, Weiwen, et al. "Personalizing fairness-aware re-ranking." arXiv preprint arXiv:1809.02921 (2018).

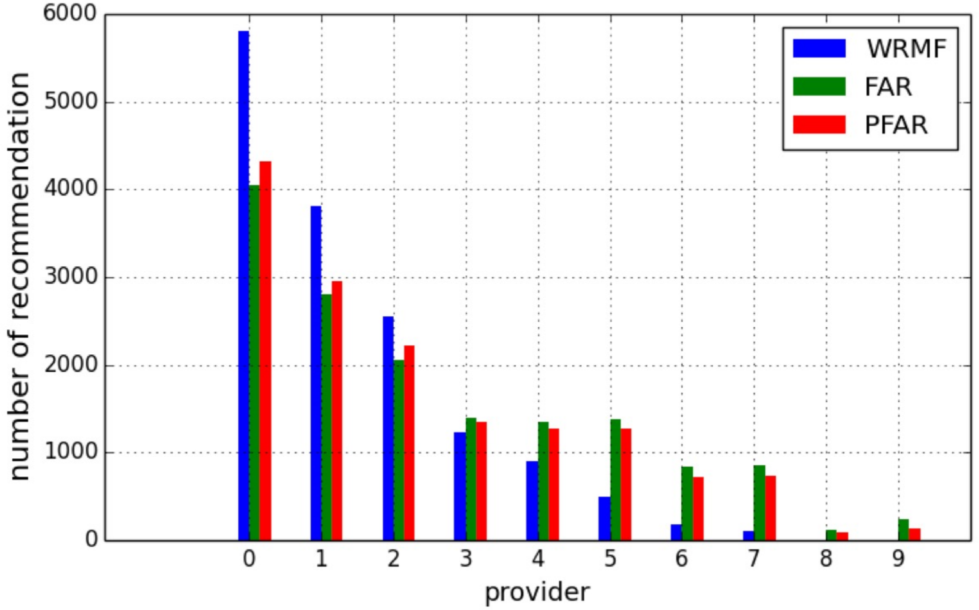
Provider Fairness in Recommendation

Application-specific fairness in recommender systems.

(3) Provider Fairness.

In a recommender system:

there could be significant differences in **the exposure rate** of items from different providers in a recommendation system [2].



Criterion: Items from different providers should receive **the same exposure rate** to the customers [1, 2, 3].

Metrics: (1) number of providers whose corresponding exposure rates are lower than a **threshold** exposure rate [1]; (2) **diversity** of providers for recommended items [2]; (3) item **exposure rate difference** between different providers [3];

[1] Boratto, Ludovico, et al. Interplay between upsampling and regularization for provider fairness in recommender systems. In UMUAI, 2020.
[2] Liu, Weiwen, et al. "Personalizing fairness-aware re-ranking." arXiv preprint arXiv:1809.02921 (2018).
[3] Patro, Gourab, et al. "Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms." In WWW, 2020.

Marketing Fairness in Recommendation

Application-specific fairness in recommender systems.

(4) Marketing Fairness. Users' interactions are biased according to the marketing strategies: under certain marketing strategy, **identity-consistent users** interact more with this item [1].

[1] Wan, Mengting, et al. "Addressing marketing bias in product recommendations." In WSDM, 2020.

Marketing Fairness in Recommendation

Application-specific fairness in recommender systems.

(4) Marketing Fairness. Users' interactions are biased according to the marketing strategies: under certain marketing strategy, **identity-consistent users** interact more with this item [1].

Identity Consistent Users: Females

Identity Consistent Users: Males

[1] Wan, Mengting, et al. "Addressing marketing bias in product recommendations." In WSDM, 2020.

Marketing Fairness in Recommendation

Application-specific fairness in recommender systems.

(4) Marketing Fairness. Users' interactions are biased according to the marketing strategies: under certain marketing strategy, **identity-consistent users** interact more with this item [1].

Identity Consistent Users: Females

Identity Consistent Users: Males

Criterion: Recommender systems should not inherit such bias from data and yield biased recommendations [1].

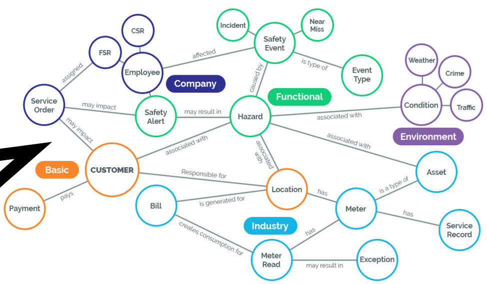
Metric: The difference of the recommendation error variance between identity-consistent and identity-inconsistent users [1].

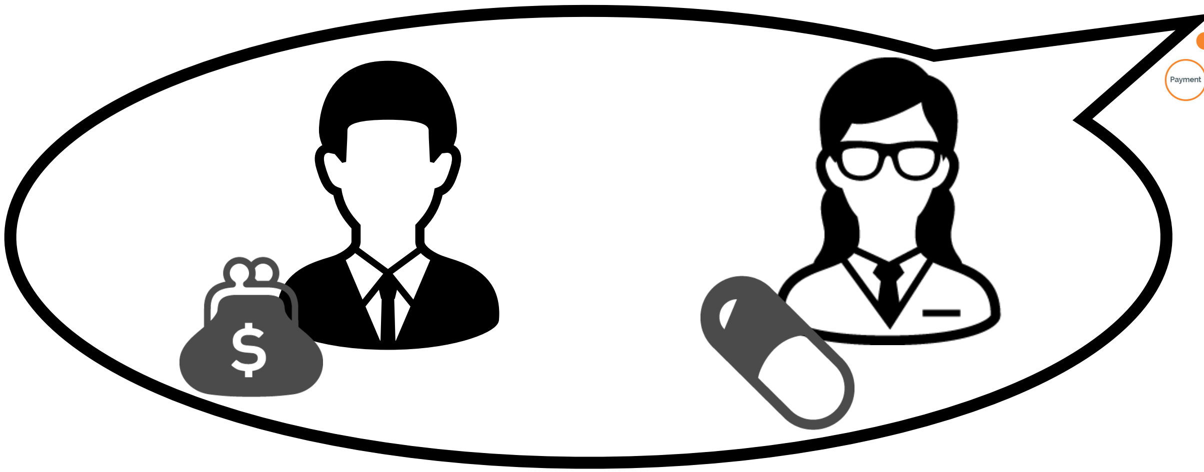
[1] Wan, Mengting, et al. "Addressing marketing bias in product recommendations." In WSDM, 2020.

Social Fairness in Knowledge Graphs

Application-specific fairness in knowledge graphs.

(1) Social Fairness.





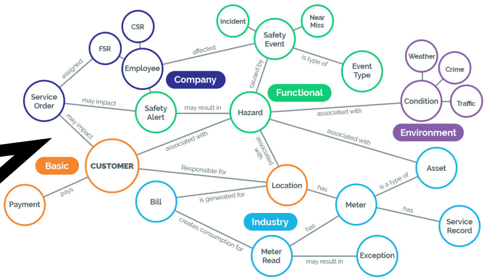
A traditional stereotype: bankers are males, while nurses are females [1].

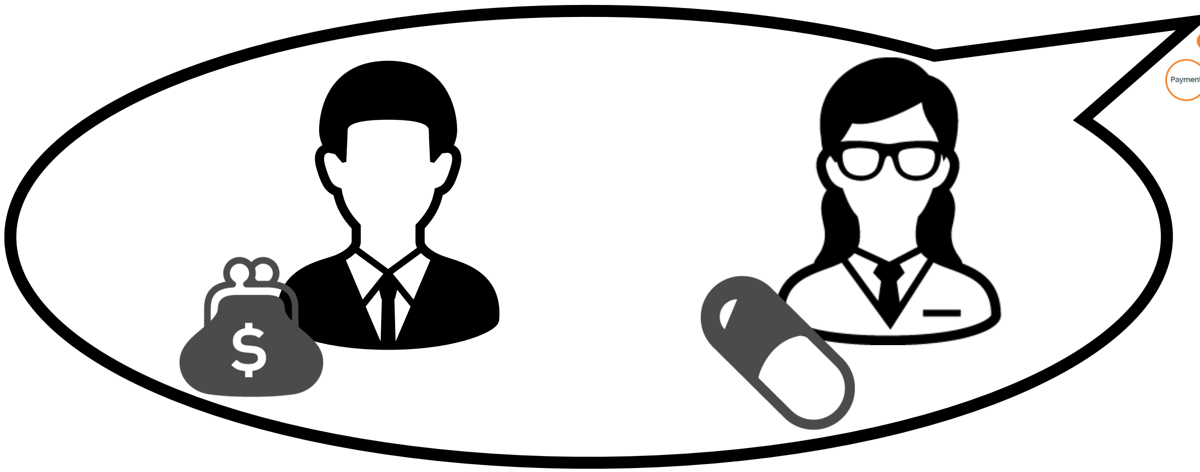
[1] Zeng, Ziqian, et al. "Fair representation learning for heterogeneous information networks." In AAAI, 2021.

Social Fairness in Knowledge Graphs

Application-specific fairness in knowledge graphs.

(1) Social Fairness.





A traditional stereotype: bankers are males, while nurses are females [1].

Criterion: The **historical biases** should not be encoded in the learned entity embeddings in knowledge graphs [1].
Metric: Distribution difference between the prediction distribution and uniform distribution over all possible sensitive feature values [2].

[1] Zeng, Ziqian, et al. “Fair representation learning for heterogeneous information networks.” In AAAI, 2021.
[2] Fisher, Joseph, et al. "Debiasing knowledge graph embeddings." In EMNLP, 2020.

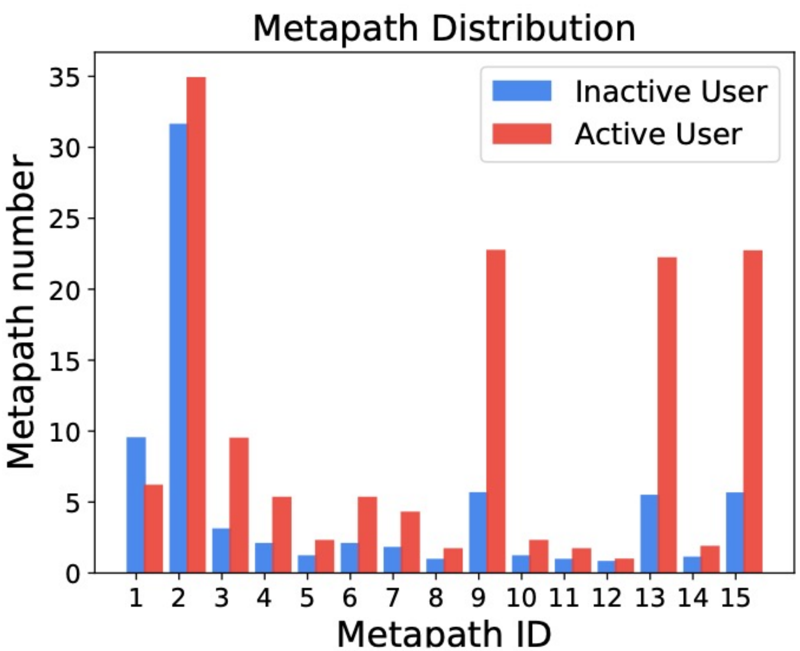
Path Diversity Fairness in Knowledge Graphs

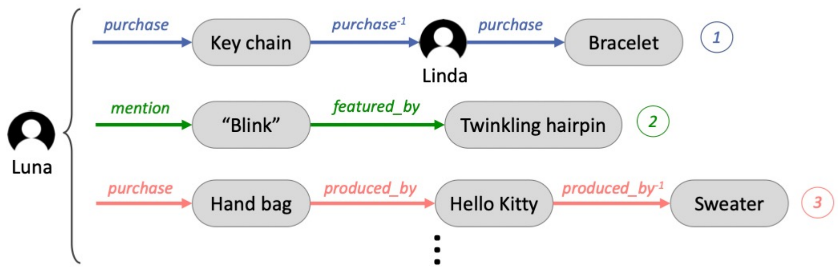
Application-specific fairness in knowledge graphs.

(2) Path Diversity Fairness.

On a **user-item knowledge graph**:

Meta-path distributions over their types can be different across different person entity groups [1].





[1] Fu, Zuohui, et al. "Fairness-aware explainable recommendation over knowledge graphs." In SIGIR, 2020.

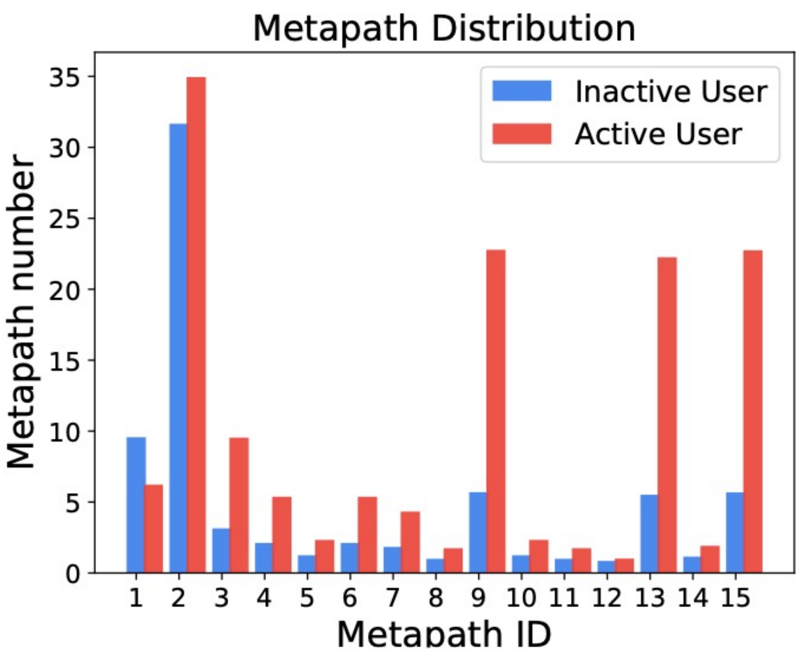
Path Diversity Fairness in Knowledge Graphs

Application-specific fairness in knowledge graphs.

(2) Path Diversity Fairness.

On a **user-item knowledge graph**:

Meta-path distributions over their types can be different across different person entity groups [1].



Criterion: The distributions of meta-paths (over their types) should be similar across different demographic subgroups in the knowledge graph [1].

Metric: The difference of Simpson's Index of Diversity (SID) between the meta-path distributions of different demographic subgroups [1].

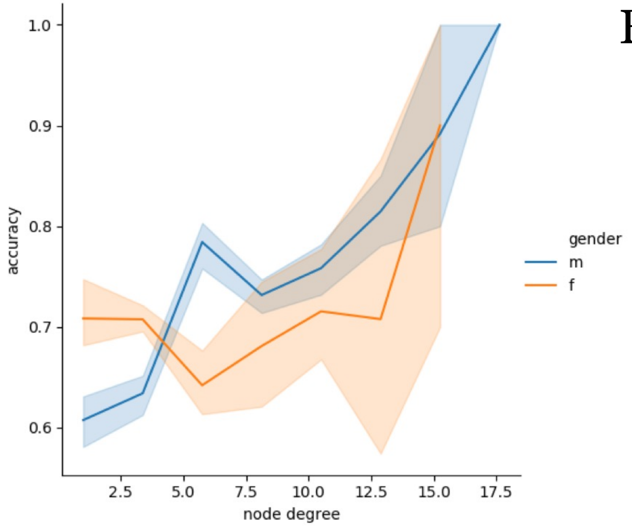
[1] Fu, Zuohui, et al. "Fairness-aware explainable recommendation over knowledge graphs." In SIGIR, 2020.

Popularity Fairness in Knowledge Graphs

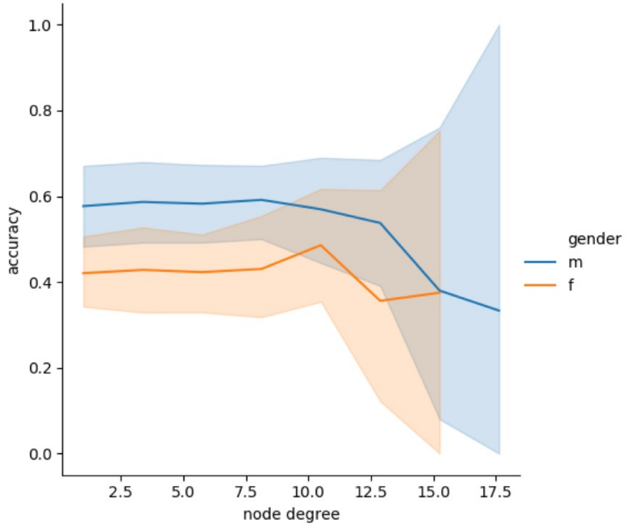
Application-specific fairness in knowledge graphs.

(3) Popularity Fairness. **Prediction for person entities** based on DBpedia.

More biased [1]:



Fairer [1]:

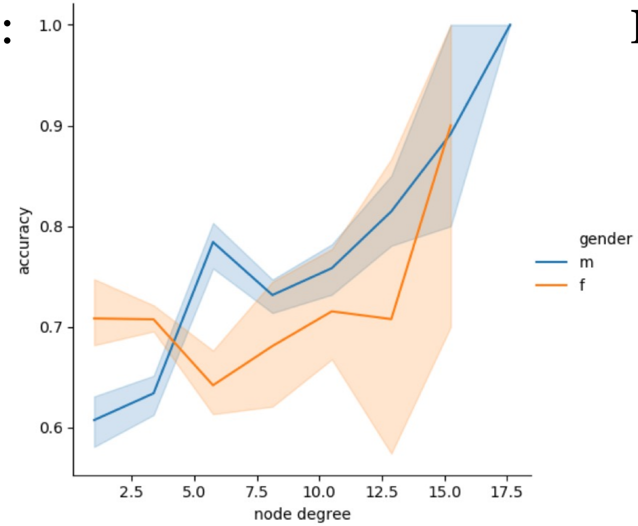


Popularity Fairness in Knowledge Graphs

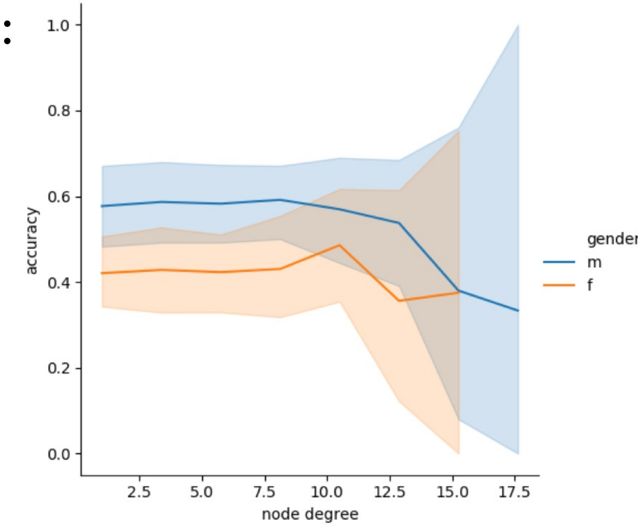
Application-specific fairness in knowledge graphs.

(3) Popularity Fairness. **Prediction for person entities** based on DBpedia.

More biased [1]:



Fairer [1]:



Criterion: The prediction accuracy under certain tasks should be uniformly distributed w.r.t. entity node popularity (e.g., defined as the entity node degree) in the knowledge graph [1].

Metric: Difference between the output distribution of accuracy w.r.t. entity popularity and a uniform distribution [1].

[1] Arduini, Mario, et al. "Adversarial learning for debiasing knowledge graph embeddings." In SIGKDD, 2020.

Outline

Background Introduction

Fairness Notions and Metrics

Theoretical Understanding of Bias

Techniques for Fair Node Embeddings

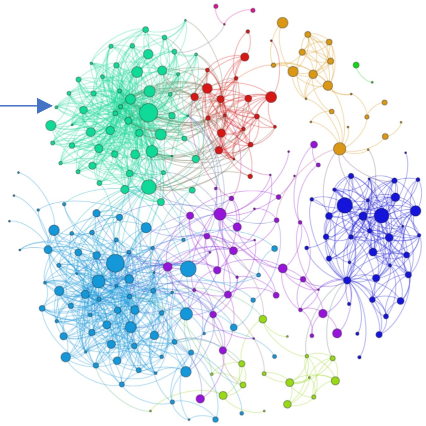
Real-World Applications

Summary, Challenges, & Future Directions

Motivation and Unique Challenges

- **Motivation:** Theoretical understanding of bias is crucial
  - Large-scale deployment in critical decision-making applications
  - Guidance for fairness-aware algorithm design
  - Explainability for the developed strategies

- **Challenge:** Analysis for tabular data cannot be directly extended to graphs
  - Non-IID structure of graph data
  - Intertwined bias from both nodal features and graph structure



- Need to develop novel analysis techniques for different learning frameworks and fairness notions

Overview

Mean-discrepancy analysis
[Li, Peizhao, et. al., ICLR, 2021]
[Kose, O. Deniz, et. al., Arxiv, 2023]

Entropy-based analysis
[Jiang, Zhimeng, et al., Arxiv, 2023]

Correlation-based analysis
[Kose, O. Deniz et. al., TNNLS, 2023]

Analysis Techniques for Understanding Bias in Graph ML

PAC-Bayesian analysis
[Ma, Jiaqi, et. al., NeurIPS, 2021]

Gradient-based analysis
[Kang, Jian et. al., WWW, 2022]

Mean-discrepancy Analysis

- Bias term: discrepancy of node representations from two sensitive groups
    - Assuming binary sensitive attribute

- Inherently related to demographic parity
    - Analytically demonstrated for both link prediction and node classification

$$\boldsymbol{h}_v^{l+1} = \sigma\left(\sum_{u \in \mathcal{N}_v} \frac{1}{\deg(v)} \cdot \boldsymbol{W}^l \boldsymbol{h}_u^l\right)$$

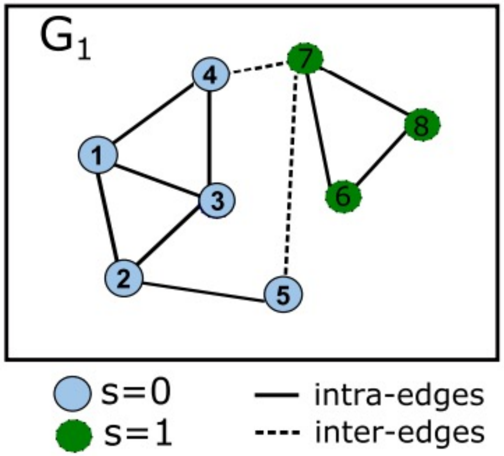
- Existing two works:
    - link prediction with mean aggregation scheme [1]
    - node classification using attention-based aggregation [2]

$$\boldsymbol{h}_v^{l+1} = \sigma\left(\sum_{u \in \mathcal{N}_v} \alpha_{vu}^l \cdot \boldsymbol{W}^l \boldsymbol{h}_u^l\right)$$

[1] Li, Peizhao, et al. "On dyadic fairness: Exploring and mitigating bias in graph connections." In ICLR, 2021.
[2]Kose, O. Deniz, et al. "FairGAT: Fairness-aware graph attention networks." Arxiv, 2023.

Intuitions from Mean-discrepancy

- These analyses show demographic parity affected by weights on intra- and inter-edges

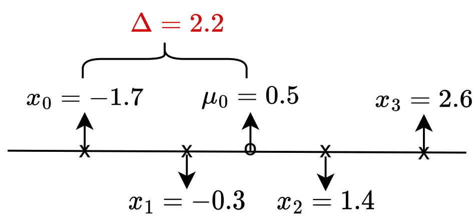


- Main finding: Balance the weights of inter- and intra-edges
  - Edge weight balancing
  - Change input graph topology via edge augmentations

Discrepancy for Mean Aggregation

- Bound $\Delta_{\mathrm{DP}}^{\mathrm{Aggr}} := \left\| \mathbb{E}_{v\sim\mathcal{V}}\left[\mathrm{Agg}(v) \mid v \in \mathcal{S}_0\right] - \mathbb{E}_{v\sim\mathcal{V}}\left[\mathrm{Agg}(v) \mid v \in \mathcal{S}_1\right]\right\|_2$

Mean aggregation at lth layer



Theorem [1]:

$$\boxed{\left\|\mathbf{x}_v - \mu_0\right\|_\infty \leq \Delta, \forall v \in \mathcal{S}_0}$$

$$\max\left\{\beta_{\min}\left\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right\|_\infty - 2\Delta, 0\right\} \leq \Delta_{\mathrm{DP}}^{\mathrm{Aggr}} \leq \beta_{\max}\left\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right\|_2 + 2\sqrt{N}\Delta$$

$$\boxed{\mathbb{E}_{v\sim\mathcal{V}}\left[\mathbf{x}_v \mid v \in \mathcal{S}_0\right]}$$

[1] Li, Peizhao, et al. "On dyadic fairness: Exploring and mitigating bias in graph connections." In ICLR, 2021.

Guidelines for Fair Link Prediction

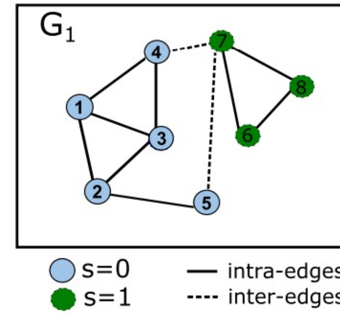
$$\max\{\beta_{\min}\|\boldsymbol{\mu}_0-\boldsymbol{\mu}_1\|_\infty-2\Delta,0\}\le\Delta_{\mathrm{DP}}^{\mathrm{Aggr}}\le\beta_{\max}\|\boldsymbol{\mu}_0-\boldsymbol{\mu}_1\|_2+2\sqrt{N}\Delta$$

$$\beta_{\min}=\min\{\beta_1,\beta_2\},\beta_{\max}=\max\{\beta_1,\beta_2\}$$

$$\boxed{m_w:=\sum_{s_v\neq s_u}A_{vu}}\qquad \boxed{\mathcal{S}_0^\chi:=\{v\in\mathcal{S}_0\mid\mathcal{N}(v)\cap\mathcal{S}_1\neq\varnothing\}}$$

$$\beta_1=\left|1-\frac{m_w}{D_{\max}}\left(\frac{1}{|\mathcal{S}_0|}+\frac{1}{|\mathcal{S}_1|}\right)\right|,\beta_2=\left|1-\frac{|\mathcal{S}_0^\chi|}{|\mathcal{S}_0|}-\frac{|\mathcal{S}_1^\chi|}{|\mathcal{S}_1|}\right|$$

$$\boxed{D_{\max}:=\max_{v\in\mathcal{V}}\deg(v)}$$



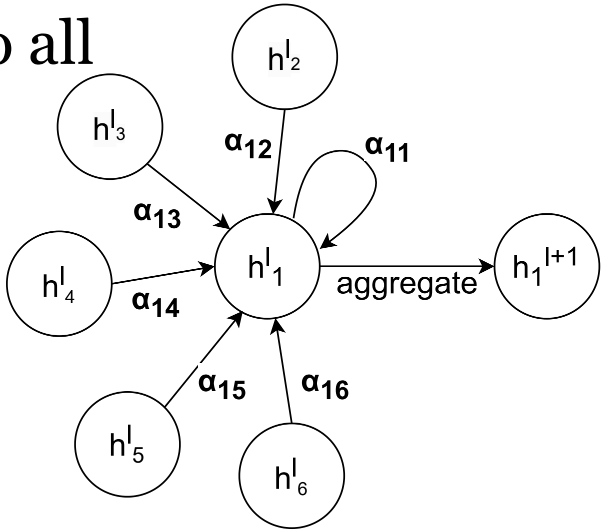
$$\mathcal{S}_0^\chi=\{4,5\}$$
$$\mathcal{S}_1^\chi=\{7\}$$

- $\beta_{\max}$: multiplying factor on the disparity of input representations
  - If topology fixed, β_2 is a constant pre-determined by input graph
  - Can modify the total weights of inter edges, m_w, to reduce β_1
  - Manipulate m_w to reduce $\beta_1 \rightarrow$ tighter upper bound for Δ_{DP}^{Aggr}

[1] Li, Peizhao, et al. "On dyadic fairness: Exploring and mitigating bias in graph connections." In ICLR, 2021.

Discrepancy Measure with Attention

- Most GNN structures assign equal weights to all neighbors



- GATs learn weights α_{vu} indicating the importance of neighbor u to node v
  - Aggregation: $\boldsymbol{h}_v^{l+1} = \sigma\left(\sum_{u \in \mathcal{N}_v} \alpha_{vu}^l \cdot \boldsymbol{W}^l \boldsymbol{h}_u^l\right)$

- For the *l*th GAT layer, [2] upper bounds the disparity of outputs for two sensitive groups:

$$\delta_h^{l+1} := \left\|\operatorname{mean}\left(\mathbf{h}_j^{l+1} \mid s_j = 0\right) - \operatorname{mean}\left(\mathbf{h}_j^{l+1} \mid s_j = 1\right)\right\|_2$$

- [2] further shows δ_h^{L+1} is equivalent to demographic parity for node classification, when the output of final layer is a probability for class label 1 (e.g., sigmoid in the last layer)

[2] Kose, O. Deniz, et al. "FairGAT: Fairness-aware graph attention networks." Arxiv, 2023.

Mean Discrepancy for Attention

Theorem [2]:

spectral norm

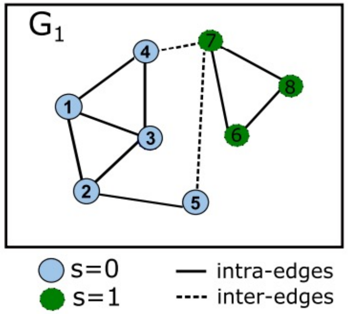
$$\delta_h^{l+1} \leq L\left(\sigma_{\max}\left(\mathbf{W}^l\right)\left|\left(R_1^\chi \alpha^\chi + R_0^\chi \alpha^\chi - 1\right)\right| \delta_h^l + c\right)$$

Lipschitz constant of the nonlinear activation

$$\alpha^\chi = \sum_{a \in \mathcal{N}(k) \cap S_i} \alpha_{ka} \quad \text{for} \quad v_k \in \mathcal{S}_j, i \neq j \quad \text{where} \quad \alpha^\chi + \alpha^\omega = 1$$

Total amount of attention assigned to inter-edges

$$R_1^\chi := \frac{\left|S_1^\chi\right|}{\left|\mathcal{S}_1\right|}, R_0^\chi := \frac{\left|S_0^\chi\right|}{\left|\mathcal{S}_0\right|}$$



$$\mathcal{S}_0^\chi = \{4, 5\}$$
$$\mathcal{S}_1^\chi = \{7\}$$

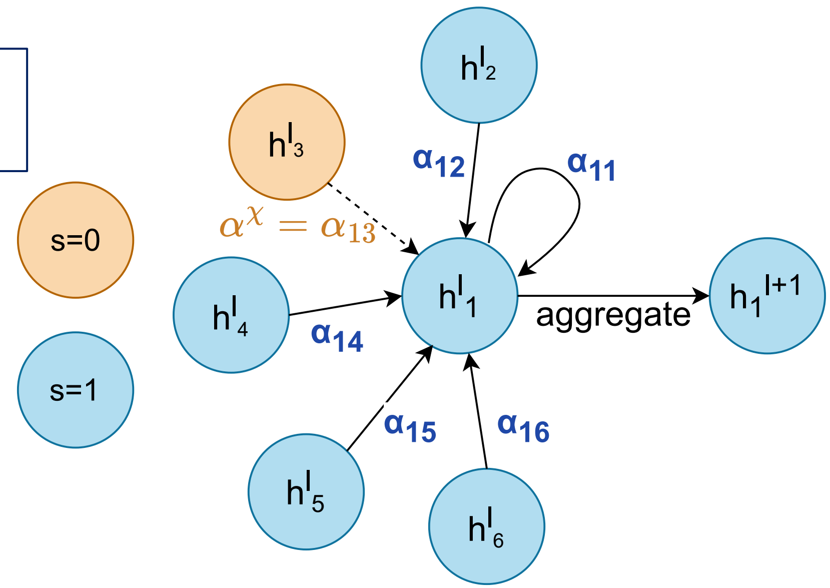
[2] Kose, O. Deniz, et al. "FairGAT: Fairness-aware graph attention networks." Arxiv, 2023.

Guidelines for Fair Attention

$$\delta_h^{l+1} \leq L\left(\sigma_{\max}\left(\mathbf{W}^l\right) \left|\left(R_1^{\chi}\alpha^{\chi} + R_0^{\chi}\alpha^{\chi} - 1\right)\right| \delta_h^l + c\right)$$

$$\boxed{\alpha^{\chi} = \sum\nolimits_{a \in \mathcal{N}(k) \cap S_i} \alpha_{ka} \quad \text{for} \quad v_k \in \mathcal{S}_j, i \neq j}$$

$$R_1^{\chi} := \frac{\left|S_1^{\chi}\right|}{|\mathcal{S}_1|}, R_0^{\chi} := \frac{\left|S_0^{\chi}\right|}{|\mathcal{S}_0|}$$

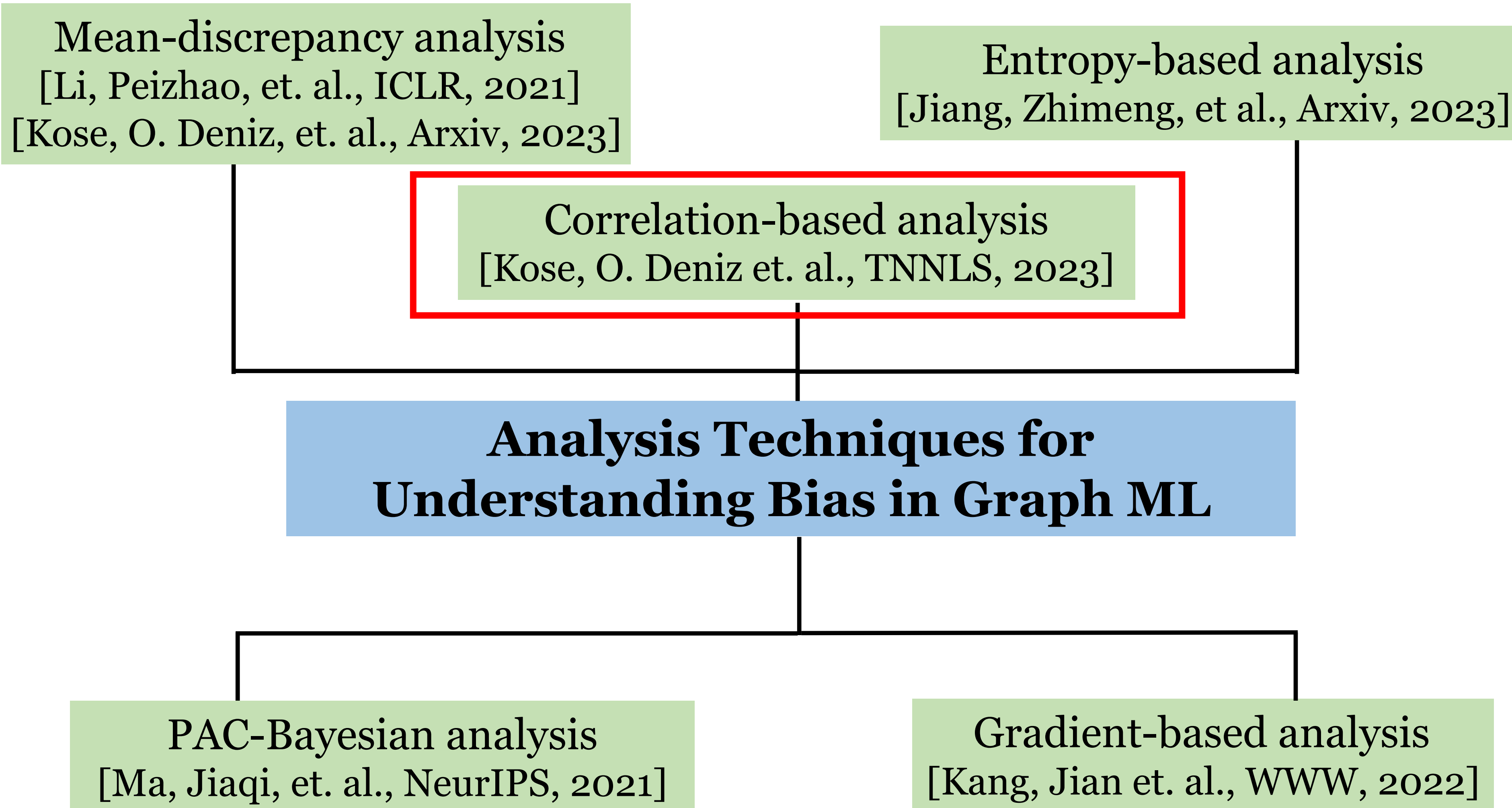


- Reduce $\left|\left(R_1^{\chi}\alpha^{\chi} + R_0^{\chi}\alpha^{\chi} - 1\right)\right|$ for a tighter upper bound
  - FairGAT [2] provides a novel attention strategy that minimizes this term

$$\boxed{\begin{aligned} \min_{\alpha^{\chi}} \quad & \left|R_1^{\chi}\alpha^{\chi} + R_0^{\chi}\alpha^{\chi} - 1\right| \\ \text{s.t.} \quad & 0 \leq \alpha^{\chi} \leq 1 \end{aligned}}$$

[2] Kose, O. Deniz, et al. "FairGAT: Fairness-aware graph attention networks." Arxiv, 2023.

Overview



Mean-discrepancy analysis
[Li, Peizhao, et. al., ICLR, 2021]
[Kose, O. Deniz, et. al., Arxiv, 2023]
Entropy-based analysis
[Jiang, Zhimeng, et al., Arxiv, 2023]
Correlation-based analysis
[Kose, O. Deniz et. al., TNNLS, 2023]
Analysis Techniques for Understanding Bias in Graph ML
PAC-Bayesian analysis
[Ma, Jiaqi, et. al., NeurIPS, 2021]
Gradient-based analysis
[Kang, Jian et. al., WWW, 2022]

Correlation-based Analysis [1]

- Correlation between features and sensitive attributes leads to bias

- More problematic for graphs
  - Generally, the neighbors share the same sensitive attribute
  - Information aggregation among neighbors $\longrightarrow$ Indirect use of sensitive attributes in learning!
  - Aggregated representations are correlated with sensitive attributes

- **Bias measure [1]:** Correlation between aggregated representations, $\mathbf{Z}^l = \mathbf{D}^{-1}(\mathbf{A}+\mathbf{I})\mathbf{H}^{l-1}$, and sensitive attributes $\mathbf{s}$

Degree matrix

Input representations at the *l*th layer

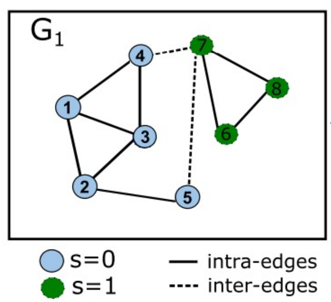
[1] Kose, O. Deniz, et al. "Demystifying and mitigating bias for node representation learning." In TNNLS, 2023.

Correlation-based Analysis: Intuitions

- Factors of bias amplification
  - Distributions of nodal features from different sensitive groups

$$\boldsymbol{\mu}_j := \mathbb{E}_{\mathbf{h}_n \sim U}\left[\overbrace{\mathbf{x}_n}^{\text{features for node n}} \mid n \in \overbrace{\mathcal{S}_j}^{\text{set of nodes with sensitive attribute j}}\right], \quad j = \{0, 1\}$$

  - Node distribution



$$\mathcal{S}_0^{\chi} = \{4, 5\}$$
$$\mathcal{S}_1^{\chi} = \{7\}$$

  - Edge distribution

- Graph data augmentations on input graph

[1] Kose, O. Deniz, et al. "Demystifying and mitigating bias for node representation learning." In TNNLS, 2023.

Correlation for Mean Aggregation

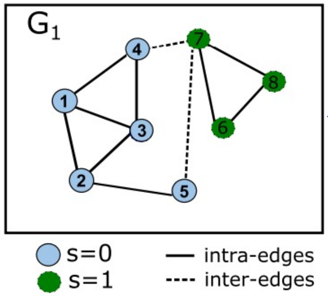
- **Approach [1]:** Bound $||\boldsymbol{\rho}||_1$ with $\rho_i = \mathrm{Corr}(\underbrace{\mathbf{z}_{:,i}}_{\text{i-th aggregated feature}}, \mathbf{s})$ for $i = \{1 \cdots F\}$

- **Theorem [1]:** $||\boldsymbol{\rho}||_1 \leq ||\mathbf{c}||_1(||\boldsymbol{\delta}||_1 \max(\gamma_1, \gamma_2) + 2N\Delta)$

  - $\boldsymbol{\delta} := \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$

  - $$\gamma_1 := \left|1 - \frac{\overbrace{|\mathcal{S}_0^\chi|}^{\text{nodes with at least one inter-edge}}}{|\mathcal{S}_0|} - \frac{|\mathcal{S}_1^\chi|}{|S_1|}\right|$$

  - $$\gamma_2 = \left|1 - 2\min\left(\mathrm{mean}\left(\frac{\overbrace{d_m^\chi}^{\text{inter degree of node m}}}{d_m^\chi + \underbrace{d_m^\omega}_{\text{intra degree of node m}}}\middle|v_m \in \mathcal{S}_0\right), \mathrm{mean}\left(\frac{d_n^\chi}{d_n^\chi + d_n^\omega}\middle|v_n \in \mathcal{S}_1\right)\right)\right|$$



$$\mathcal{S}_0^\chi = \{4, 5\}$$
$$\mathcal{S}_1^\chi = \{7\}$$

[1] Kose, O. Deniz, et al. “Demystifying and mitigating bias for node representation learning.” In TNNLS, 2023.

Guidelines from Correlation Analysis

- Terms, $||\boldsymbol{\delta}_1||, \gamma_1, \gamma_2$, all depend on the input graph structure and nodal features
  - Design augmentation on input graph to reduce these terms → Decrease upper bound on correlation
  - Optimal, fair augmentation strategies to lower bias terms [1]

$$\boldsymbol{\delta} := \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \qquad \boldsymbol{\mu}_j := \mathbb{E}_{\mathbf{h}_n \sim U}\left[\overbrace{\mathbf{x}_n}^{\text{features of node n}} \mid n \in \overbrace{\mathcal{S}_j}^{\text{set of nodes with sensitive attribute j}}\right], \quad j = \{0, 1\}$$

Nodal feature augmentation

$$\gamma_1 := \left| 1 - \frac{\overbrace{|\mathcal{S}_0^{\chi}|}^{\text{at least one inter-edge}}}{|\mathcal{S}_0|} - \frac{|\mathcal{S}_1^{\chi}|}{|S_1|} \right|$$

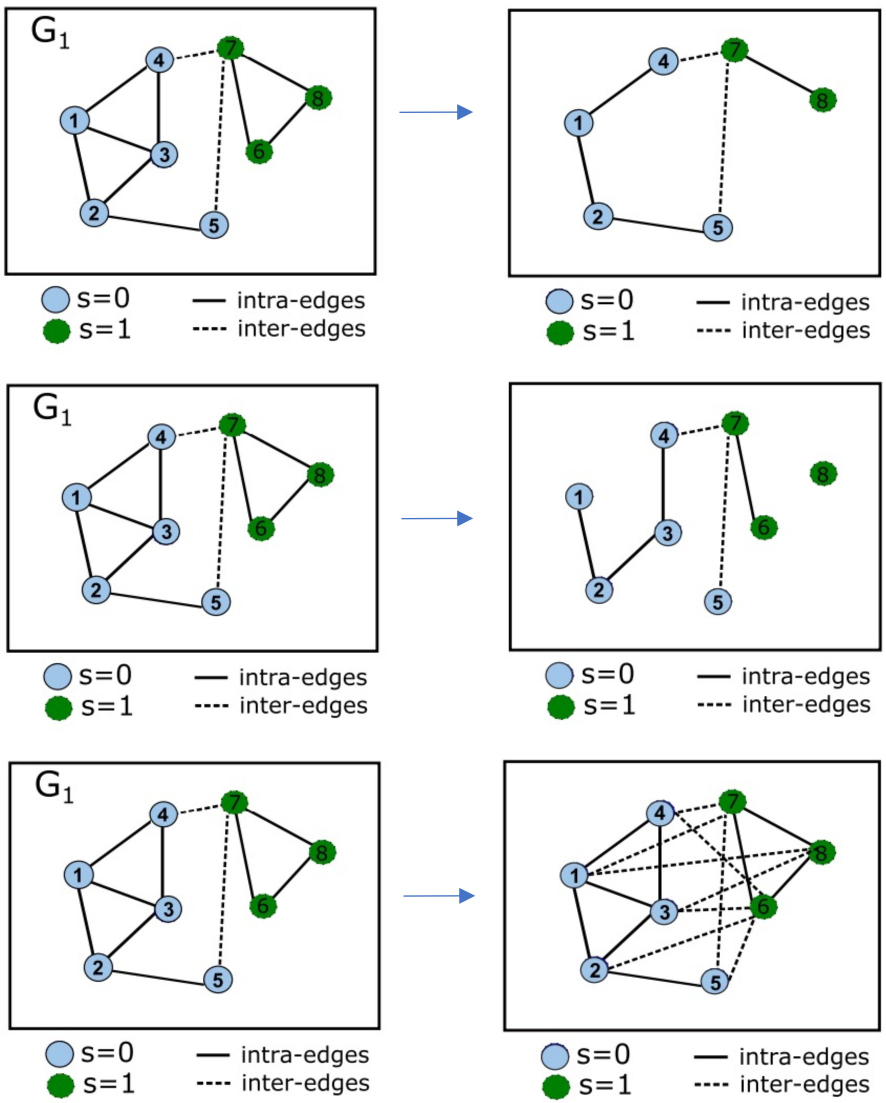
Augmentation on nodes

$$\gamma_2 = \left| 1 - 2\min\left(\operatorname{mean}\left(\frac{\overbrace{d_m^{\chi}}^{\text{inter degree of node m}}}{d_m^{\chi} + \underbrace{d_m^{\omega}}_{\text{intra degree of node m}}} \middle| v_m \in \mathcal{S}_0\right), \operatorname{mean}\left(\frac{d_n^{\chi}}{d_n^{\chi} + d_n^{\omega}} \middle| v_n \in \mathcal{S}_1\right)\right)\right|$$

edge augmentation

[1] Kose, O. Deniz, et al. "Demystifying and mitigating bias for node representation learning." In TNNLS, 2023.

Augmented Graph Examples



Node sampling reducing γ_1

Edge deletion reducing γ_2

Edge addition reducing γ_2

[1] Kose, O. Deniz, et al. "Demystifying and mitigating bias for node representation learning." In TNNLS, 2023.

Overview

Mean-discrepancy analysis
[Li, Peizhao, et. al., ICLR, 2021]
[Kose, O. Deniz, et. al., Arxiv, 2023]

Entropy-based analysis
[Jiang, Zhimeng, et al., Arxiv, 2023]

Correlation-based analysis
[Kose, O. Deniz et. al., TNNLS, 2023]

Analysis Techniques for Understanding Bias in Graph ML

PAC-Bayesian analysis
[Ma, Jiaqi, et. al., NeurIPS, 2021]

Gradient-based analysis
[Kang, Jian et. al., WWW, 2022]

Entropy-based Analysis [1]

- Bias measure: mutual information between node representations and sensitive attributes

- Idea: examine the change in mutual information before/after mean aggregation over graph structure
  - Identify the factors increases mutual information

- Mutual information is intractable to estimate
  - [1] upper bounds mutual information, where the bound is used as bias measure

$$I(\mathbf{s}, \mathbf{X}) \leq -(1-c)\ln\left[(1-c) + c\exp\left(-D_{KL}\left(P_1 \| P_2\right)\right)\right]$$
$$-c\ln\left[c + (1-c)\exp\left(-D_{KL}\left(P_2 \| P_1\right)\right)\right] \triangleq \mathrm{Bias}(\mathbf{s}, \mathbf{X})$$

KL-divergence

nodal features following a GMM distribution

$$c = \mathbb{E}_i\left[\mathbb{P}\left(s_i = 1\right)\right]$$
$$P_1 \triangleq f_{\mathbf{X}}\left(\mathbf{X}_i = \mathbf{x} \mid s_i = -1\right) \sim \mathcal{N}\left(\mu_1, \mathbf{\Sigma}_1\right)$$
$$P_2 \triangleq f_{\mathbf{X}}\left(\mathbf{X}_i = \mathbf{x} \mid s_i = 1\right) \sim \mathcal{N}\left(\mu_2, \mathbf{\Sigma}_2\right)$$

[1] Jiang, Zhimeng, et al. "Topology matters in fair graph learning: A theoretical pilot study." Arxiv, 2023.

Entropy-based Analysis: Intuitions

- Bias amplifying factors in graph-based aggregation
  - Node number
  - Density of graph connectivity $\rho_d = \mathbb{E}_{ij}\left[\mathbb{P}\left(\mathbf{A}_{ij} = 1\right)\right]$
  - Sensitive attribute homophily coefficient $\epsilon_{\text{sens}} = \mathbb{P}\left(s_i = s_j \mid \mathbf{A}_{ij} = 1\right)$
- Modify graph structure based on these factors

[1] Jiang, Zhimeng, et al. "Topology matters in fair graph learning: A theoretical pilot study." Arxiv, 2023.

Guidelines from Entropy-based Analysis

- Theorem [1]: In mean-aggregation over graph topology, bias increases if $\boxed{(v_1 - v_2)^2 \min\{\beta_1, \beta_2\} > 1}$

$$v_1 = \frac{(N_1-1)p_{\text{intra}}+1}{\beta_1} \qquad p_{\text{intra}} = \frac{\rho_d \epsilon_{\text{sens}}}{c^2+(1-c)^2}$$

probability of intra-edges

$$v_2 = \frac{(N_1-1)p_{inter}}{\beta_2} \qquad p_{\text{inter}} = \frac{\rho_d(1-\epsilon_{\text{sens}})}{2c(1-c)}$$

probability of inter-edges

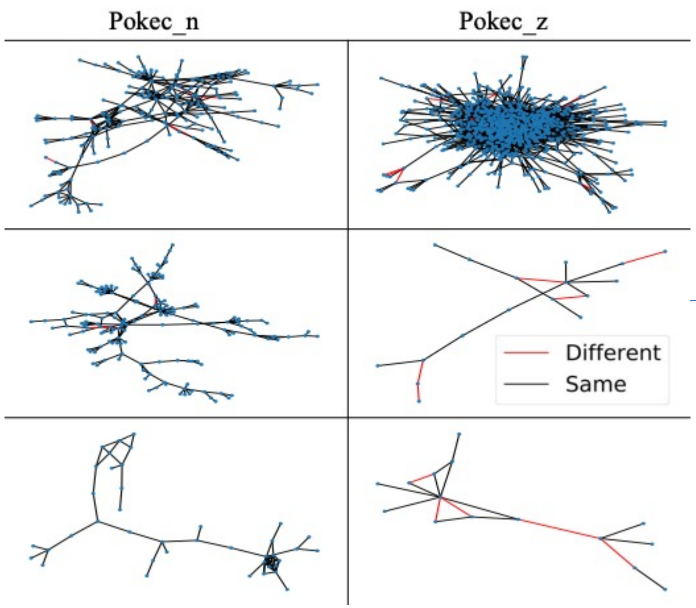
$$\beta_1 = N_{-1}p_{\text{inter}} + (N_1 - 1)\,p_{\text{intra}} + 1 \qquad N_1 = Nc$$

$$\beta_2 = N_{-1}p_{intra} + (N_1 - 1)\,p_{inter} + 1 \qquad N_{-1} = N(1-c)$$

- Bias enlarges as node number, N, increases
- Denser graph connectivity, higher ρ_d, increases bias
- For extremely large or small sensitive attribute homophily coefficient, i.e. ε_{sens} ->1/0, bias increases!

[1] Jiang, Zhimeng, et al. "Topology matters in fair graph learning: A theoretical pilot study." Arxiv, 2023.

Edge Distribution of Real-World Networks



Number of intra-edges is significantly larger compared to number of inter-edges

- For real-world networks, $\varepsilon_{sens} \rightarrow 1$, which leads to enhanced bias based on entropy analysis!
- Balanced inter- and intra-edges is critical for real-world applications!

[1] Jiang, Zhimeng, et al. "Topology matters in fair graph learning: A theoretical pilot study." Arxiv, 2023.

Overview

Mean-discrepancy analysis
[Li, Peizhao, et. al., ICLR, 2021]
[Kose, O. Deniz, et. al., Arxiv, 2023]

Entropy-based analysis
[Jiang, Zhimeng, et al., Arxiv, 2023]

Correlation-based analysis
[Kose, O. Deniz et. al., TNNLS, 2023]

Analysis Techniques for Understanding Bias in Graph ML

PAC-Bayesian analysis
[Ma, Jiaqi, et. al., NeurIPS, 2021]

Gradient-based analysis
[Kang, Jian et. al., WWW, 2022]

PAC-Bayesian Analysis [1]

- [1] derives a PAC-Bayesian analysis for the generalization ability of GNNs on node-level tasks with non-IID assumptions

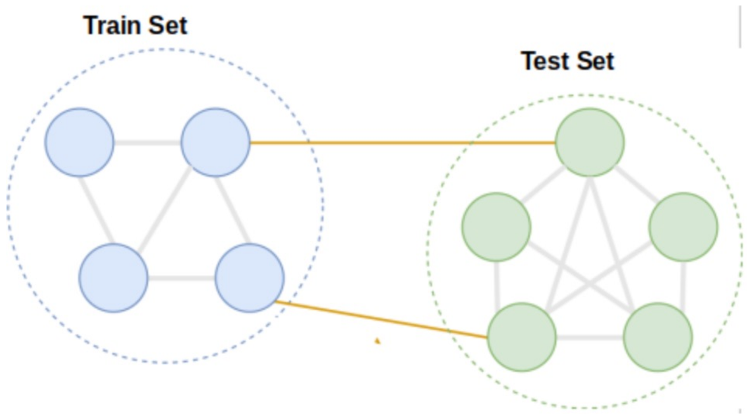
- Bias measure: Accuracy disparity on test set between different sensitive groups

[1] Ma, Jiaqi, et al. "Subgroup Generalization and Fairness of Graph Neural Networks." In NeurIPS, 2021.

PAC-Bayesian Analysis: Intuitions

- Generalization of trained model on a subset of test nodes is related to geodesic distance between training and test nodes

- Selection of training set for a similar generalizability on each group in test set



[1] Ma, Jiaqi, et al. "Subgroup Generalization and Fairness of Graph Neural Networks." In NeurIPS, 2021.

Subgroup Generalization Bound for GNNs

- Theorem [1]:

$$\mathcal{L}_m^0(\tilde{h}) \le \widehat{\mathcal{L}}_{tr}^{\gamma}(\tilde{h}) + O\left(cK\epsilon_m + \frac{b\sum_{l=1}^{L}\left\|\widetilde{W}_l\right\|_F^2}{(\gamma/8)^{2/L}|\mathcal{S}_0|^{\alpha}}(\epsilon_m)^{2/L} + \frac{1}{|\mathcal{S}_0|^{1-2\alpha}} + \frac{1}{|\mathcal{S}_0|^{2\alpha}}\ln\frac{LC(2B_m)^{1/L}}{\gamma^{1/L}\delta}\right)$$

- Upper bound on the generalization error of trained classifier for any subgroup in terms of data distribution- and model-related parameters
- ε_m is useful for a fairness-aware training data selection

$$\epsilon_m := \max_{j\in\mathcal{S}_m} \min_{i\in\mathcal{S}_{tr}} \left\|\mathbf{z}_i - \mathbf{z}_j\right\|_2$$

Distance to training set

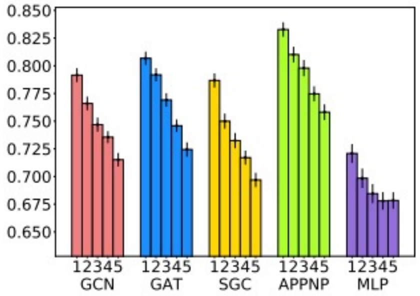
aggregated representation for node i

- There is a better generalization guarantee for subgroups that are closer to the training set
  - Bias for subgroups that are far away from the training set

[1] Ma, Jiaqi, et al. "Subgroup Generalization and Fairness of Graph Neural Networks." In NeurIPS, 2021.

Training Set Matters

- Geodesic distance (length of the shortest path) between two nodes is positively related to corresponding ε_m
  - Test nodes with larger geodesic distance to the training set tend to suffer from lower accuracy



Bars labeled 1 to 5 illustrate test accuracy for subgroups with increasing geodesic distance to training set.

- Selection of training set plays an important role on fairness
  - An unevenly selected training set, leaving part of the test nodes far away, may cause a large accuracy disparity
  - Can guide a fairness-aware training set selection strategy

[1] Ma, Jiaqi, et al. "Subgroup Generalization and Fairness of Graph Neural Networks." In NeurIPS, 2021.

Overview

Mean-discrepancy analysis
[Li, Peizhao, et. al., ICLR, 2021]
[Kose, O. Deniz, et. al., Arxiv, 2023]

Entropy-based analysis
[Jiang, Zhimeng, et al., Arxiv, 2023]

Correlation-based analysis
[Kose, O. Deniz et. al., TNNLS, 2023]

Analysis Techniques for Understanding Bias in Graph ML

PAC-Bayesian analysis
[Ma, Jiaqi, et. al., NeurIPS, 2021]

Gradient-based analysis
[Kang, Jian et. al., WWW, 2022]

Gradient-based Analysis [1]

- Rawlsian Difference Principle: achieves equality by maximizing the welfare of the worst-off groups

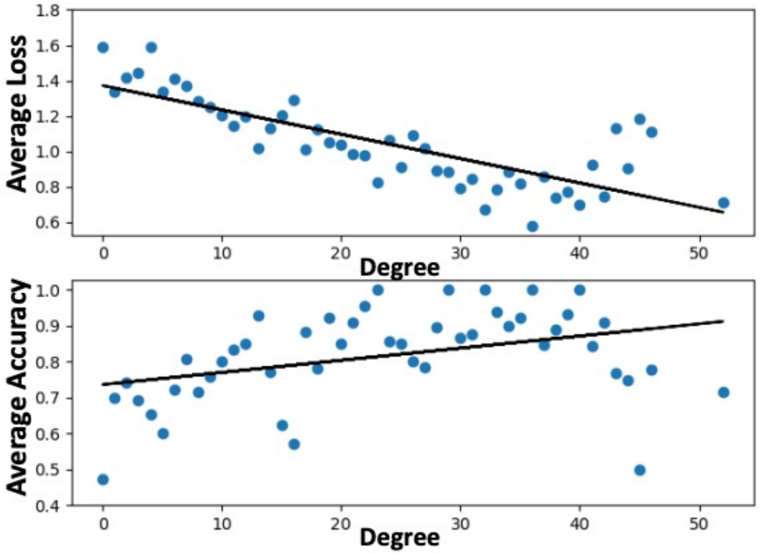
- Bias measure [1]: Variance of losses corresponding to different sensitive groups
  - A mathematical formulation for Rawlsian principle

- Analysis technique: find root cause of bias by analyzing mathematically gradient of loss wrt weight parameters
  - Key component in training is gradient

[1] Kang, Jian, et al. "RawlsGCN: Towards Rawlsian Difference Principle on Graph Convolutional Networks." In WWW, 2022.

Degree-related Bias

- [1] focuses on degree-related bias
  - GNN is often biased towards benefiting high-degree nodes



- Question: why the loss of a GNN varies among nodes with different degrees after training?

[1] Kang, Jian, et al. "RawlsGCN: Towards Rawlsian Difference Principle on Graph Convolutional Networks." In WWW, 2022.

Gradient-based Analysis: Intuitions

- Degree of a node effects its importance on the updates of weights during training

- Can be solved by equalizing each degree to a constant
  - Normalized adjacency matrix

[1] Kang, Jian, et al. "RawlsGCN: Towards Rawlsian Difference Principle on Graph Convolutional Networks." In WWW, 2022.

Gradient Analysis for GNNs

gradient of loss wrt weight matrix of l-th GNN layer

Theorem [1]: $\frac{\partial J}{\partial \mathbf{W}^{(l)}} = \sum_{j=1}^{n} \deg(j)\mathbf{I}_j^{(row)} = \sum_{i=1}^{n} \deg(i)\mathbf{I}_i^{(col)}$

$$\mathbf{I}_j^{(\mathrm{row})} = \left(\mathbf{H}^{(l-1)}[j,:]\right)^T \mathbb{E}_{i\sim p_{\mathcal{N}}(j)}\left[\frac{\partial J}{\partial \mathbf{E}^{(l)}[i,:]}\right]$$

row-wise influence matrix of node j

Input node representations to l-th GNN layer

$$\mathbf{I}_i^{(col)} = \left(\mathbb{E}_{j\sim p_{\mathcal{N}}(i)}\left[\mathbf{H}^{(l-1)}[j,:]\right]\right)^T \frac{\partial J}{\partial \mathbf{E}^{(l)}[i,]}$$

column-wise influence matrix of node i

Probability distribution on the neighborhood of node i

$$\mathbf{E}^{(l)} = \hat{\mathbf{A}}\mathbf{H}^{(l-1)}\mathbf{W}^{(l)}$$

$$\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\tilde{\mathbf{D}}^{-\frac{1}{2}}$$

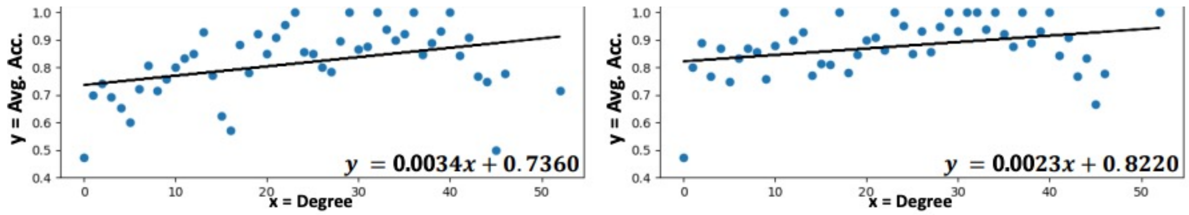
[1] Kang, Jian, et al. “RawlsGCN: Towards Rawlsian Difference Principle on Graph Convolutional Networks.” In WWW, 2022.

Guidelines for Degree-Bias Mitigation

Theorem [1]: $\frac{\partial J}{\partial \mathbf{W}^{(l)}} = \sum_{j=1}^{n} \deg(j)\mathbf{I}_j^{(row)} = \sum_{i=1}^{n} \deg(i)\mathbf{I}_i^{(col)}$

- This analysis shows that node degrees serve as importance scores of node influence matrices for corresponding gradient
  - Higher node degree implies more importance on the gradient
  - Provides explainability for degree-related bias

- Solution: Normalize adjacency matrix such that each node has equal importance in updating the weight parameters
  - Rows and columns of **A** should sum up to a constant
  - [1] employs an iterative algorithm (Sinkhorn-Knopp algorithm) to balance **A** for a constant degree for each node



(a) GCN (b) RawlsGCN-Graph

[1] Kang, Jian, et al. "RawlsGCN: Towards Rawlsian Difference Principle on Graph Convolutional Networks." In WWW, 2022.

Conclusions

- Multiple studies demonstrate the sources of bias via following different analysis techniques
  - Improves explainability aspect of fairness-aware ML on graphs
  - Essential for large-scale deployment of learning algorithms

- Provided theoretical results can guide novel fairness-aware algorithm design

- **Open problems**
  - Multiple, non-binary sensitive attributes
  - Different aggregation schemes
  - Less restrictive assumptions on data distribution

Outline

Background Introduction

Fairness Notions and Metrics

Theoretical Understanding of Bias

Techniques for Fair Node Embeddings

Real-World Applications

Summary, Challenges, & Future Directions

Node Embeddings

- Mappings into a low-dimensional space
  - Protect similarities in network structure & nodal features
- Different approaches based on different similarity definitions
  - SOTA: GNN-based approaches
- Can be employed in several downstream tasks

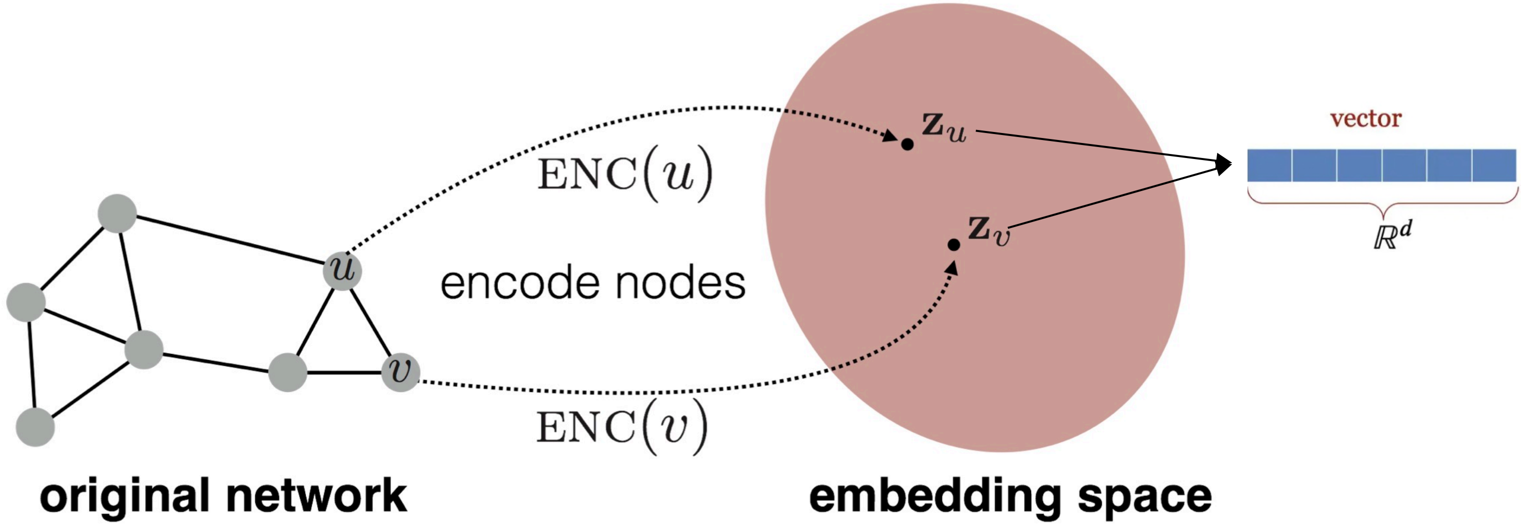
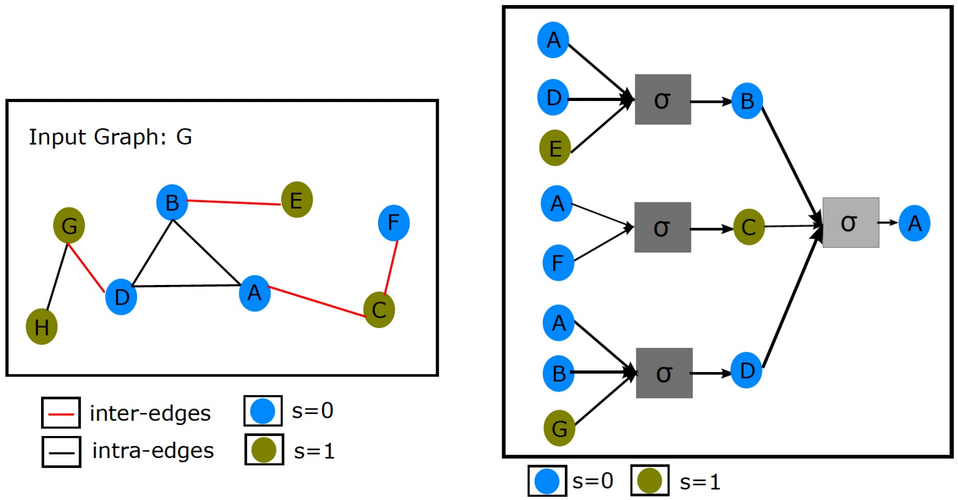


*Figure is modified from snap-stanford.github.io

Bias in Node Embeddings

- Bias in nodal features will be encoded in node embeddings

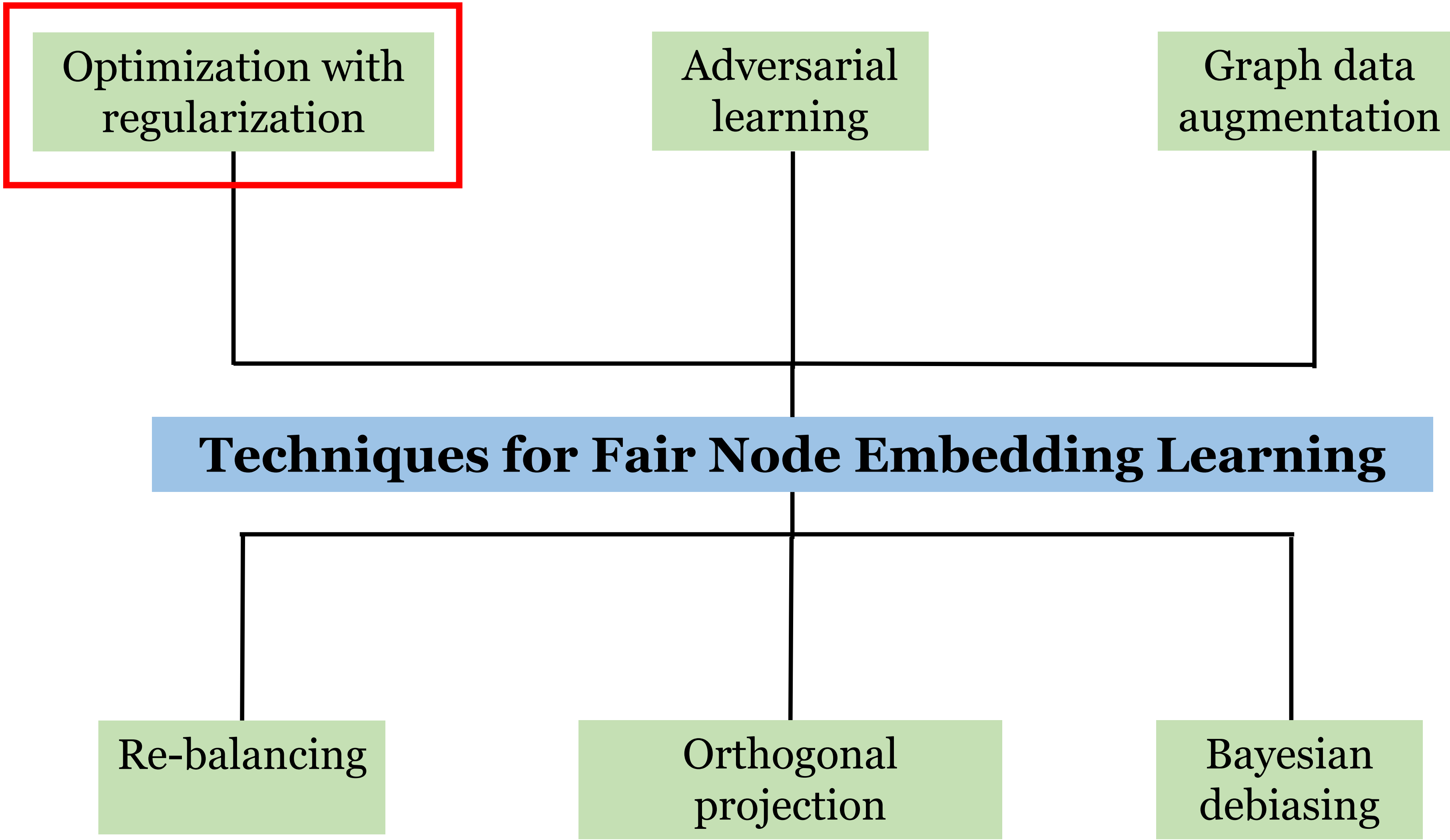
- Aggregate information from neighbors → node embeddings



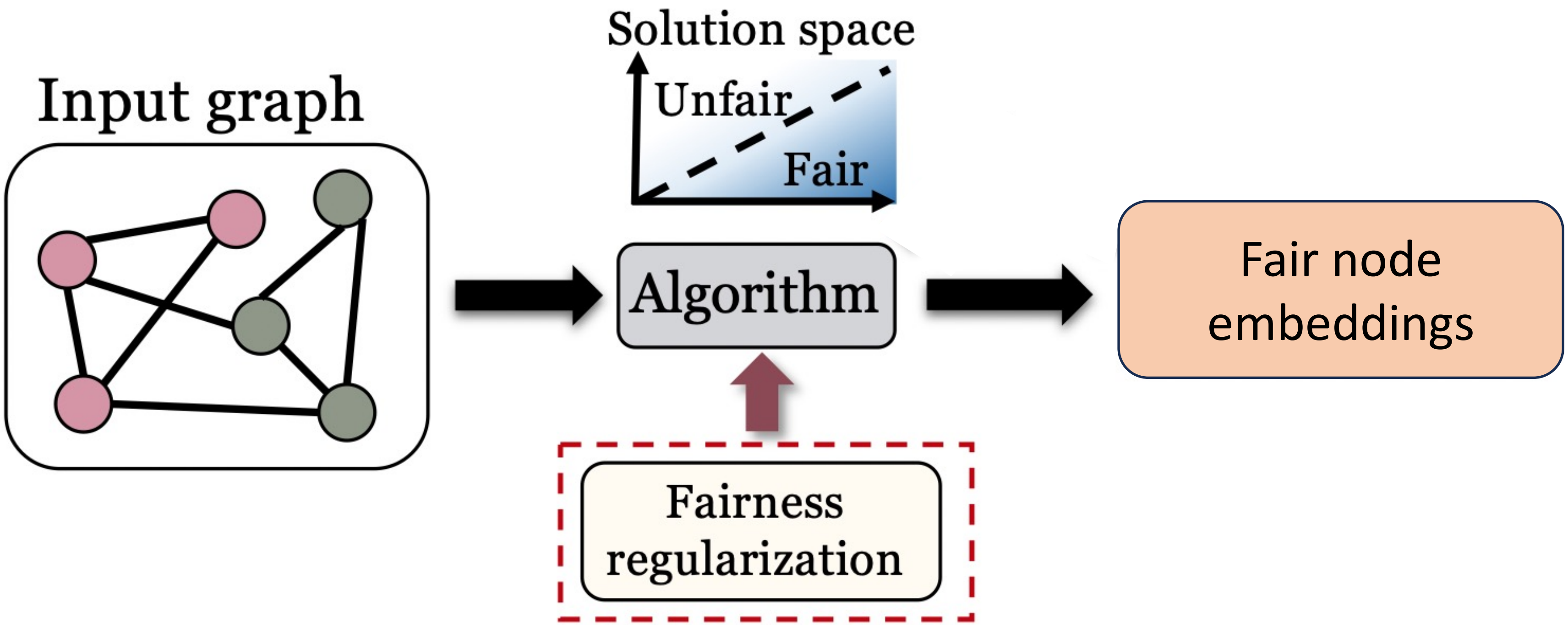
- Neighbors generally with same sensitive attributes
    - Embeddings correlated with sensitive attributes
    - Bias in graph structure propagated towards node embeddings

- Intertwined bias from both nodal features and graph structure

Overview


Optimization with regularization
Adversarial learning
Graph data augmentation
Techniques for Fair Node Embedding Learning
Re-balancing
Orthogonal projection
Bayesian debiasing

Optimization with Regularization



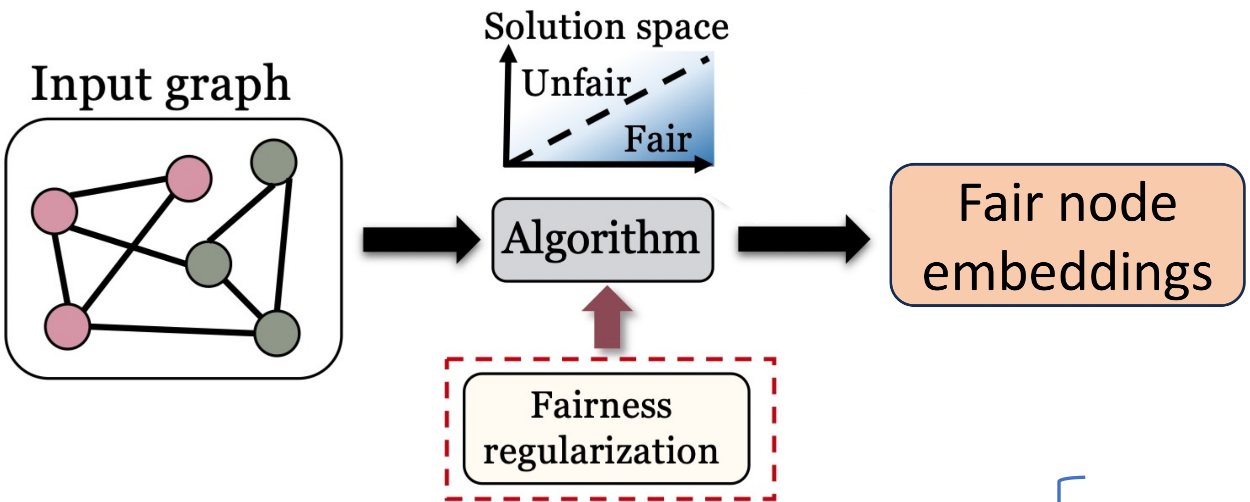
- Bias mitigation:

$$\mathcal{L} = \mathcal{L}_{utility} + \lambda \mathcal{L}_{fair}$$

Similarity of Node Embeddings

Distribution of Node Embeddings

Optimization with Regularization



- Bias mitigation: $\mathcal{L} = \mathcal{L}_{utility} + \lambda \mathcal{L}_{fair}$

Similarity of Node Embeddings

Distribution of Node Embeddings

Node embedding matrix

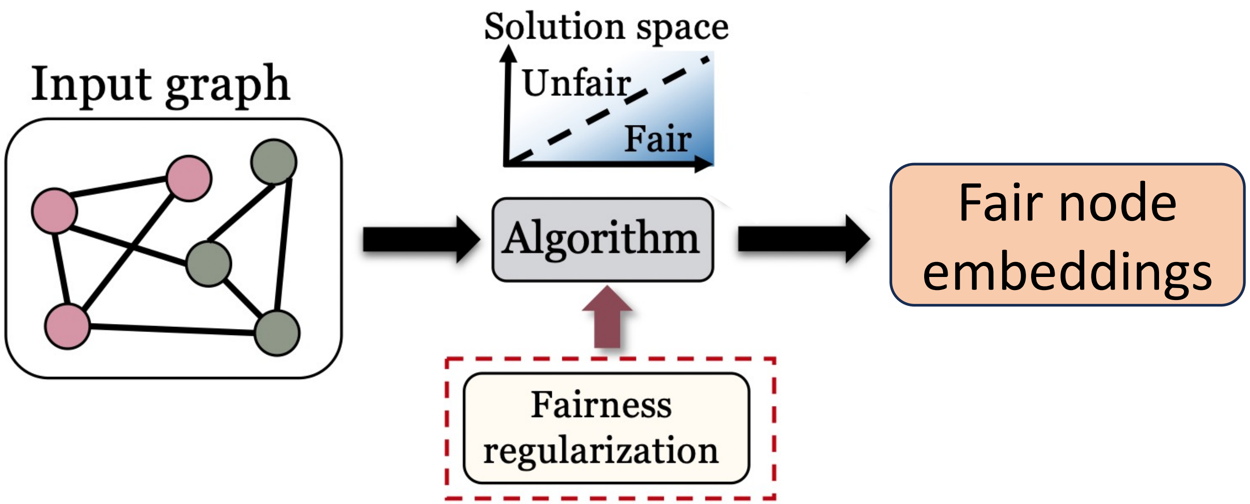
- [1]: $\mathcal{L}_{fair} := \operatorname{Tr}\left(\mathbf{H}^{\top}\mathbf{L}_{\mathbf{S}}\mathbf{H}\right)$

Laplacian matrix of similarity matrix

Individual fairness-based regularization: Similar embeddings for similar nodes

[1] Kang, Jian, et al. "Inform: Individual fairness on graph mining." In SIGKDD, 2020.

Optimization with Regularization



- Bias mitigation: $\mathcal{L} = \mathcal{L}_{utility} + \lambda \mathcal{L}_{fair}$
  - Similarity of Node Embeddings
  - Distribution of Node Embeddings

Cross-entropy loss

Prob. node i is more similar to node j than node m based on node embedding similarity

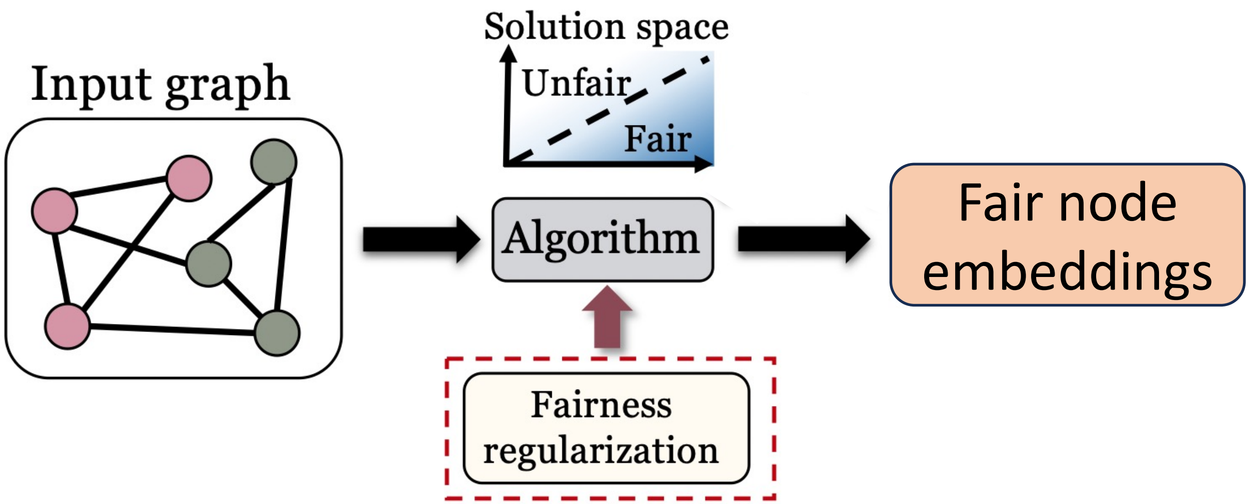
- [1]: $\mathcal{L}_{fair} = \sum_i \sum_{j,m} \mathcal{L}_{CE}(P_{j,m}, \hat{P}_{j,m})$

Ranking based individual fairness: Rankings provided by oracle similarity matrix and node embeddings consistent

Prob. node i is more similar to node j than node m based on oracle similarity matrix

[1] Dong, Yushun, et al. "Individual fairness for graph neural networks: A ranking based approach." In KDD, 2021.

Optimization with Regularization



- Bias mitigation: $\mathcal{L} = \mathcal{L}_{utility} + \lambda \mathcal{L}_{fair}$
  - Similarity of Node Embeddings
  - Distribution of Node Embeddings

sensitive groups

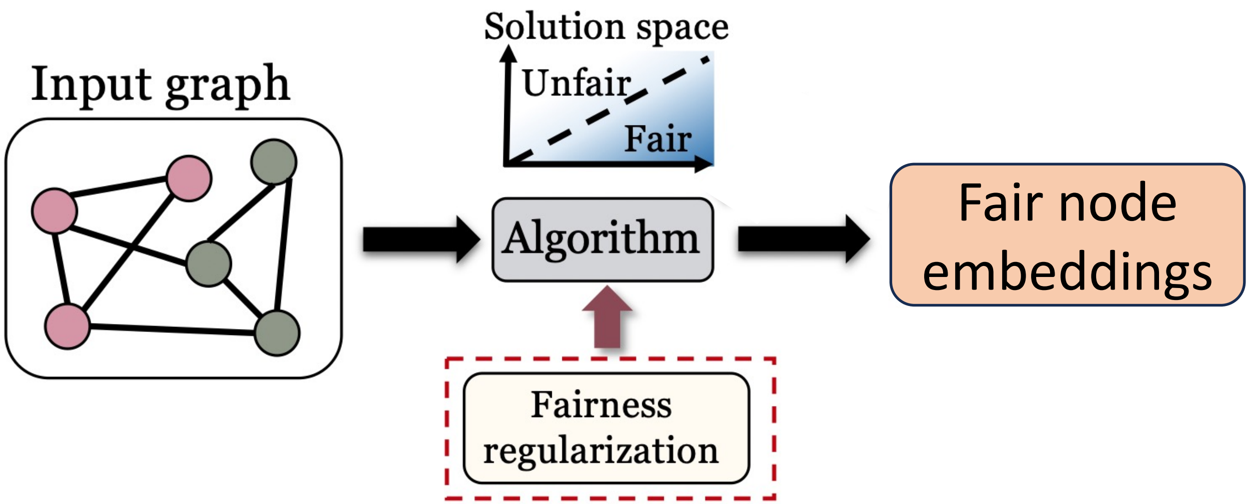
- [1]: $\mathcal{L}_{fair} = \sum_{p,q}^{1 \le p < q \le G} \left(\frac{U_p}{U_q} - 1\right)^2 + \left(\frac{U_q}{U_p} - 1\right)^2$

Group disparity of individual fairness: Similar individual fairness level for each sensitive group

Individual unfairness of embeddings in group q based on Lipschitz condition.

[1] Song, Weihao, et al. "GUIDE: Group Equality Informed Individual Fairness in Graph Neural Networks." In KDD, 2022.

Optimization with Regularization



- Bias mitigation: $\mathcal{L} = \mathcal{L}_{utility} + \lambda \mathcal{L}_{fair}$

Similarity of Node Embeddings

Distribution of Node Embeddings

Distribution of node embeddings from sensitive group 0

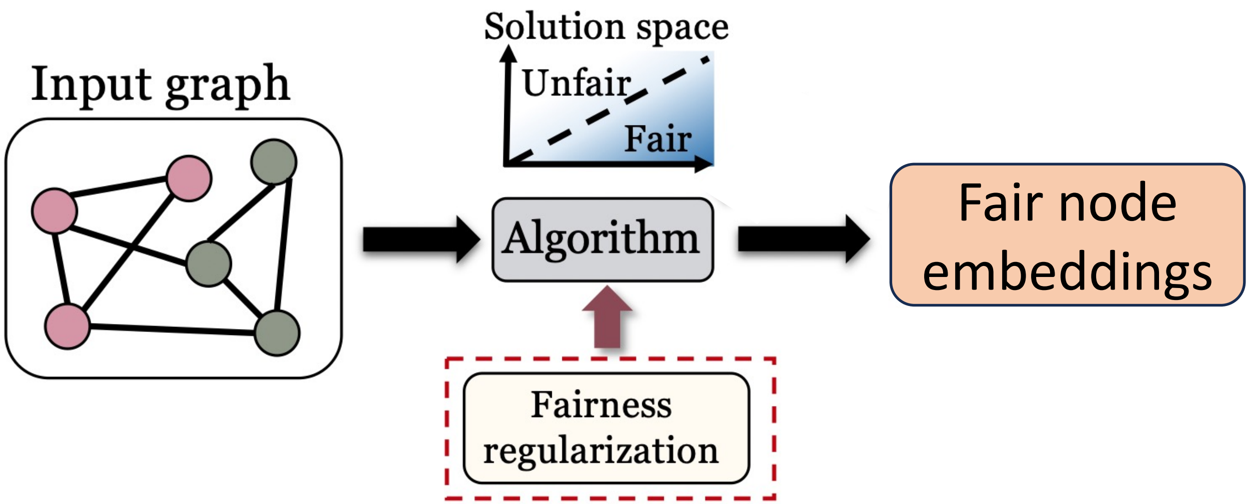
- [1]: $\mathcal{L}_{fair} = D_W(f_0, f_1)$

Squared quadratic Wasserstein distance

Group fairness:
Brings distributions of node embeddings from different sensitive groups closer

[1] Fan, Wei, et al. "Fair Graph Auto-Encoder for Unbiased Graph Representations with Wasserstein Distance." In ICDM, 2021.

Optimization with Regularization



- Bias mitigation: $\mathcal{L} = \mathcal{L}_{utility} + \lambda\mathcal{L}_{fair}$

Similarity of Node Embeddings
Distribution of Node Embeddings

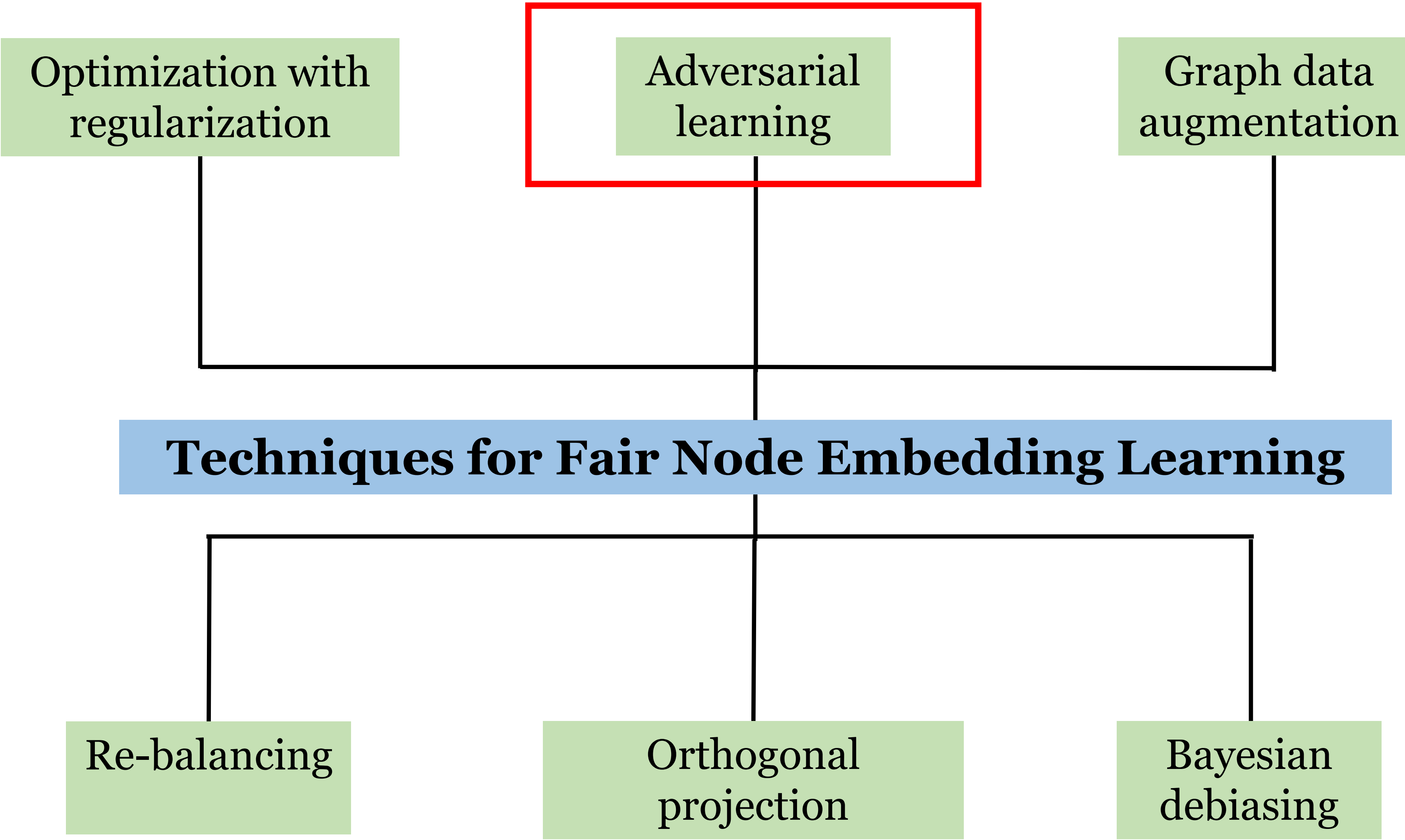
Sample mean of embeddings from sensitive group 0:
Governed by learnable parameters of a novel normalization layer

- [1]: $\mathcal{L}_{fair} = \left\|\boldsymbol{\mu}^{(0)} - \boldsymbol{\mu}^{(1)}\right\|_2^2$

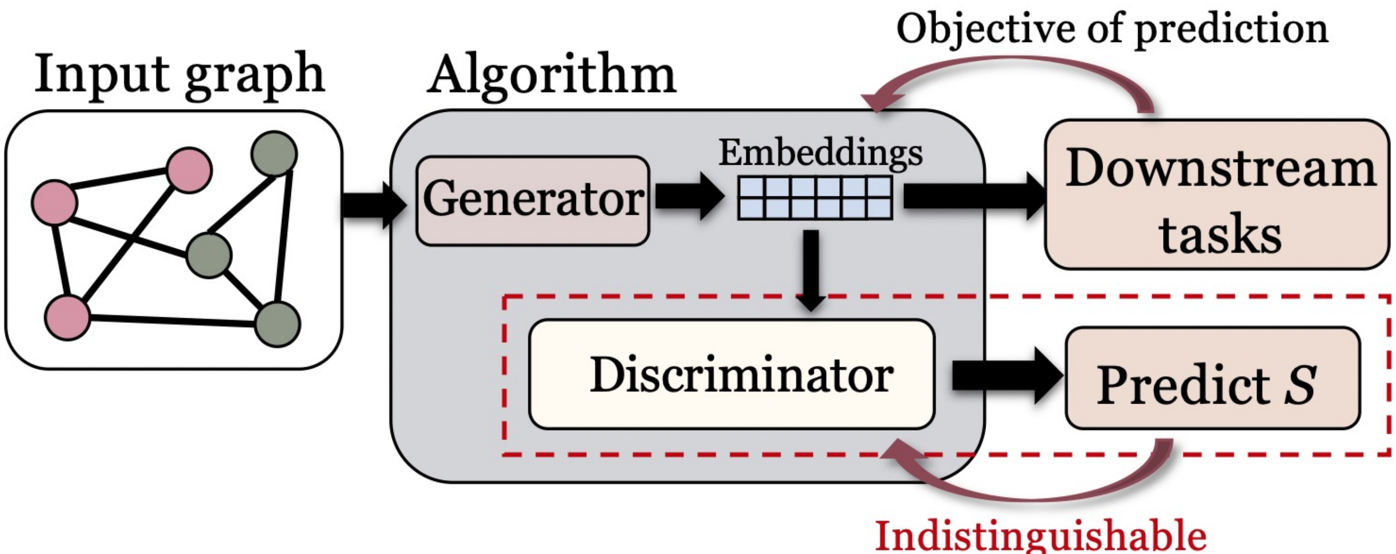
Group fairness:
Brings distributions of node embeddings from different sensitive groups closer

[1] Kose, O. Deniz, et al. "Fast& Fair: Training Acceleration and Bias Mitigation for GNNs." In TMLR, 2023.

Overview


Optimization with regularization
Adversarial learning
Graph data augmentation
Techniques for Fair Node Embedding Learning
Re-balancing
Orthogonal projection
Bayesian debiasing

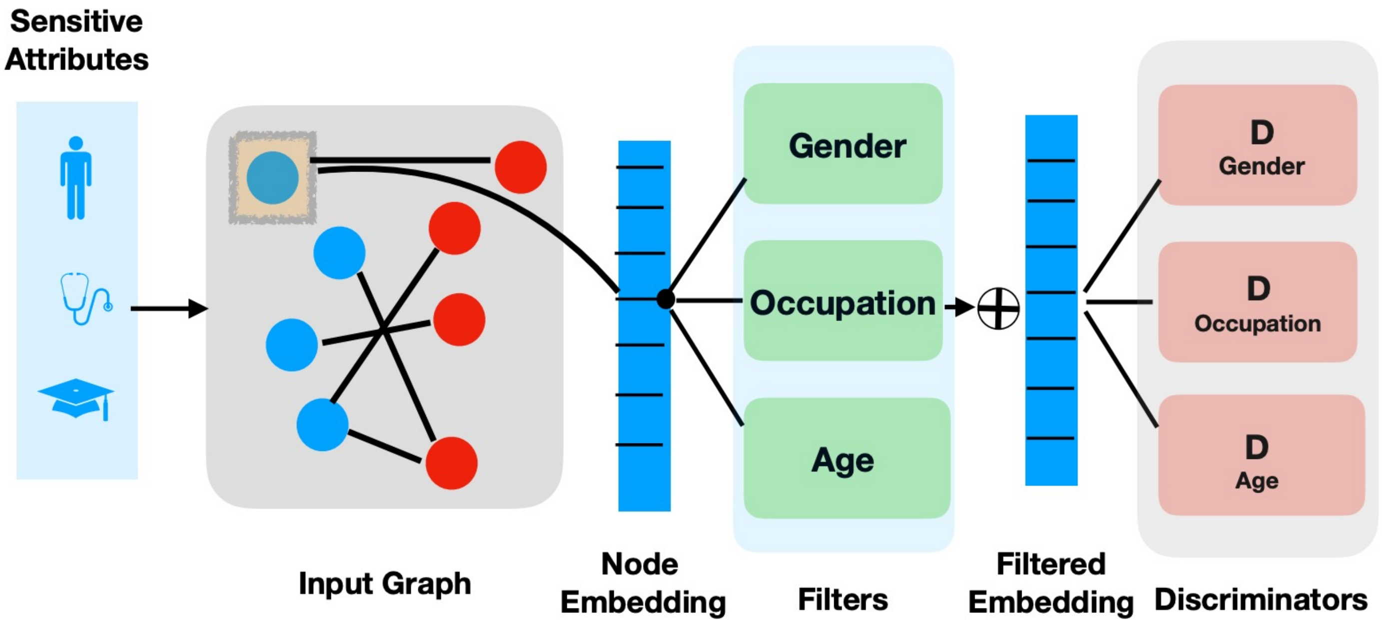
Adversarial Learning



- Two main components:
  - **Generator:** generate node embeddings for downstream tasks
  - **Discriminator:** distinguish the embeddings between demographic subgroups
- Downstream task → node classification [1]

[1] Dai, Enyan, et al. "Say No to the Discrimination: Learning Fair Graph Neural Networks with Limited Sensitive Attribute Information." In WSDM, 2021.

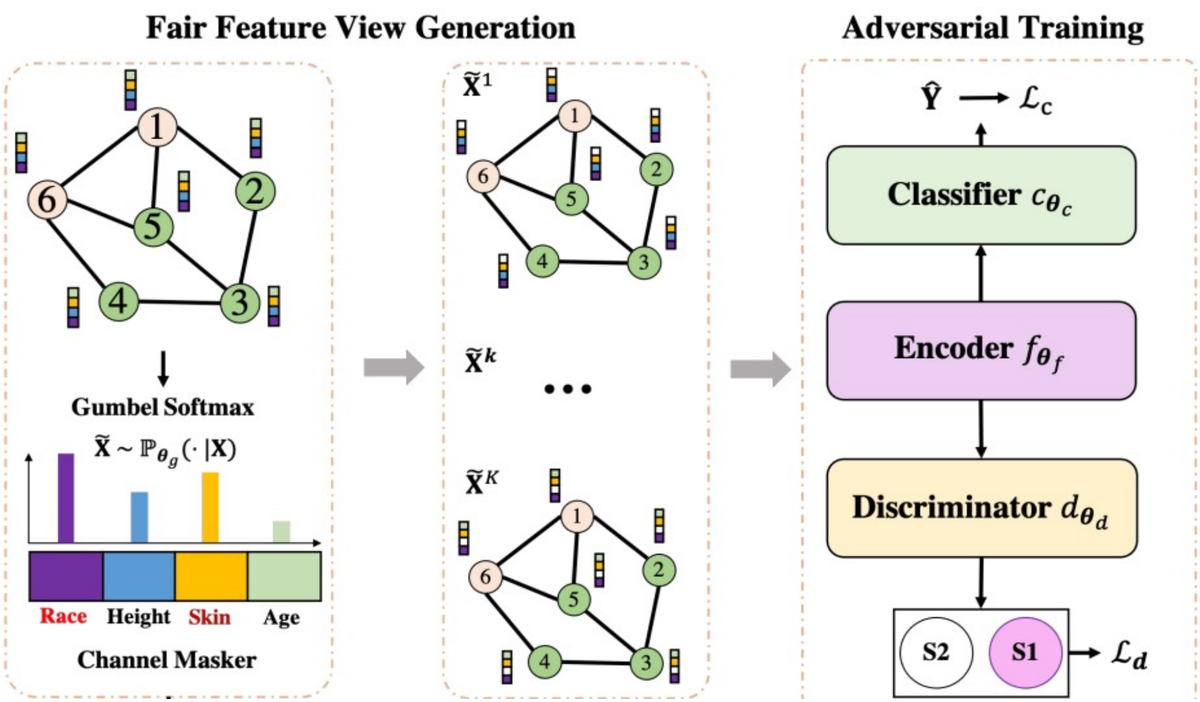
Adversarial Learning



- Two main components:
  - **Generator:** generate node embeddings for downstream tasks
  - **Discriminator:** distinguish the embeddings between demographic subgroups
- [1] considers **multiple** sensitive attributes

[1] Bose, Avishek, et al. "Compositional fairness constraints for graph embeddings." In ICML, 2019.

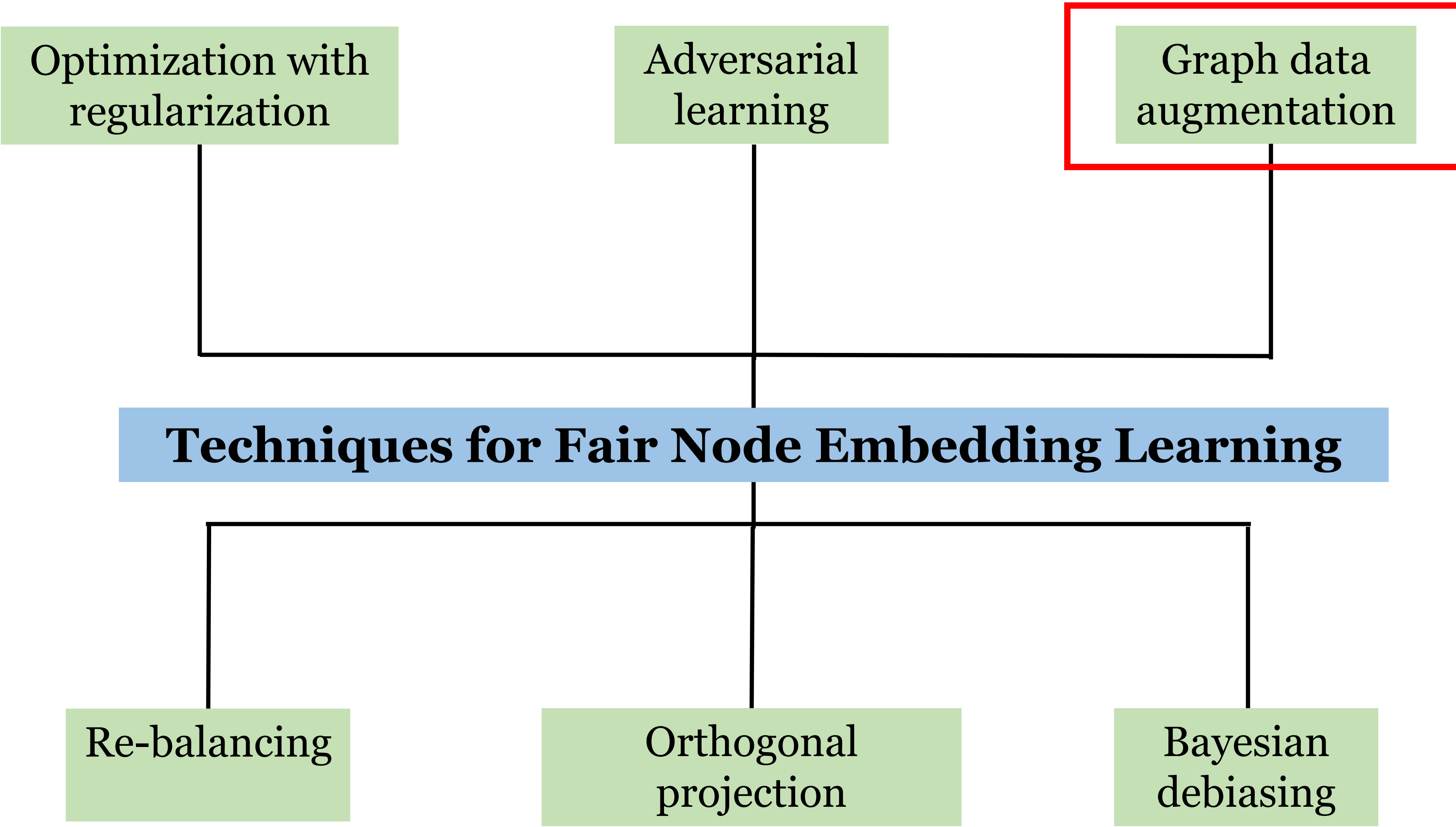
Adversarial Debiasing for Graphs



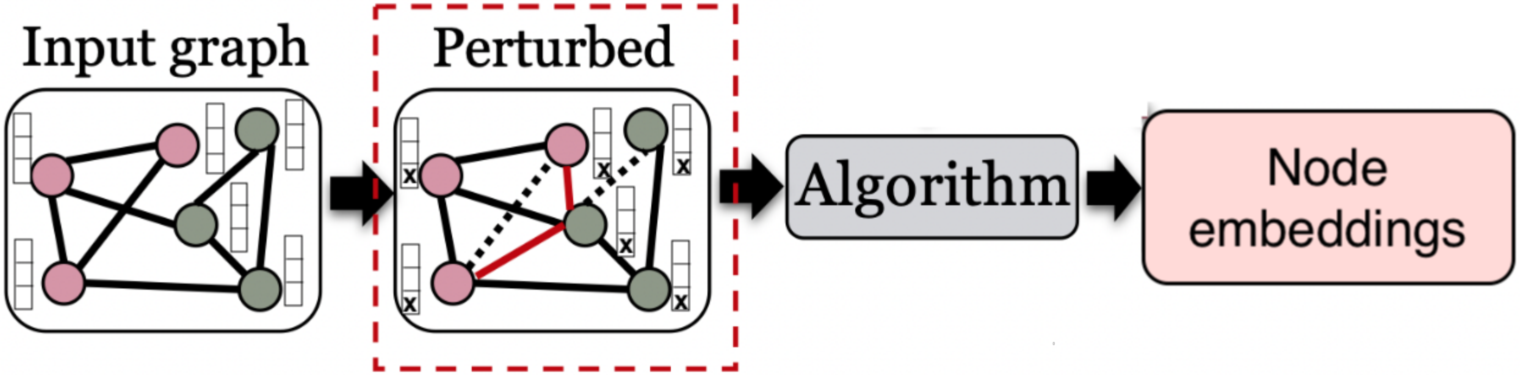
- Via adversarial learning, generate fair views of input graph
  - Generate node embeddings based on fair graph views
- Learn feature masks to prevent sensitive information leakage [1]
- In addition to feature mask, re-wire adjacency matrix [2]

[1] Wang, Yu, et al. "Improving Fairness in Graph Neural Networks via Mitigating Sensitive Attribute Leakage." In KDD, 2022.
[2] Ling, Hongyi, et al. "Learning Fair Graph Representations via Automated Data Augmentations." In ICLR, 2023.

Overview


Optimization with regularization
Adversarial learning
Graph data augmentation
Techniques for Fair Node Embedding Learning
Re-balancing
Orthogonal projection
Bayesian debiasing

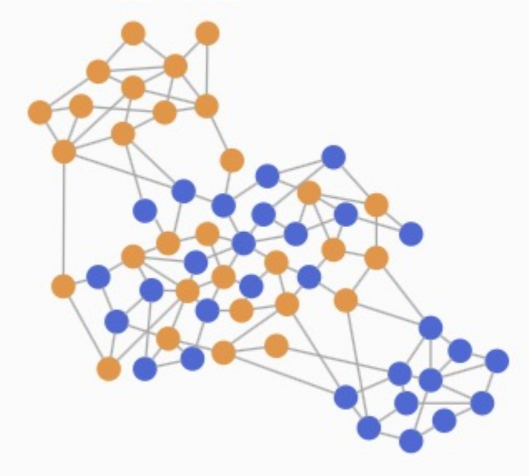
Graph Data Augmentation



- Graph data augmentation: Corrupt graph structure and/or nodal features
    - Introduced for better robustness
    - Can be used to eliminate the bias amplifying factors in graph structure and nodal features
- Input augmented graph for fair node embeddings
    - Hand-crafted, heuristic edge augmentation & feature masking
    - Theory-based augmentation design
    - Automated augmentation
    - Counterfactual fairness-based augmentation design

Observations for Sources of Bias

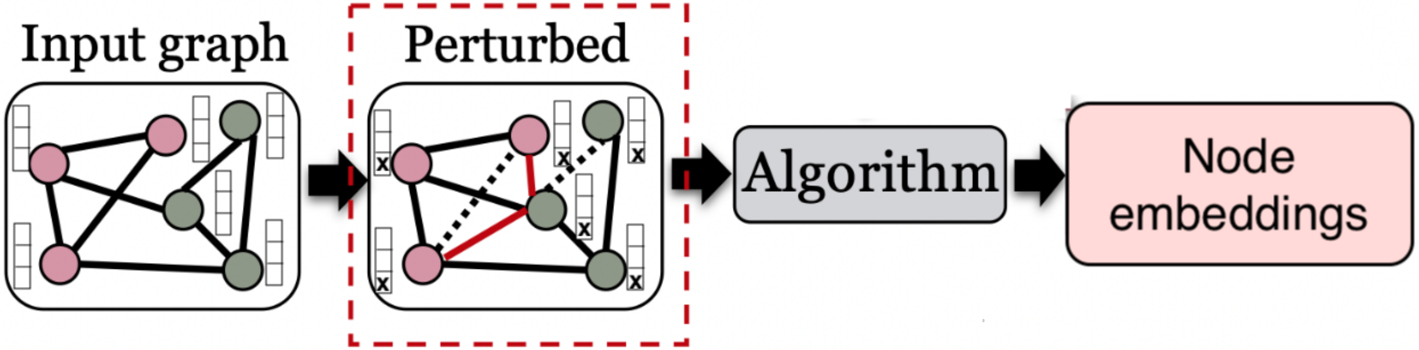
- **Biased graph structure**
  - Clear **community structure** between two groups of nodes with different sensitive attribute (i.e., yellow and blue)



- **Biased nodal features**
  - Features correlated with sensitive attributes lead to **intrinsic bias**

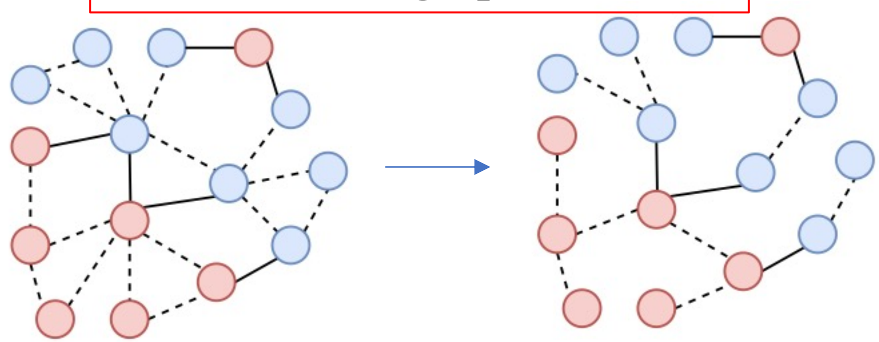
- **Possible solutions**
  - Edge augmentation: Balance inter/intra edges
  - Feature masking: Mask features highly correlated with sensitive attribute

Edge Augmentation for Group Fairness



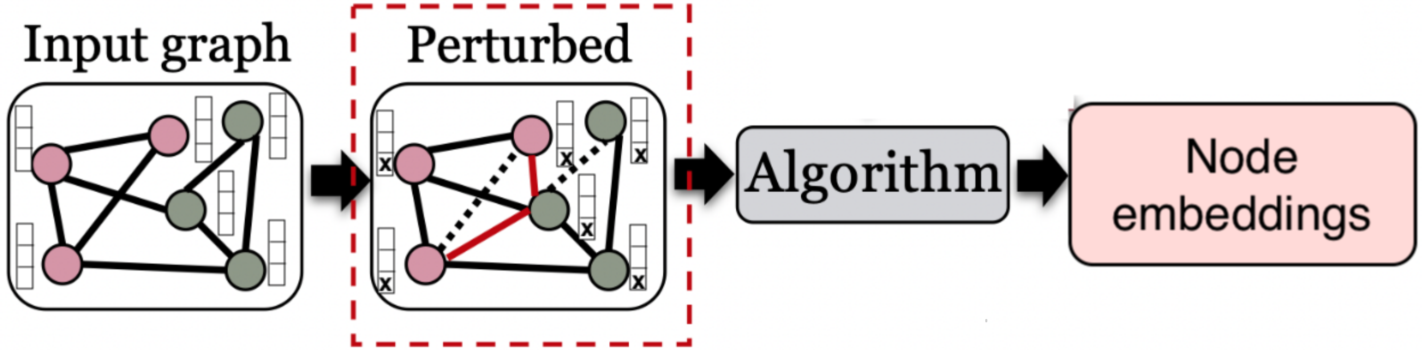
- Group fairness
  - Intuitional edge deletion designs based on observations for sources of structural bias [1], [2], [3]
  - Hand-crafted edge deletion strategies for balanced inter and intra edges

Reduce bias in graph structure



[1] Kose, O. Deniz, et al. "Fair Contrastive Learning on Graphs." In TSIPN, 2022.
[2] Kose, O. Deniz, et al. "Fairness-aware Adaptive Network Link Prediction." In EUSIPCO, 2022.
[3] Spinelli, Indro, et al. "Biased edge dropout for enhancing fairness in graph representation learning." In TAI, 2021.

Feature Masking for Group Fairness

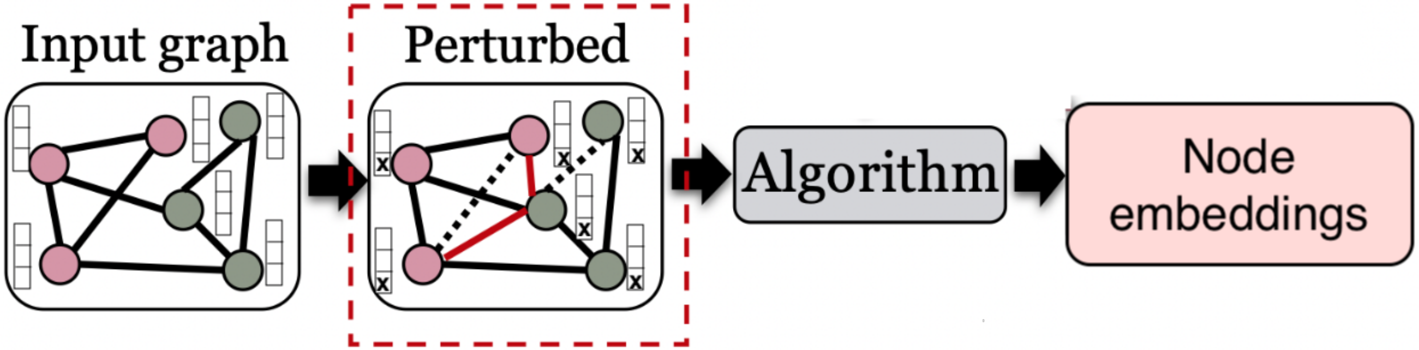


- Group fairness
  - Hand-crafted feature masking strategies based on observations for sources of nodal feature bias [1], [2]
  - Intuition: features correlated with sensitive attributes propagate bias

Mask correlated features with higher probabilities

[1] Kose, O. Deniz, et al. "Fair Contrastive Learning on Graphs." In TSIPN, 2022.
[2] Kose, O. Deniz, et al. "Fairness-aware Adaptive Network Link Prediction." In EUSIPCO, 2022.

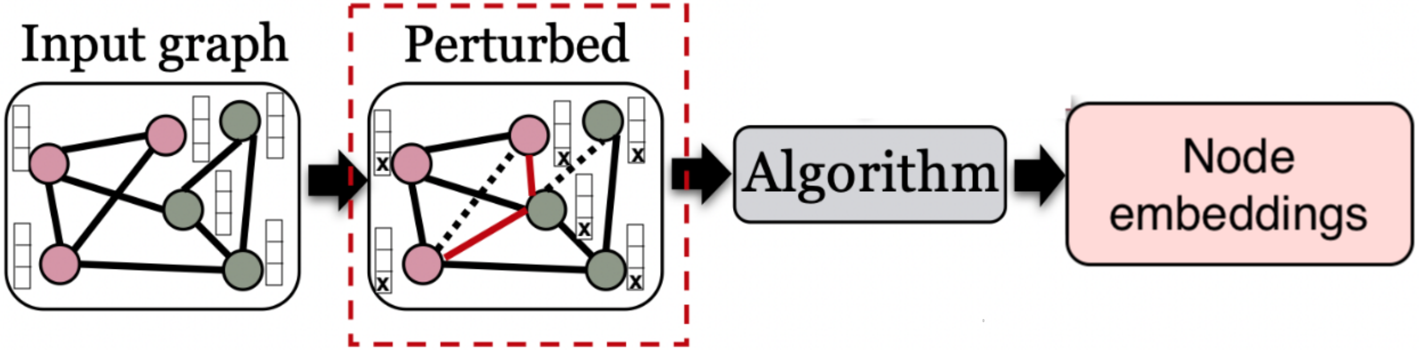
Theory-based Augmentation



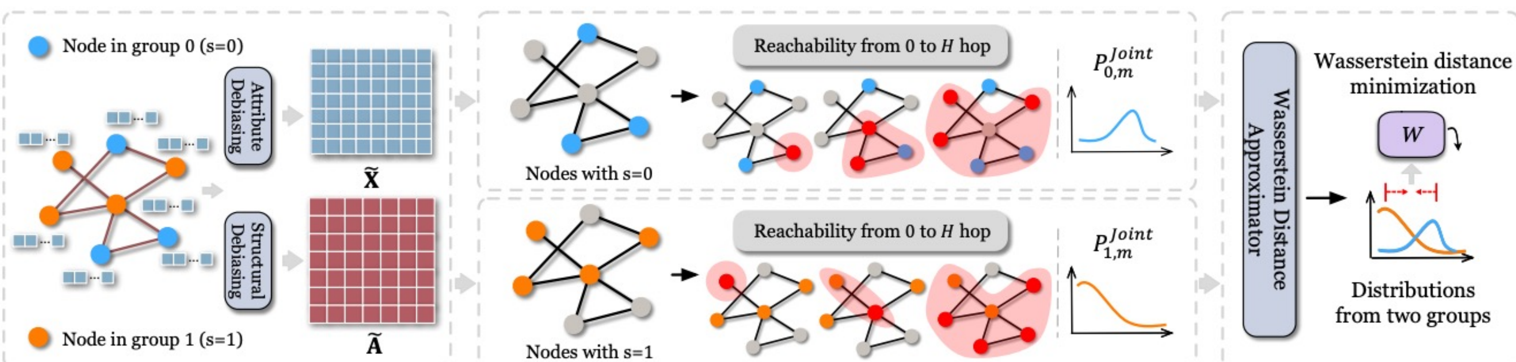
- Group fairness
  - [1] theoretically identifies bias amplifying factors in mean aggregation
  - Manually designs feature masking, node sampling, edge augmentation strategies [1]
    - Each augmentation targets different bias amplifying terms
    - Augmentations minimize the corresponding bias factors

[1] Kose, O. Deniz, et al. "Demystifying and Mitigating Bias for Node Representation Learning." In TNNLS, 2023

Automated Augmentation for Group Fairness

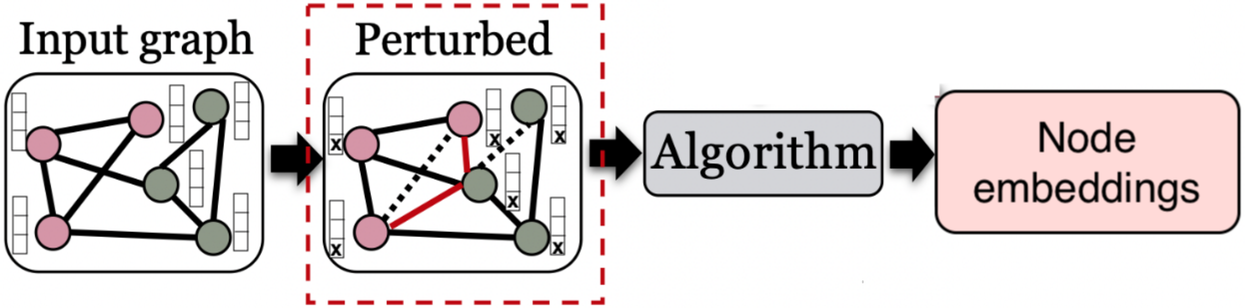


- Group fairness
  - Instead of manual design, optimize augmentations with a fairness loss
  - Automated augmentations on nodal features and graph structure [1]
    - **Fairness loss:** Wasserstein distance between node embeddings' distributions from different sensitive groups



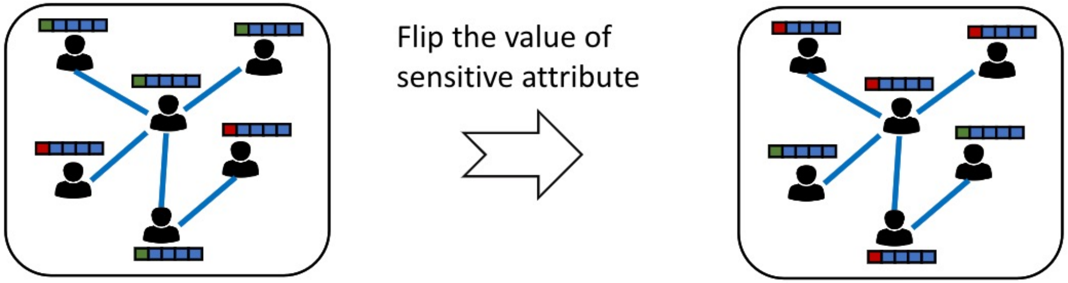
[1] Dong, Yushun, et al. "EDITS: Modeling and Mitigating Data Bias for Graph Neural Networks." In WWW, 2022.

Augmentation for Counterfactual Fairness



- Counterfactual fairness [1]: Node embeddings should be same after flipping sensitive attribute, while everything else is fixed.

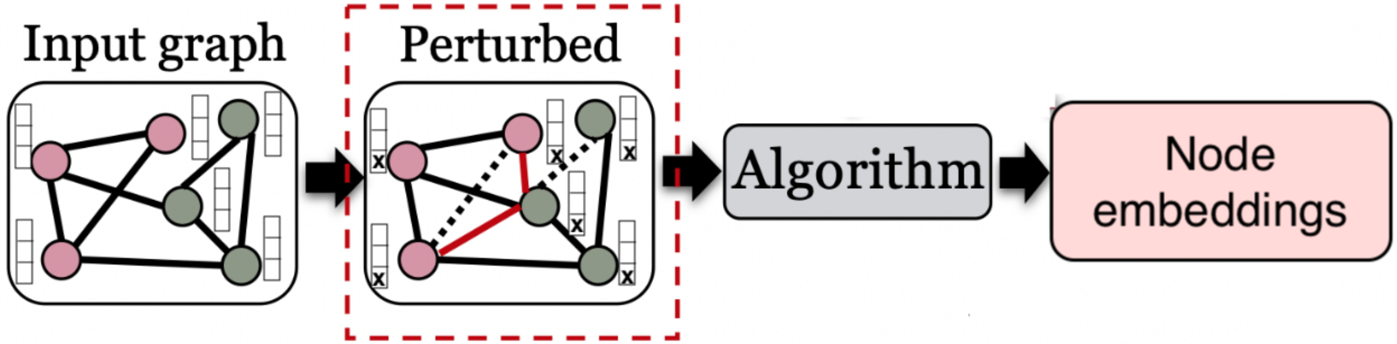
- Design [1]:
  - Flip sensitive attributes in augmented graph
  - Bring embeddings of original and augmented graph closer



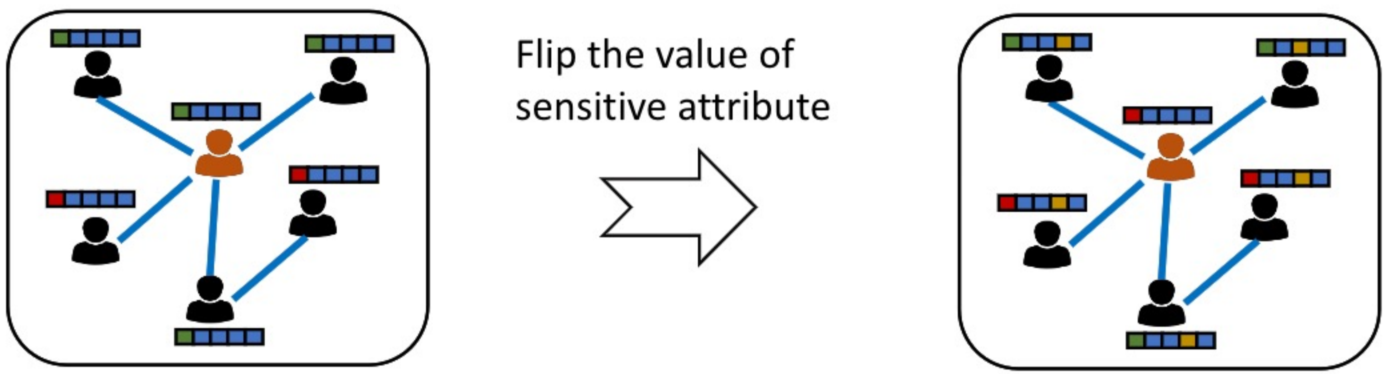
- Intuition: Embeddings must be independent of sensitive attributes

[15] Agarwal, Chirag, et al. "Towards a unified framework for fair and stable graph representation learning." In UAI, 2021.

Augmentation for Counterfactual Fairness

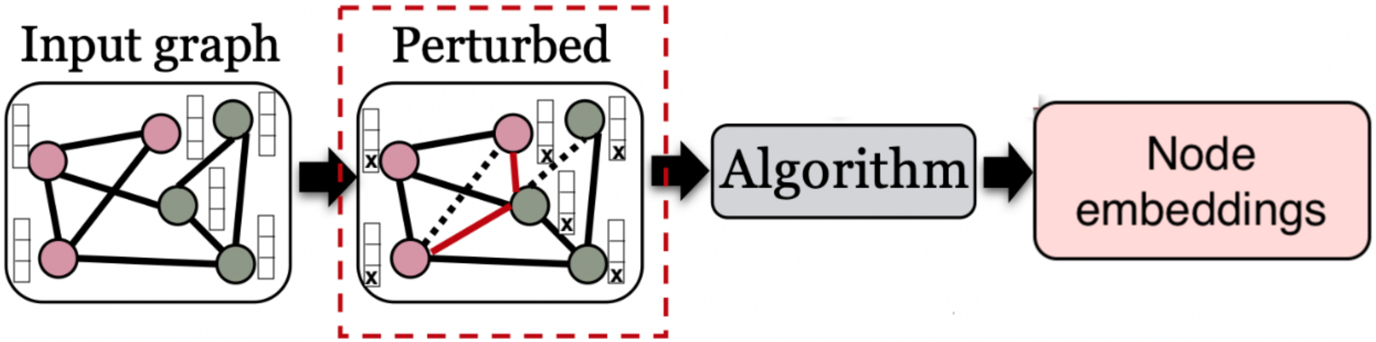


- [1] extends counterfactual fairness definition on graphs
  - The effect of sensitive attributes of neighbors

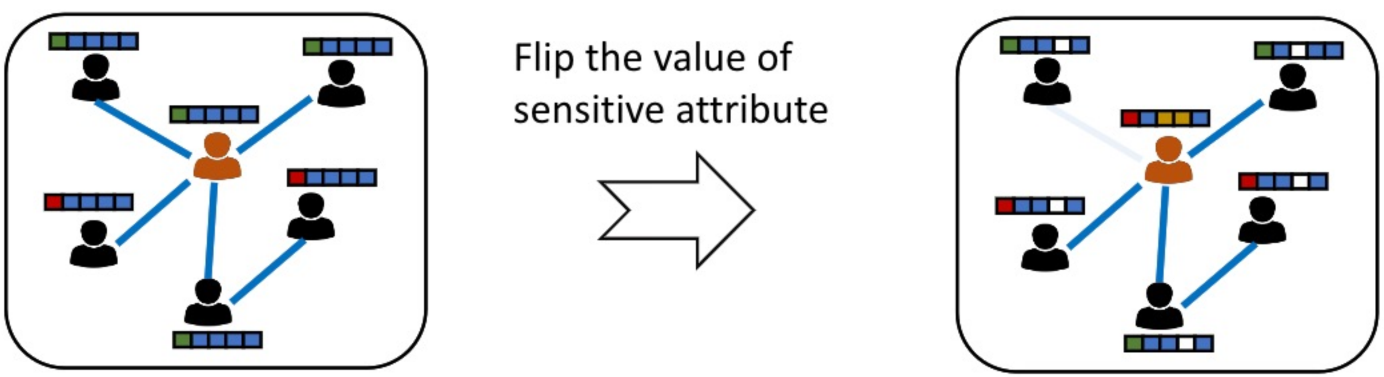


[1] Ma, Jing, et al. “Learning fair node representations with graph counterfactual fairness.” In WSDM,2022.

Augmentation for Counterfactual Fairness

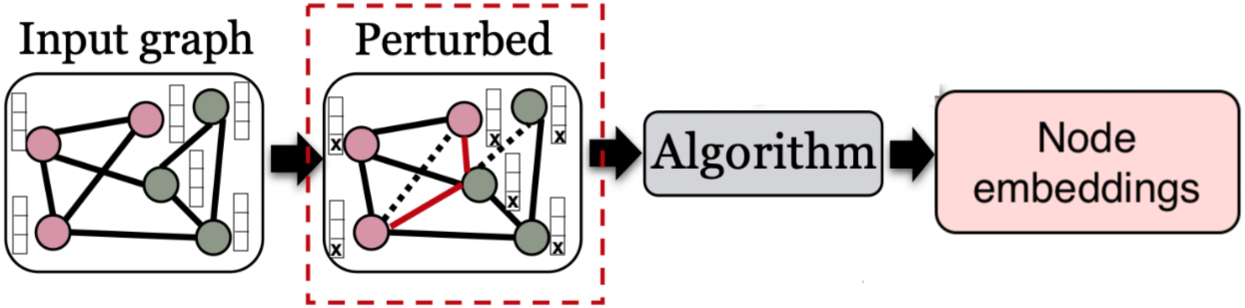


- [1] extends counterfactual fairness definition on graphs
  - The effect of sensitive attributes of neighbors
  - Causal effect from sensitive attributes on other variables like nodal features and graph adjacency



[1] Ma, Jing, et al. "Learning fair node representations with graph counterfactual fairness." In WSDM,2022.

Augmentation for Counterfactual Fairness



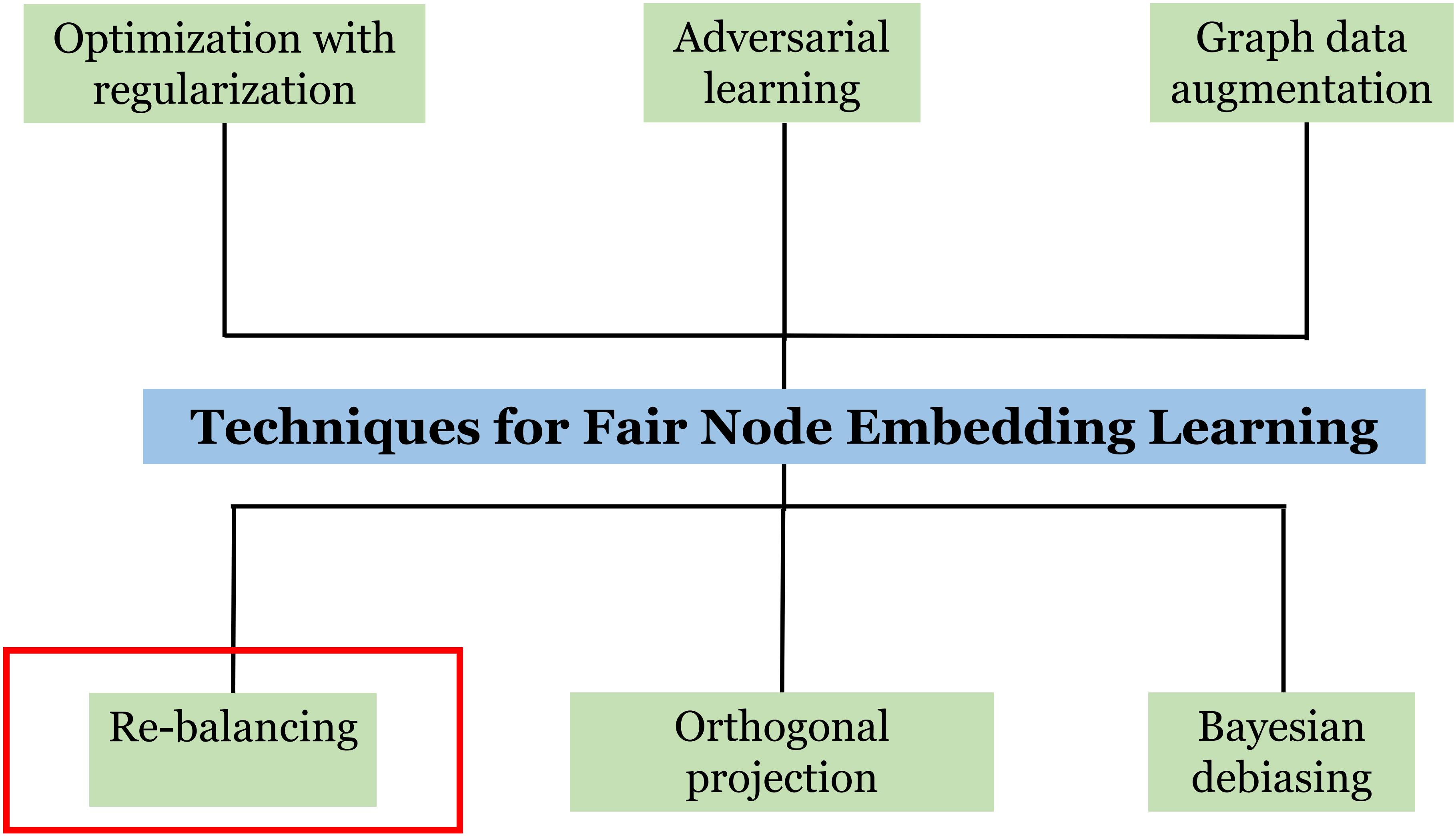
- [1] extends counterfactual fairness definition on graphs
  - The effect of sensitive attributes of neighbors
  - Causal effect from sensitive attributes on other variables like nodal features and graph adjacency

- Automized augmentation to generate counterfactual subgraphs for each node [1]
  - Optimization via an adversarial loss

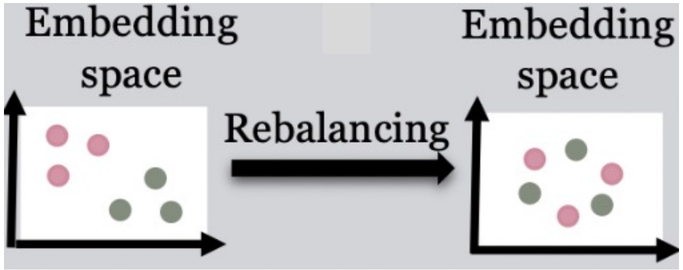
- Bring embeddings based on counterfactual subgraphs closer

[1] Ma, Jing, et al. "Learning fair node representations with graph counterfactual fairness." In WSDM,2022.

Overview

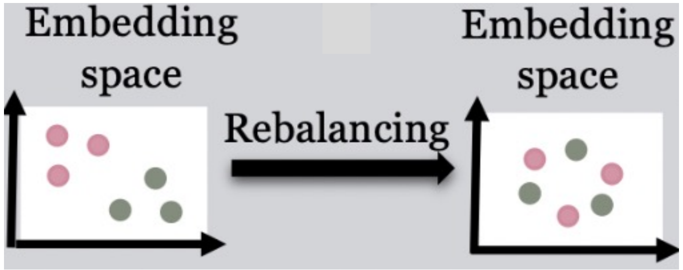

Optimization with regularization
Adversarial learning
Graph data augmentation
Techniques for Fair Node Embedding Learning
Re-balancing
Orthogonal projection
Bayesian debiasing

Rebalancing: Path-based Rebalancing

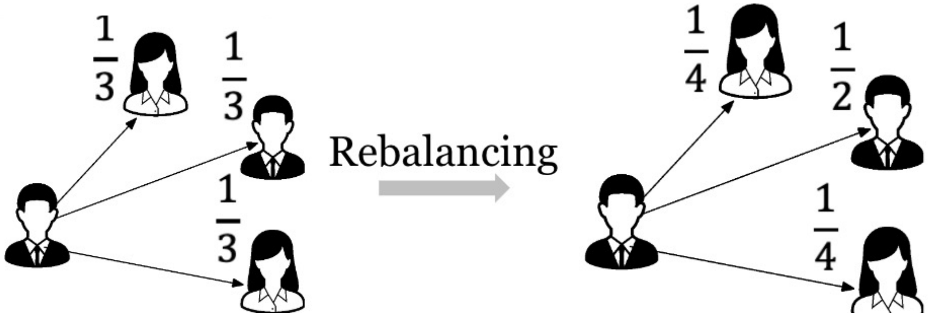


- Aim: similar embedding distributions for different sensitive groups
- Method: Re-distribute weights of edges without topology change
- Group fairness:
  - Path-based rebalancing
  - Edge-based rebalancing
- Degree-based fairness: Re-weight existing edges

Rebalancing: Path-based Rebalancing

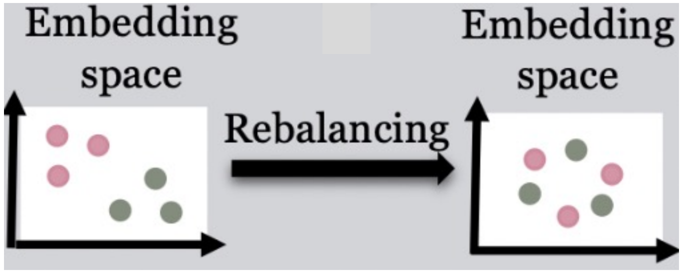


- Aim: similar embedding distributions for different sensitive groups
- Method: Re-distribute weights of edges without topology change
- Group fairness: balanced weights for inter- and intra-edges
  - Path-based rebalancing for random walk-based embeddings [1], [2]



[1] Rahman, Tahleen, et al. "Fairwalk: Towards fair graph embedding." In IJCAI, 2019.
[2] Khajehnejad, Ahmad, et. al. "CrossWalk: Fairness-enhanced Node Representation Learning." In AAAI, 2022.

Rebalancing: Edge-based Rebalancing

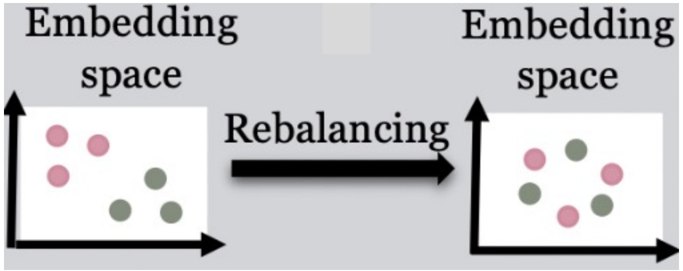


- Aim: similar embedding distributions for different sensitive groups
- Method: Re-distribute weights of edges without topology change
- Group fairness: balanced weights for inter- and intra-edges
  - Edge-based rebalancing: Re-distribute weights of edges by optimizing for dyadic loss [1]

Similar probabilities for inter and intra edges

[1] Li, Peizhao, et al. "On dyadic fairness: Exploring and mitigating bias in graph connections." In ICLR, 2021.

Rebalancing for Degree-based Fairness

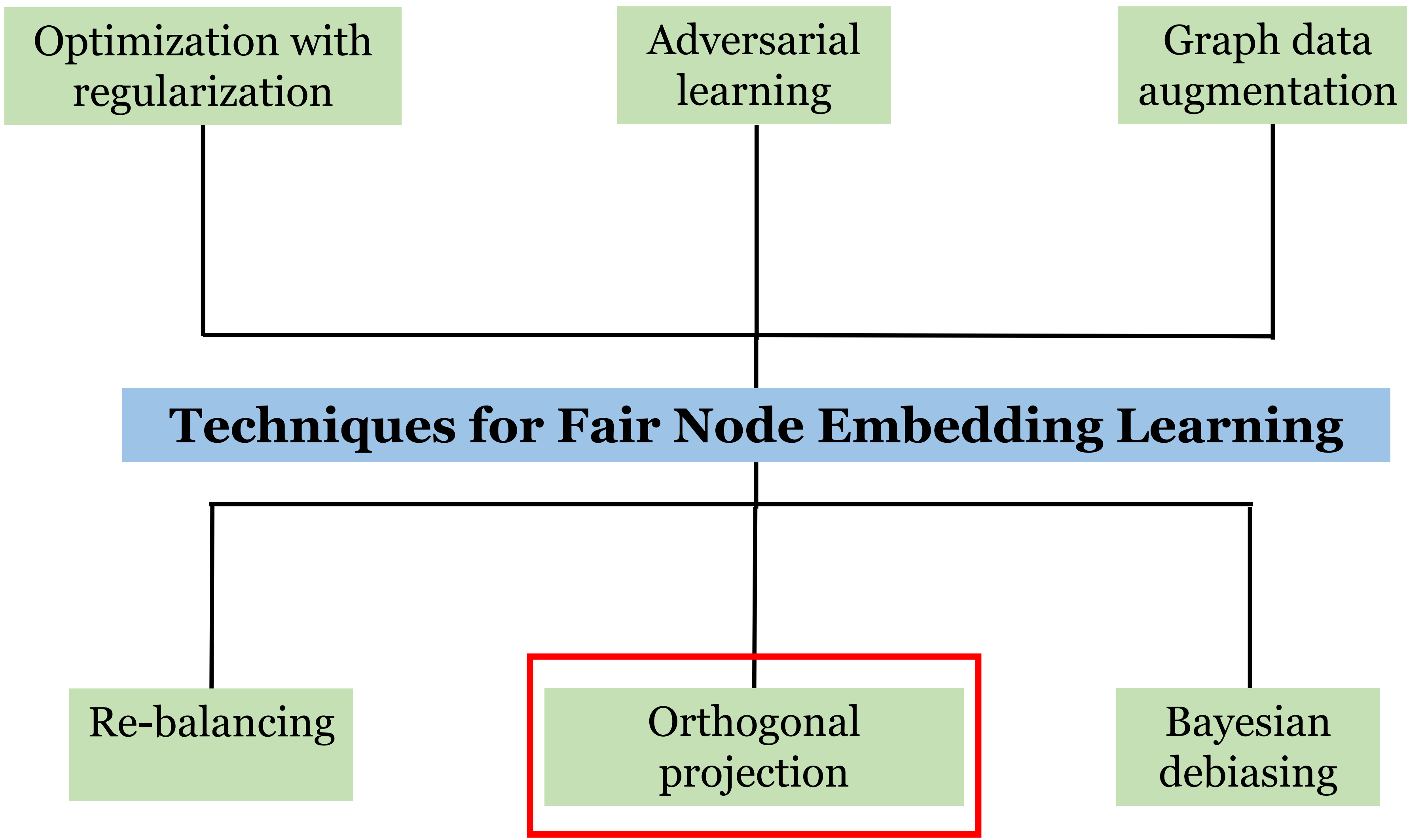


- Aim: similar embedding distributions for different sensitive groups
- Method: Re-distribute weights of edges without topology change
- Degree-based fairness: balanced weights for a constant degree [1]

Rebalances effect of each node in optimization

[1] Kang, Jian, et al. "RawlsGCN: Towards Rawlsian Difference Principle on Graph Convolutional Networks." In WWW, 2022.

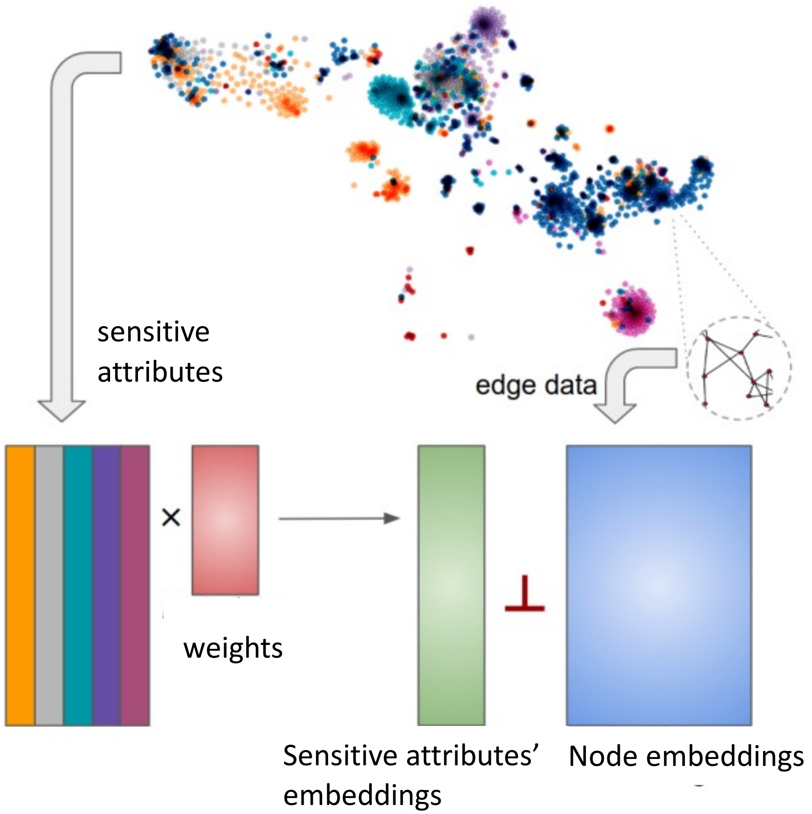
Overview


Optimization with regularization
Adversarial learning
Graph data augmentation
Techniques for Fair Node Embedding Learning
Re-balancing
Orthogonal projection
Bayesian debiasing

Orthogonal Projection

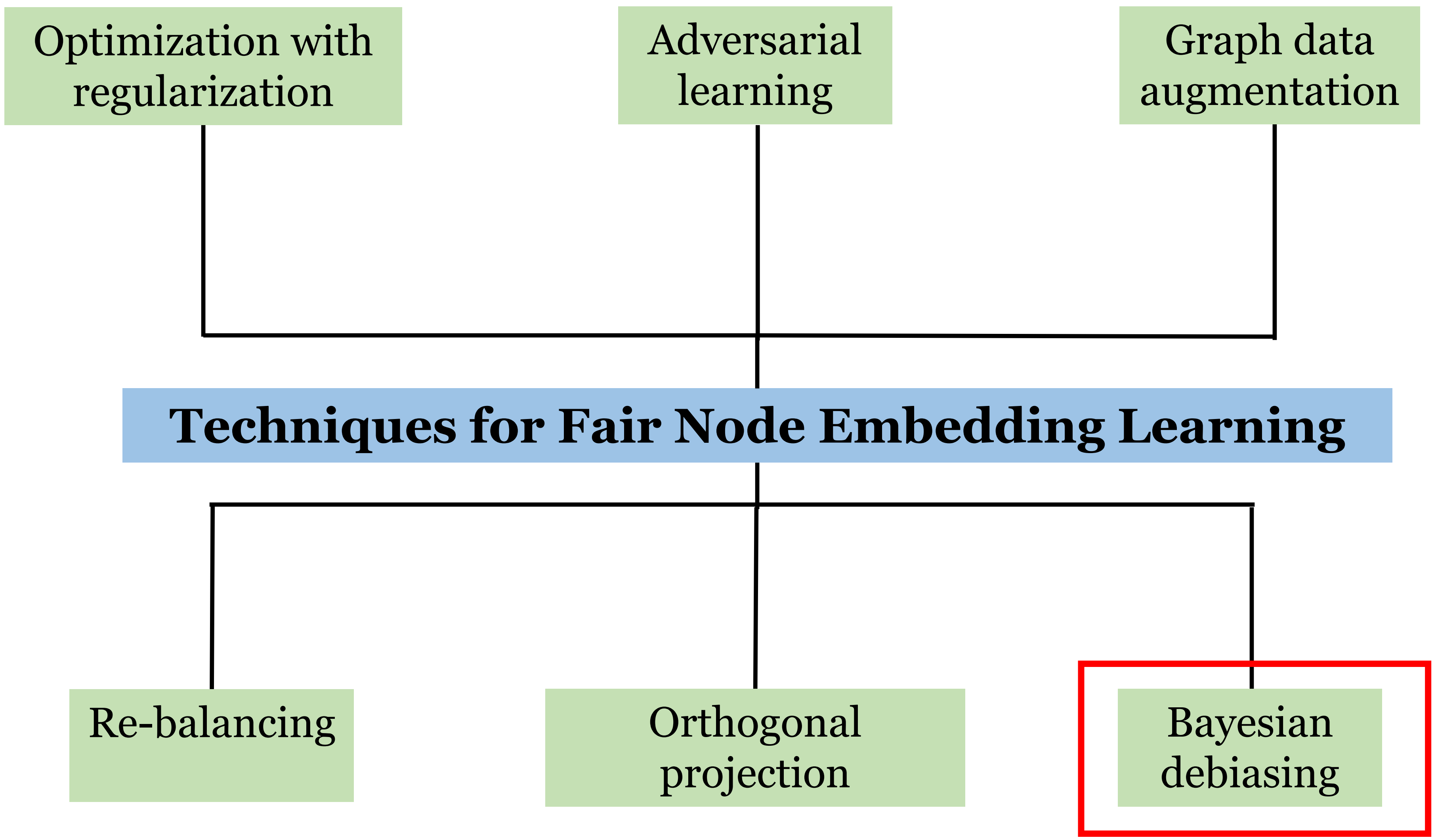
- Node embeddings are on a hyperplane orthogonal to that of sensitive attributes' embeddings
  - enforce linear independence between the two embedding spaces

- Linear debiasing approach

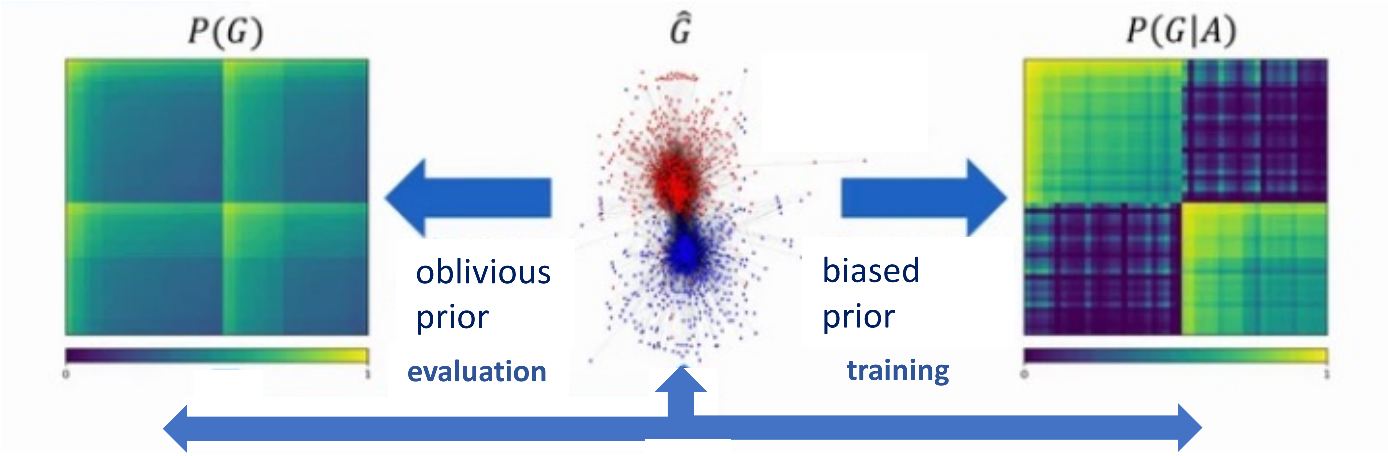


[1] Palowitch, John, et al. "Debiasing Graph Representations via Metadata-Orthogonal Training." In ASONAM 2020.

Overview


Optimization with regularization
Adversarial learning
Graph data augmentation
Techniques for Fair Node Embedding Learning
Re-balancing
Orthogonal projection
Bayesian debiasing

Bayesian Debiasing: DeBayes [1]



- Developed for Bayesian node embedding learning

- **Idea:** model sensitive information in prior distribution of graph as strongly as possible
    - Node embeddings no longer need to represent sensitive information

- Use sensitive information-agnostic prior in evaluation

[1] Buyl, Maarten, et al. "DeBayes: a Bayesian Method for Debiasing Network Embeddings." In ICML, 2020.

Conclusions

- Node embeddings are powerful
  - Carry information of structure and nodal features
  - Facilitate several downstream graph-based tasks

- Essential to prevent bias propagation towards node embeddings

- Six main approaches based on different techniques
  - Optimization with regularization
  - Adversarial learning
  - Graph data augmentation
  - Re-balancing
  - Orthogonal projection
  - Bayesian debiasing

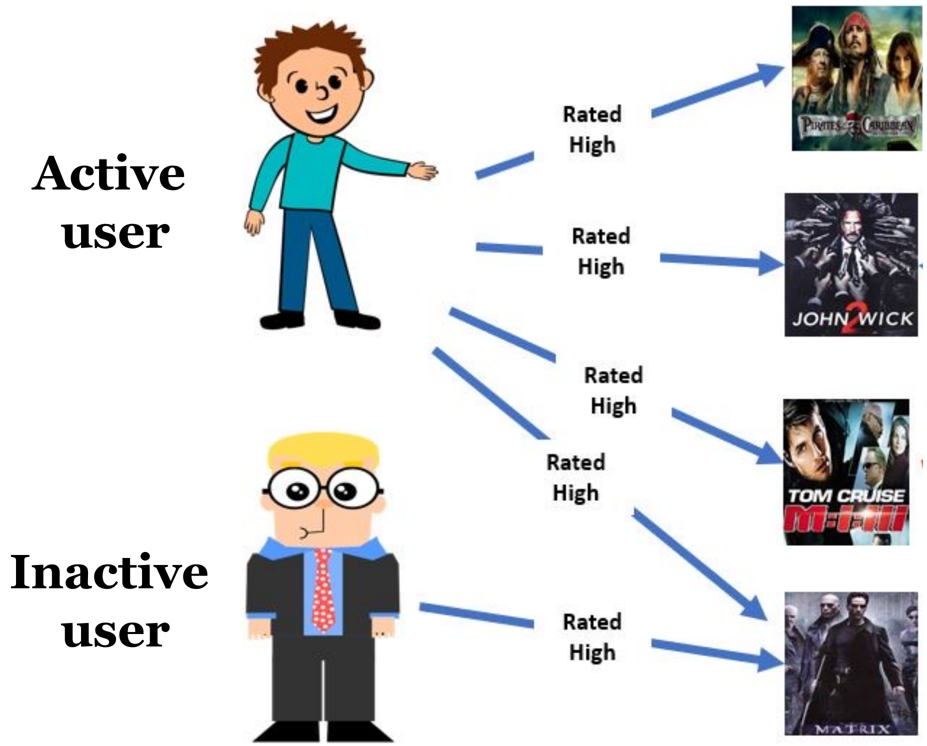
Outline

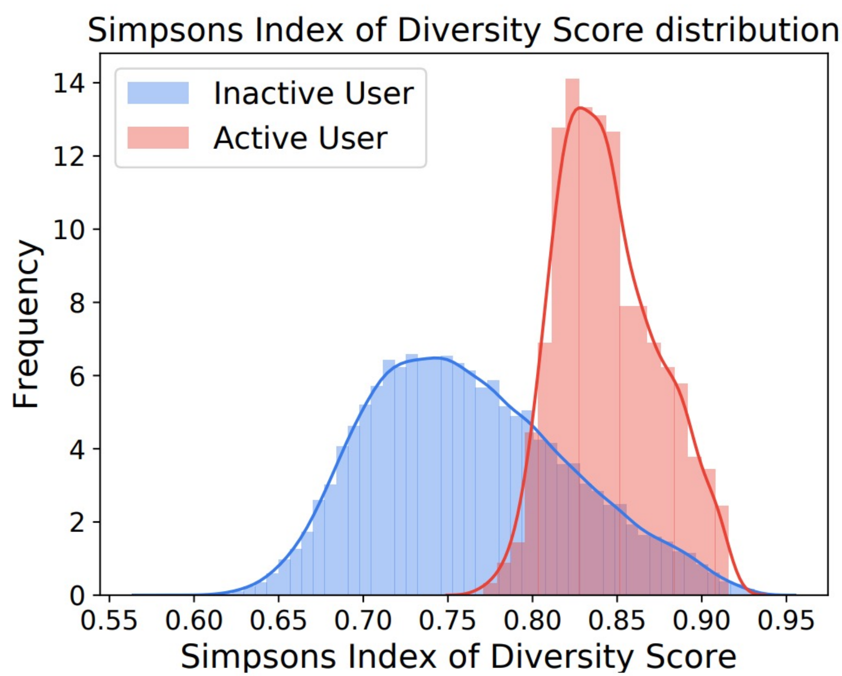
- Background Introduction
- Fairness Notions and Metrics
- Theoretical Understanding of Bias
- Techniques for Fair Graph ML
- Real-World Applications
- Summary, Challenges, & Future Directions

User Fairness in Recommender System

User Fairness: the recommendation quality for different users should be similar.

Example: Active/inactive users



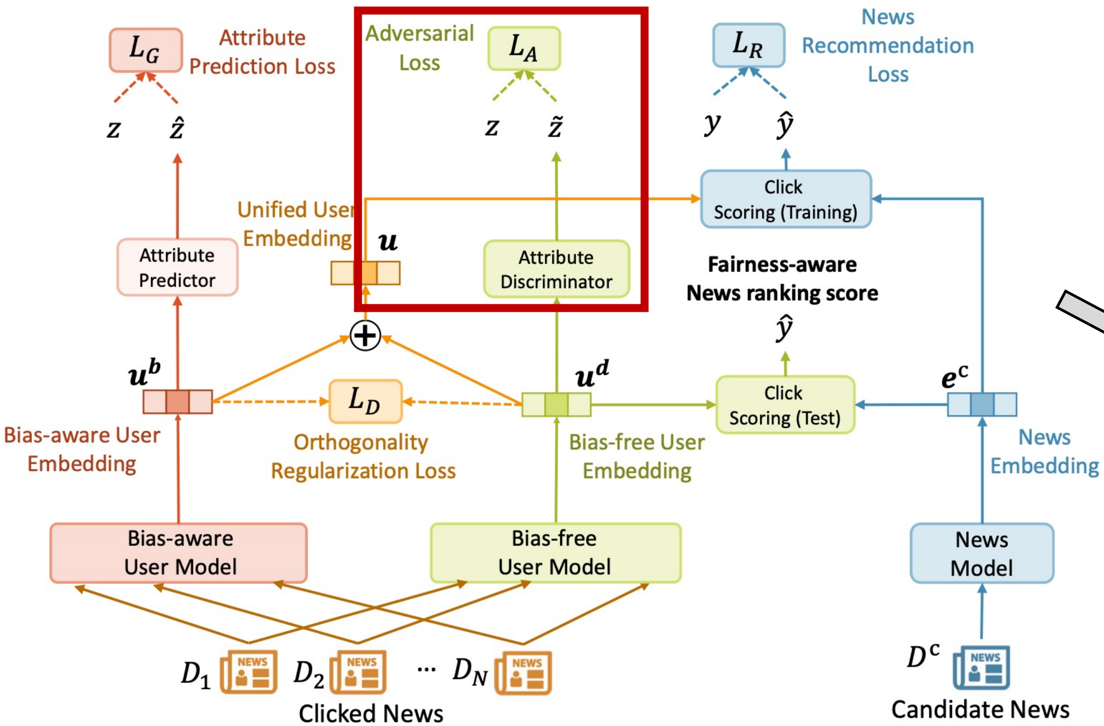


Statistics of Amazon Beauty dataset [1]

[1] Fu, Zuohui, et al. “Fairness-aware explainable recommendation over knowledge graphs.” In SIGIR, 2020.

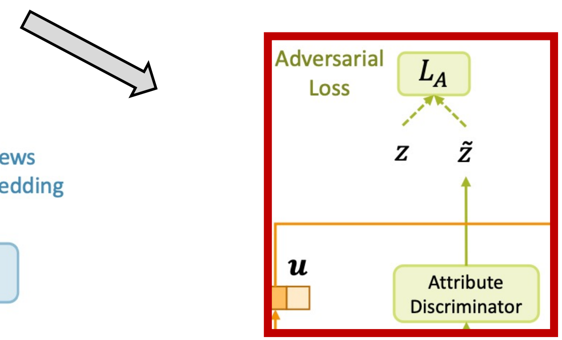
Fulfilling User Fairness

Adversarial learning-based method: avoid delivering news with biased service quality towards certain demographic subgroups.



The architecture of FairRec [1]

The goal of the adversary: excluding the sensitive information from user embeddings.



[1] Wu, Chuhan, et al. "Fairness-aware news recommendation with decomposed adversarial learning." In AAAI, 2021

Popularity Fairness in Recommender System

Popularity Fairness: popular instances should not be over-emphasized compared with other instances.

Example: filter bubble problems.

Example measurement [1]:

$$Q_{\text{fairness}} = \frac{1}{2\boxed{|\mathcal{E}|}} \sum_{i,j} \left(\mathbf{A}_{i,j} - \frac{\boxed{d_i} d_j}{2|\mathcal{E}|}\right) \boxed{\delta} \left(\boxed{M(v_i)}, \boxed{M(v_j)}\right)$$

degree of node i

group membership for user i and j

Total edge number

Kronecker delta function

[1] Masrour, Farzan, et al. "Bursting the filter bubble: Fairness-aware network link prediction." In AAAI, 2020.

Popularity Fairness in Recommender System

Popularity Fairness: popular instances should not be over-emphasized compared with other instances.

Example: filter bubble problems.

Example measurement [1]:

$$Q_{\text{fairness}} = \frac{1}{2|\mathcal{E}|}\sum_{i,j}\left(\mathbf{A}_{i,j} - \frac{d_i d_j}{2|\mathcal{E}|}\right)\delta(M(v_i), M(v_j))$$

degree of node i

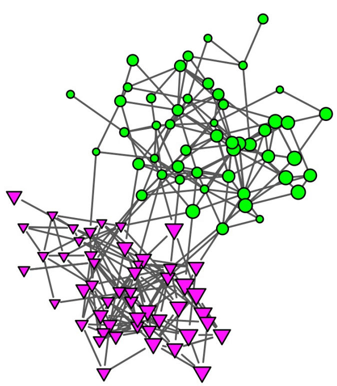
group membership for user i and j

Total edge number

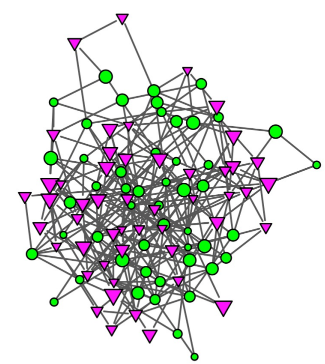
Kronecker delta function

A lower value indicates more inter-group edges, which implies that those less-popular instances are encouraged to connect with other instances.

More biased:



Less biased:



[1] Masrour, Farzan, et al. "Bursting the filter bubble: Fairness-aware network link prediction." In AAAI, 2020.

Popularity Fairness in Recommender System

(1) Regularization-based method: mitigating bias by adding a regularization term, which is relatively easy to use.

An example [1] of regularization for popularity fairness:

$$\mathscr{L}_{fair} = \mathrm{Corr}_{\mathrm{P}}(\boxed{\hat{\mathbf{r}}_+}, \boxed{\mathbf{p}_+})$$

the vector of predicted **relevance scores** for positive user-item pairs

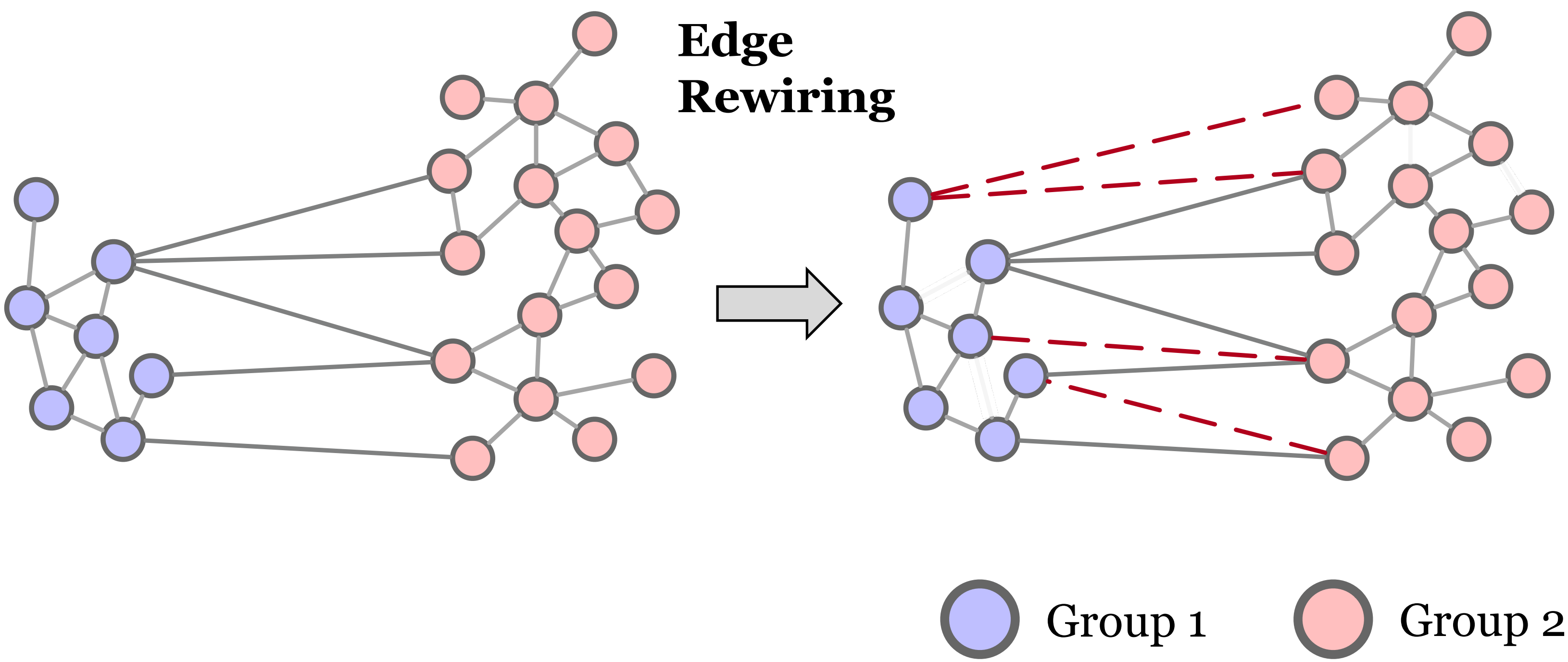
the vector of the feedback number (i.e., **popularity**) received by the items in user-item pairs

This regularization relieves the effect that popular items tend to receive higher relevance scores.

[1] Zhu, Ziwei, et al. "Popularity-opportunity bias in collaborative filtering." In WSDM, 2021.

Popularity Fairness in Recommender System

(2) Edge Rewiring-based method: Based on link prediction results, a proportion of links are rewired (i.e., flipped) in a greedy manner to achieve popularity fairness [1].



[1] Masrour, Farzan, et al. "Bursting the filter bubble: Fairness-aware network link prediction." In AAAI, 2020.

Provider Fairness in Recommender System

Provider Fairness: items from different providers should receive the same exposure rate to the customers.

- **Example of metric 1**: set a minimum exposure guarantee for all providers and used the number of unsatisfied providers to measure provider fairness [1].

[1] Patro, Gourab K., et al. "Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms." In WWW, 2020.

Provider Fairness in Recommender System

Provider Fairness: items from different providers should receive the same exposure rate to the customers.

- **Example of metric 1**: set a minimum exposure guarantee for all providers and used the number of unsatisfied providers to measure provider fairness [1].
- **Example of metric 2**: average number of providers appearing in recommendations [2].

[1] Patro, Gourab K., et al. "Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms." In WWW, 2020.
[2] Liu, Weiwen, et al. "Personalizing fairness-aware re-ranking." arXiv preprint arXiv:1809.02921 (2018).

Provider Fairness in Recommender System

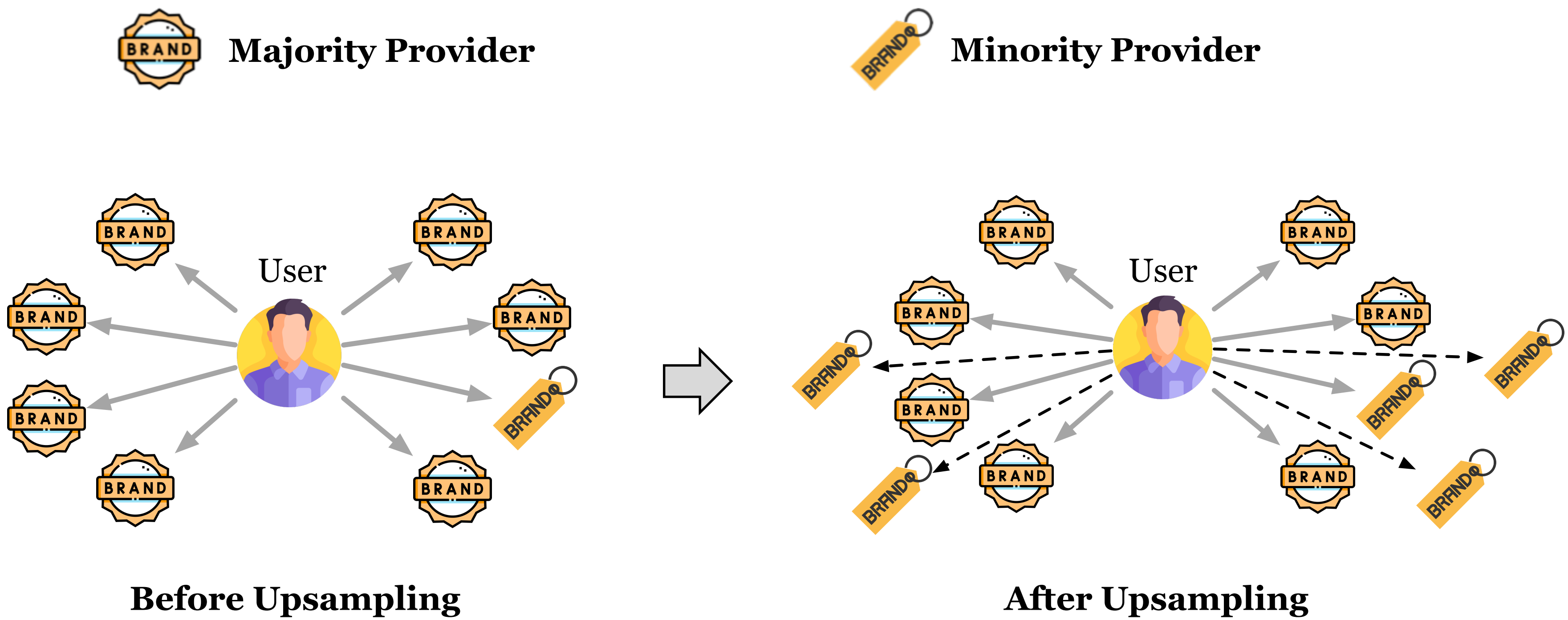
Provider Fairness: items from different providers should receive the same exposure rate to the customers.

- **Example of metric 1**: set a minimum exposure guarantee for all providers and used the number of unsatisfied providers to measure provider fairness [1].
- **Example of metric 2**: average number of providers appearing in recommendations [2].
- **Example of metric 3**: measure both the user-item relevance difference and item exposure rate difference between different providers [3].

[1] Patro, Gourab K., et al. "Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms." In WWW, 2020.
[2] Liu, Weiwen, et al. "Personalizing fairness-aware re-ranking." arXiv preprint arXiv:1809.02921 (2018).
[3] Boratto, Ludovico, et al. "Interplay between upsampling and regularization for provider fairness in recommender systems." In UMUAI, 2020.

Fulfilling Provider Fairness

Rebalancing-based method: upsampling interactions between users and items from minority providers [1].

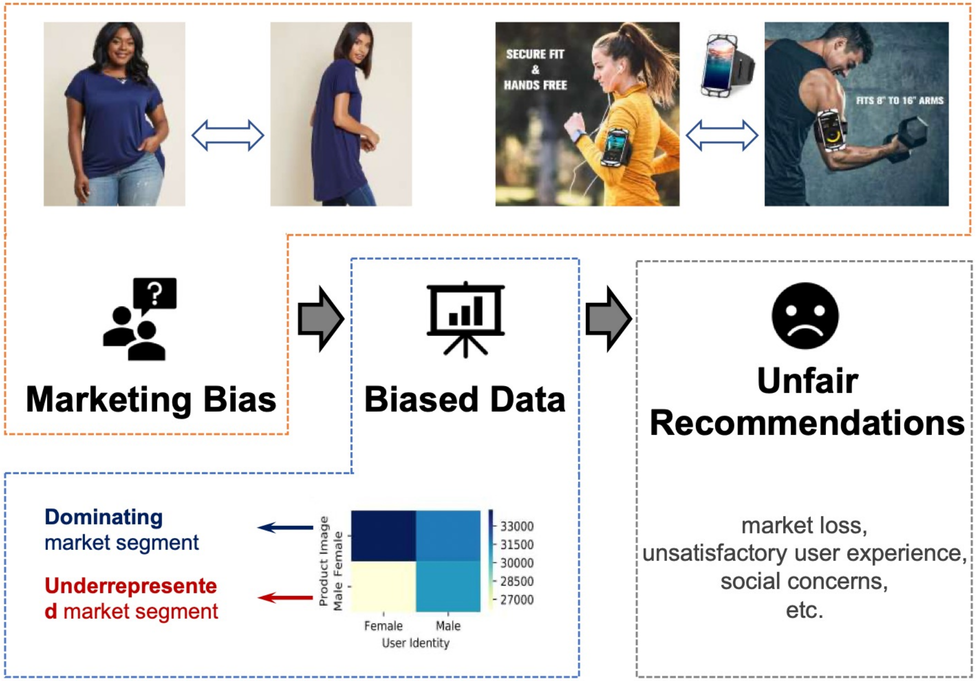


[1] Boratto, Ludovico, et al. "Interplay between upsampling and regularization for provider fairness in recommender systems." In UMUAI, 2021

Marketing Fairness in Recommender System

Marketing Fairness: users are less likely to interact with items whose marketing strategy is not consistent with their identity.

- **Example case**: some gender-neutral items (e.g., armband) could be marketed only with images of males.

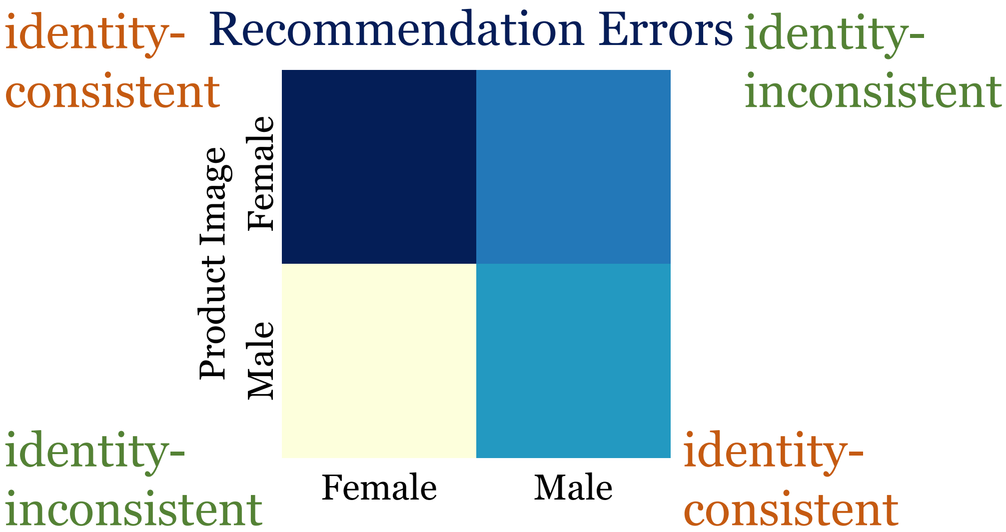


[1] Wan, Mengting, et al. "Addressing marketing bias in product recommendations." In WSDM, 2020.

Marketing Fairness in Recommender System

Marketing Fairness: users are less likely to interact with items whose marketing strategy is not consistent with their identity.

- **Example case**: some gender-neutral items (e.g., armband) could be marketed only with images of females.
- **Measurement**: variance of recommendation errors for identity-consistent and identity-inconsistent users [1].



[1] Wan, Mengting, et al. "Addressing marketing bias in product recommendations." In WSDM, 2020.

Fulfilling Marketing Fairness

Regularization-based method [1]: add an additional term to regularize the correlation between prediction errors and the distribution of market segments.

$$\mathcal{L}^* = \sum (s_{u,i} - r_{u,i})^2 + \alpha \mathcal{L}_{corr.},$$

between-segment var.: $V^{(market)} = \frac{1}{|\mathcal{D}|} \sum_{m,n} |\mathcal{D}_{m,n}| \left(\bar{e}_{m,n,\cdot} - \bar{e}\right)^2$

within-segment var.: $U^{(market)} = \frac{1}{|\mathcal{D}|} \sum_{m,n} \sum_{\substack{u \in \mathcal{U}_m, \\ i \in \mathcal{I}_n}} (e_{u,i} - \bar{e}_{m,n,\cdot})^2$

$$\mathcal{L}_{corr.} = \kappa^{(u.)} \overbrace{\frac{V^{(u.)}}{U^{(u.)}}}^{\text{error parity on user identity}} + \kappa^{(p.)} \underbrace{\frac{V^{(p.)}}{U^{(p.)}}}_{\text{error parity on product image}} + \kappa^{(market)} \overbrace{\frac{V^{(market)}}{U^{(market)}}}^{\text{error parity on market segments}}, \qquad (7)$$

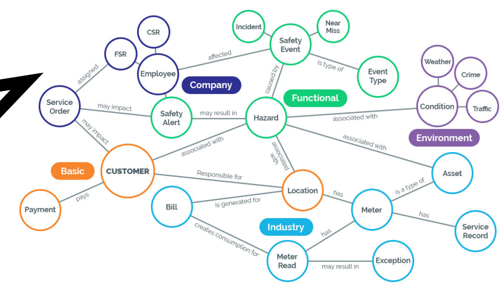
$V^{(u.)}, U^{(u.)}, V^{(p.)}, U^{(p.)}$: merging market segments within the same type of user identity groups or product image groups.

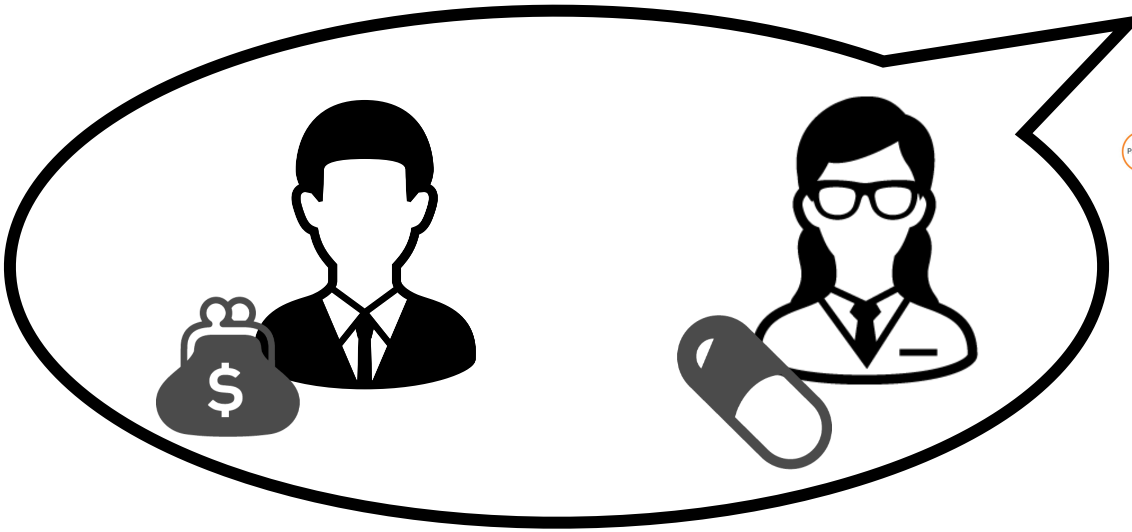
[1] Wan, Mengting, et al. "Addressing marketing bias in product recommendations." In WSDM, 2020.

Social Fairness in Knowledge Graph

Social Fairness: knowledge graph embeddings could encode historical social biases.

- **Example case**: bankers are males and nurses are female.
- **Example of measurement**: Distribution difference between the prediction distribution and uniform distribution over all possible sensitive feature values [1].





A traditional stereotype: bankers are males, while nurses are females [2].

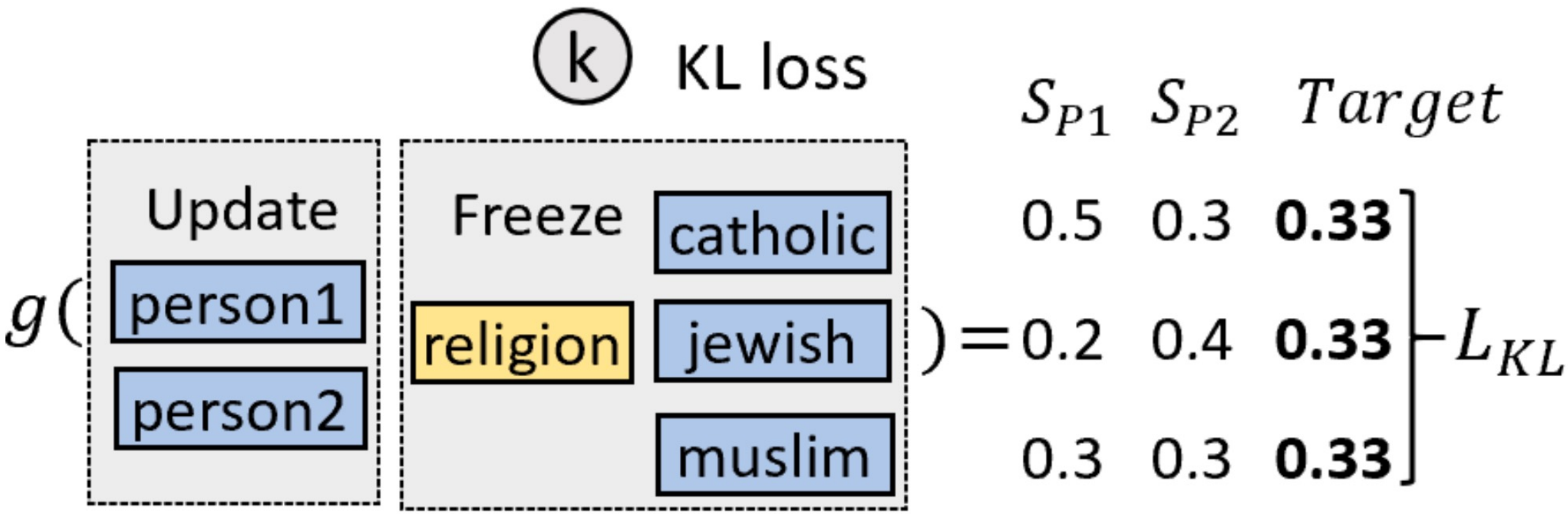
[1] Fisher, Joseph, et al. "Debiasing knowledge graph embeddings." In EMNLP, 2020.
[2] Zeng, Ziqian, et al. “Fair representation learning for heterogeneous information networks.” In AAAI, 2021.

Social Fairness in Knowledge Graph

(1) Regularization-based method

Example: Use KL-divergence between the prediction distribution and uniform distribution over all possible sensitive feature values [1].

Regularization term formulation:



[1] Fisher, Joseph, et al. "Debiasing knowledge graph embeddings." In EMNLP, 2020.

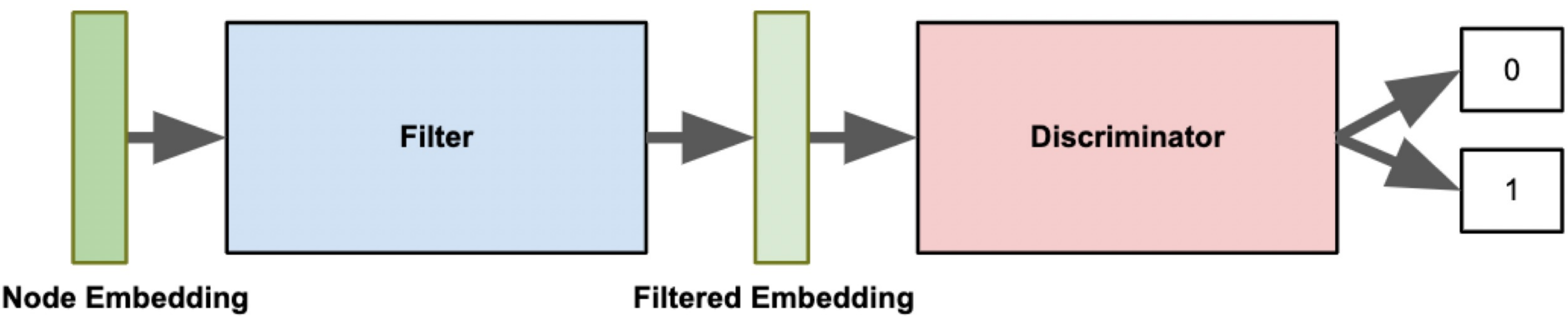
Social Fairness in Knowledge Graph

(1) Regularization-based method

Example: Use KL-divergence between the prediction distribution and uniform distribution over all possible sensitive feature values [1].

(2) Adversarial Learning-based method

Example: Use a sensitive information filter to remove social bias from the embeddings of human entities with a min-max game [2].

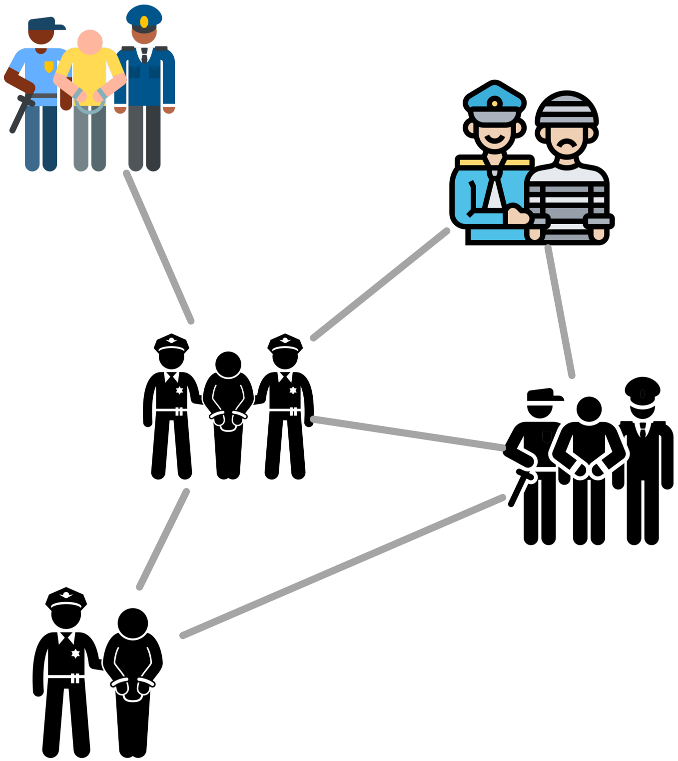


[1] Fisher, Joseph, et al. "Debiasing knowledge graph embeddings." In EMNLP, 2020.
[2] Arduini, Mario, et al. "Adversarial learning for debiasing knowledge graph embeddings." In SIGKDD, 2020.

Fairness in Criminal Justice

Criminal justice: predict whether a defendant deserves bail over a similarity network between defendants [1].

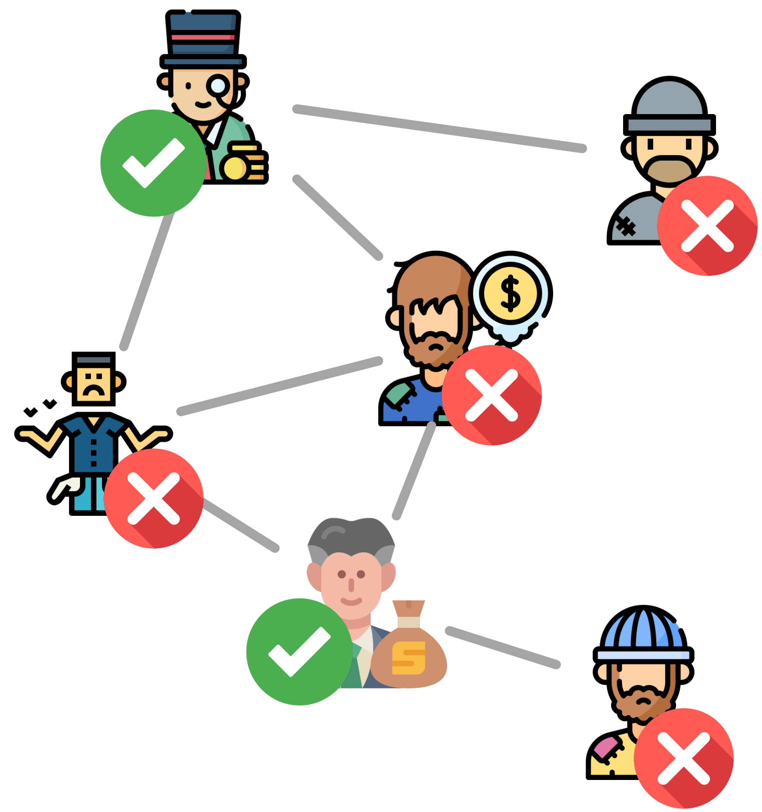
"The United States inarguably has a mass-incarceration crisis, but it is poor people and minorities who bear its brunt. Punishment profiling will exacerbate these disparities—including racial disparities. It also confirms the widespread impression that the criminal justice system is rigged against the poor [2]."



[1] Agarwal, Chirag, et al. "Towards a unified framework for fair and stable graph representation learning." In UAI 2021.
[2] Bazelon, Emily. "Sentencing by the numbers." Open Society Institute 2 (2005).

Fairness in Economics

Economics: default and credit risk prediction over the network between bank clients [1].



[1] Agarwal, Chirag, et al. "Towards a unified framework for fair and stable graph representation learning." In UAI 2021.

Fairness in Social Networks

Social Networks:

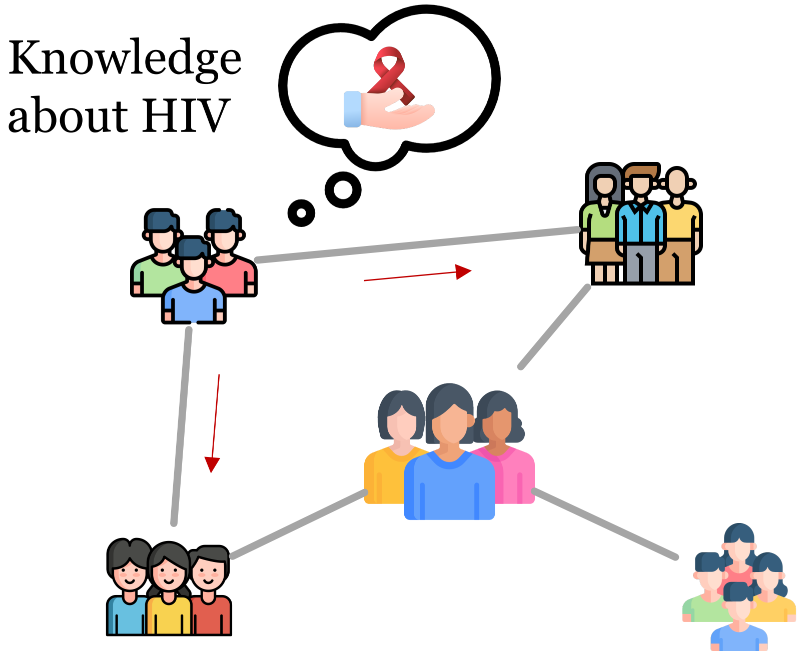
- **Information diffusion** over social networks [1].
- The **gender gap** on social media [2].
- **Fair influence maximization** on social networks [3].

[1] Balaji, T. K., et al. Machine learning algorithms for social media analysis: A survey. Computer Science Review, 2021.
[2] Dai, E., et al. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In WSDM, 2021.
[3] Khajehnejad, M., et al. Adversarial graph embeddings for fair influence maximization over social networks. In IJCAI, 2020.

Fairness in Health Care

Health Care: prevent people from HIV over real-world social connections.

Example: in the HIV prevention domain, we wish to ensure that members of racial minorities or of LGBTQ identity are not disproportionately excluded from knowledge & resources [1].



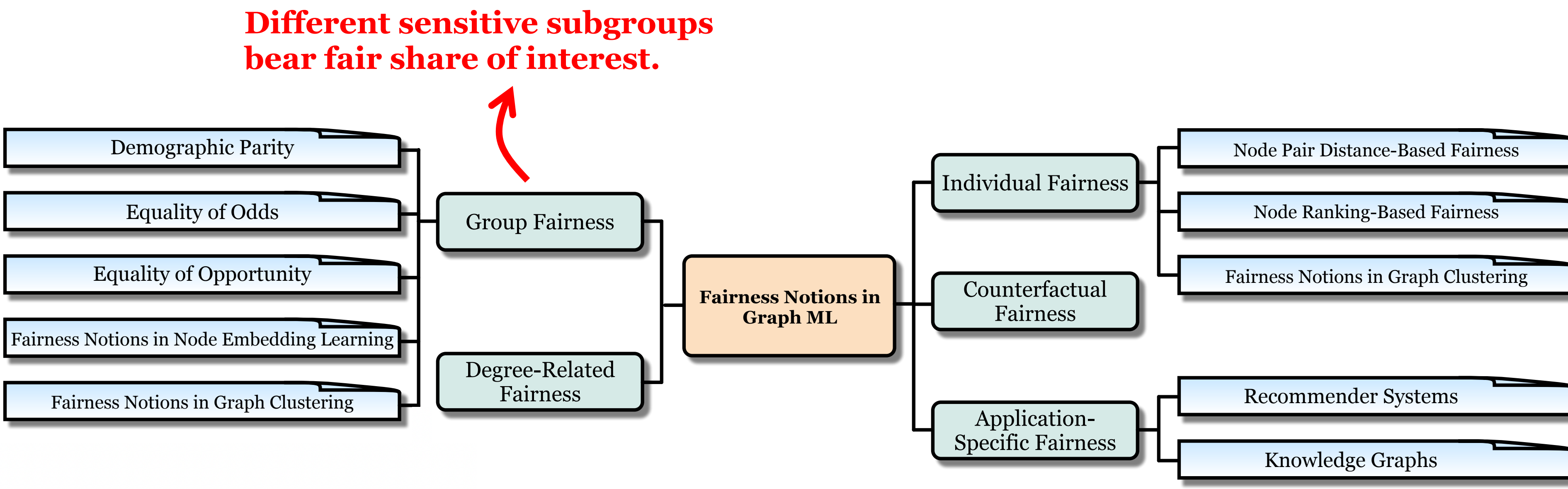
[1] Tsang, Alan, et al. "Group-fairness in influence maximization." In IJCAI, 2019.

Outline

- Background Introduction
- Fairness Notions and Metrics
- Theoretical Understanding of Bias
- Techniques for Fair Graph ML
- Real-World Applications
- Summary, Challenges, & Future Directions

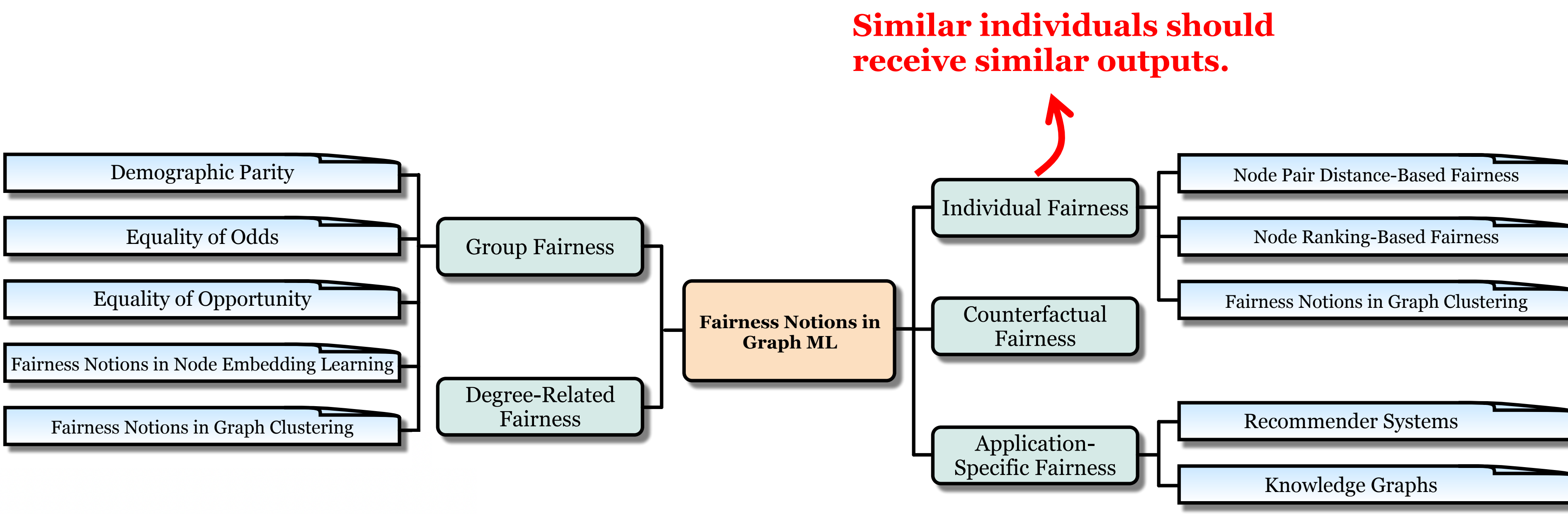
Summary on Fairness Notions

The taxonomy of fairness notions:



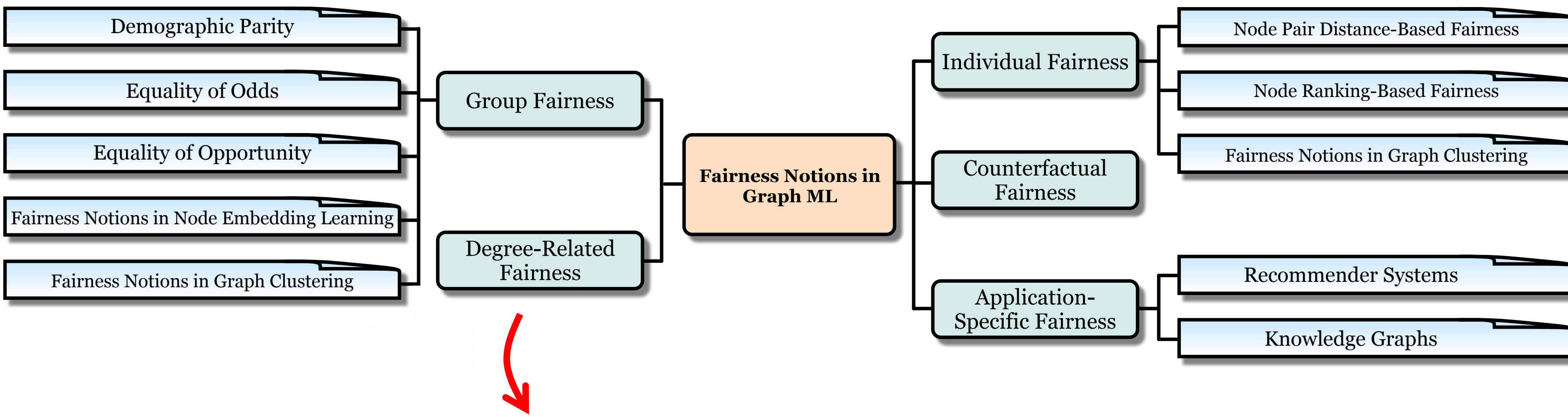
Summary on Fairness Notions

The taxonomy of fairness notions:



Summary on Fairness Notions

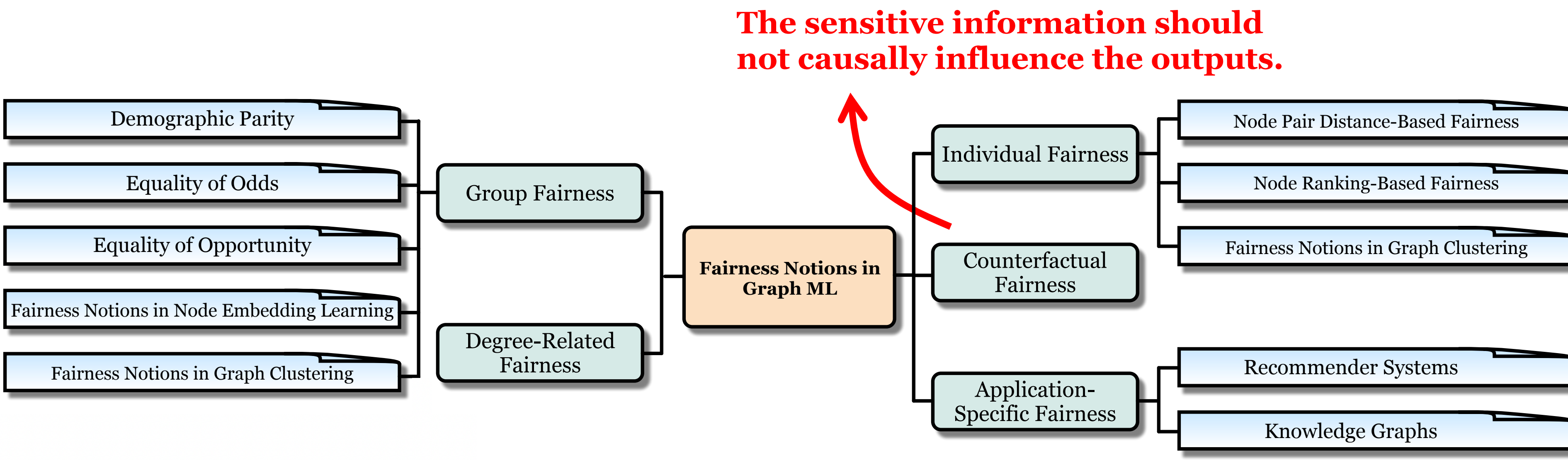
The taxonomy of fairness notions:



Nodes with different degrees should bear similar level of utility from the graph mining model.

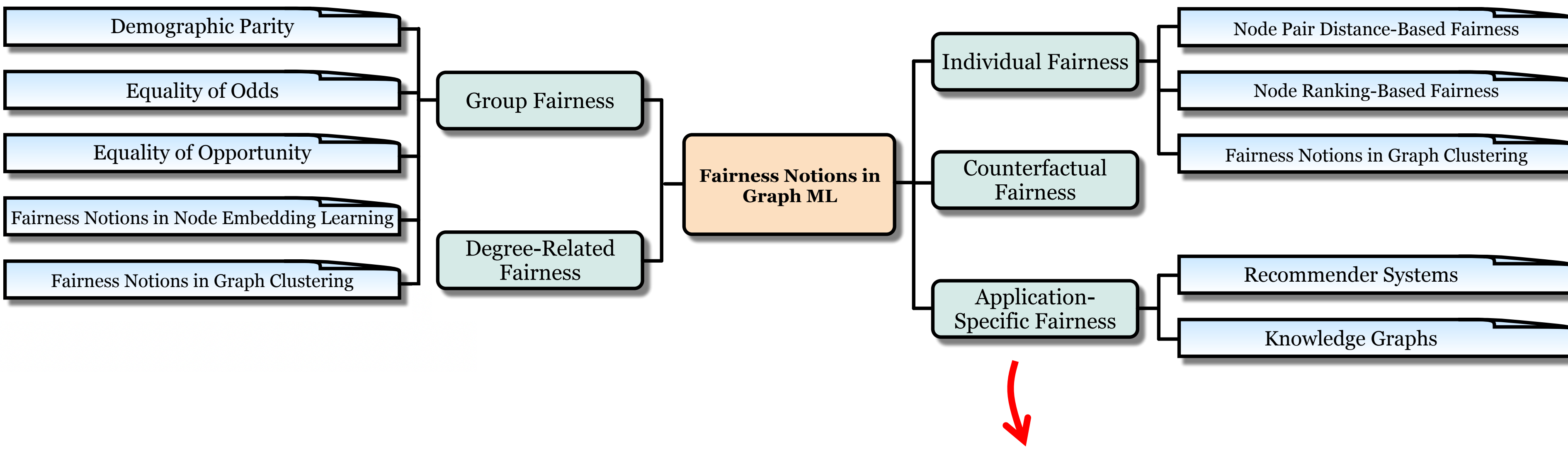
Summary on Fairness Notions

The taxonomy of fairness notions:



Summary on Fairness Notions

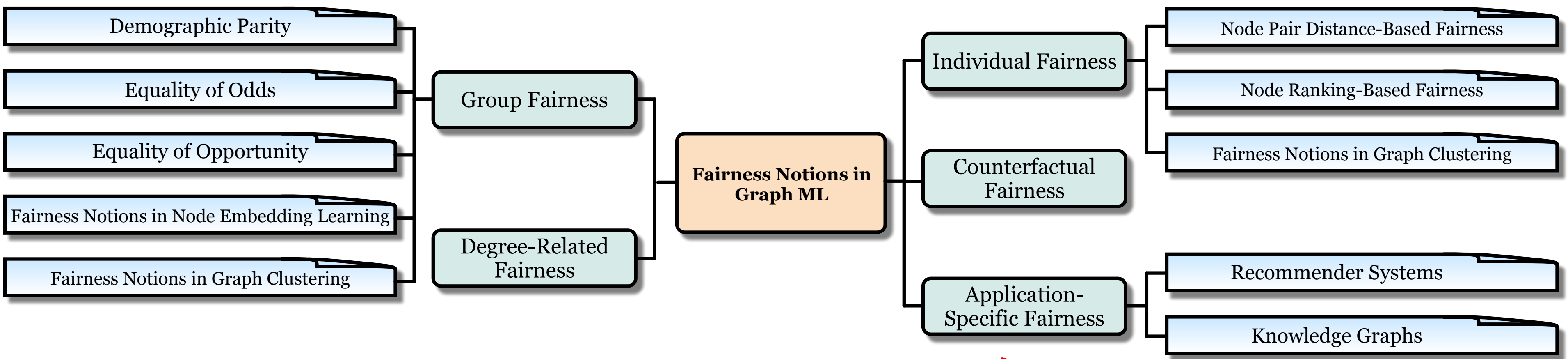
The taxonomy of fairness notions:



Recommender Systems:
(1) User Fairness
(2) Popularity Fairness
(3) Provider Fairness
(4) Marketing Fairness

Summary on Fairness Notions

The taxonomy of fairness notions:



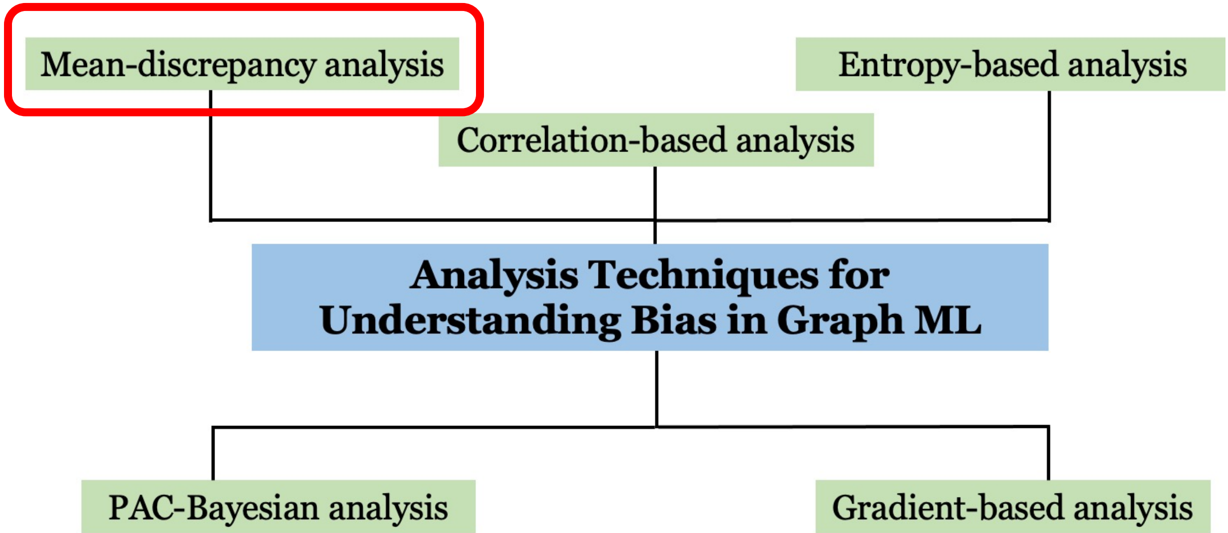
Knowledge Graphs:

(1) Social Fairness

(2) Path Diversity Fairness

(3) Popularity Fairness

Mean-discrepancy Analysis

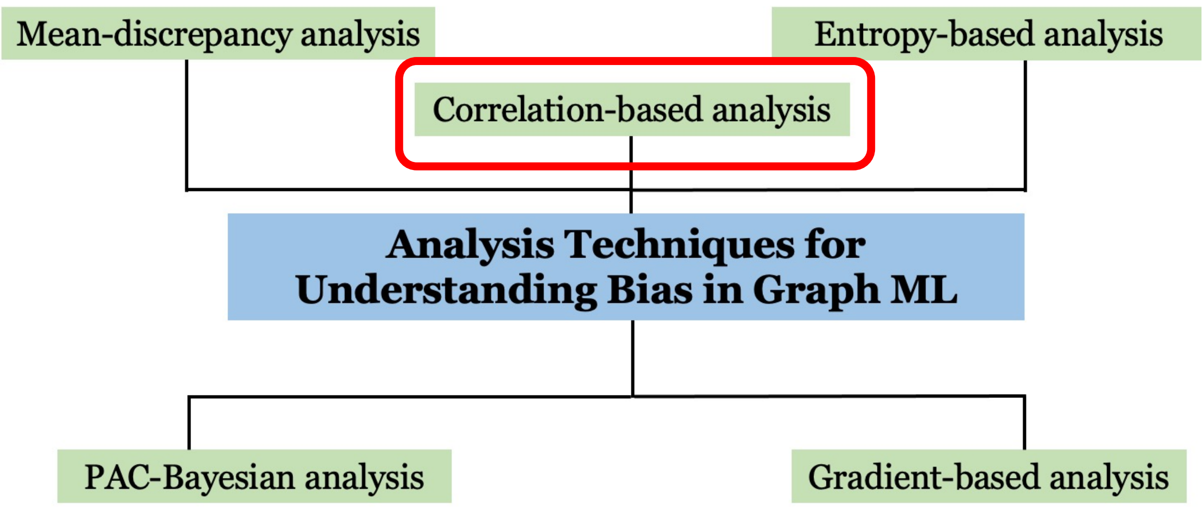


- Disparity between aggregated representations from different sensitive groups

$$\left\|\mathbb{E}_{v\sim\mathcal{V}}\left[\mathrm{Agg}(v) \mid v \in \mathcal{S}_0\right] - \mathbb{E}_{v\sim\mathcal{V}}\left[\mathrm{Agg}(v) \mid v \in \mathcal{S}_1\right]\right\|_2$$

- A measure for demographic parity for both link prediction and node classification

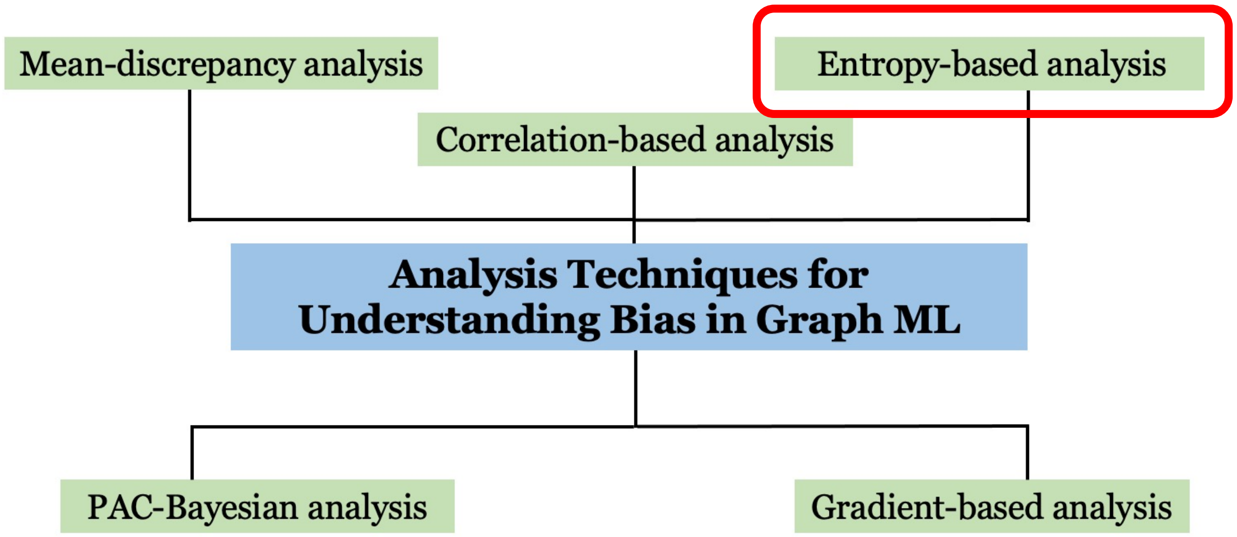
Correlation-based Analysis



- Features correlated with sensitive attributes lead to intrinsic bias
- Correlation between aggregated features and sensitive attributes

$$||\boldsymbol{\rho}||_1 \text{ with } \rho_i = \mathrm{Corr}(\underbrace{\mathbf{z}_{:,i}}_{\text{i-th aggregated feature}}, \mathbf{s})$$

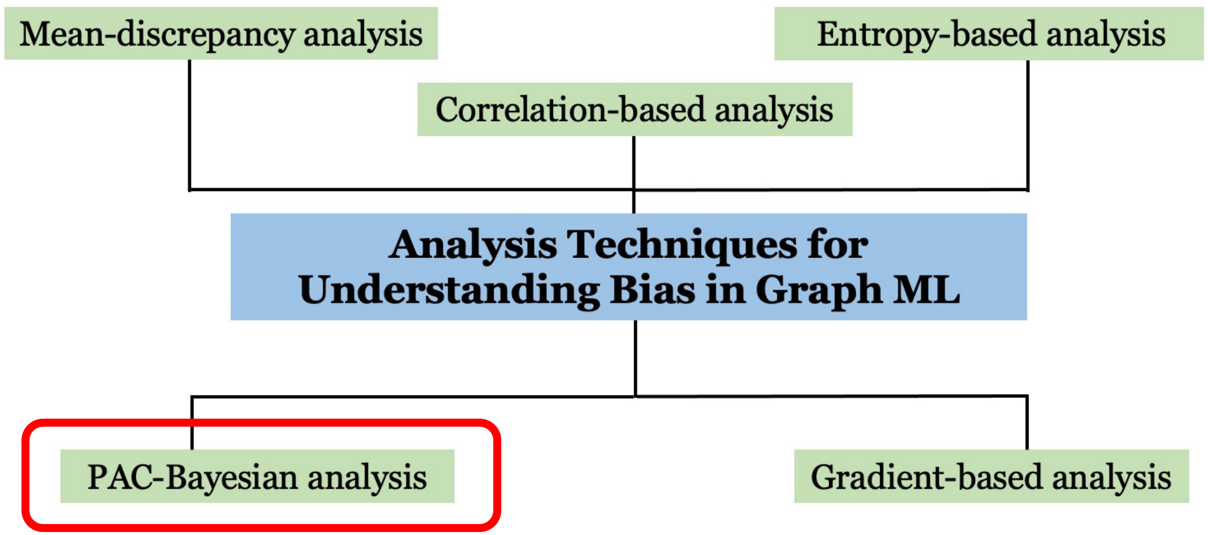
Entropy-based Analysis



- Mutual information between aggregated representations and sensitive attributes

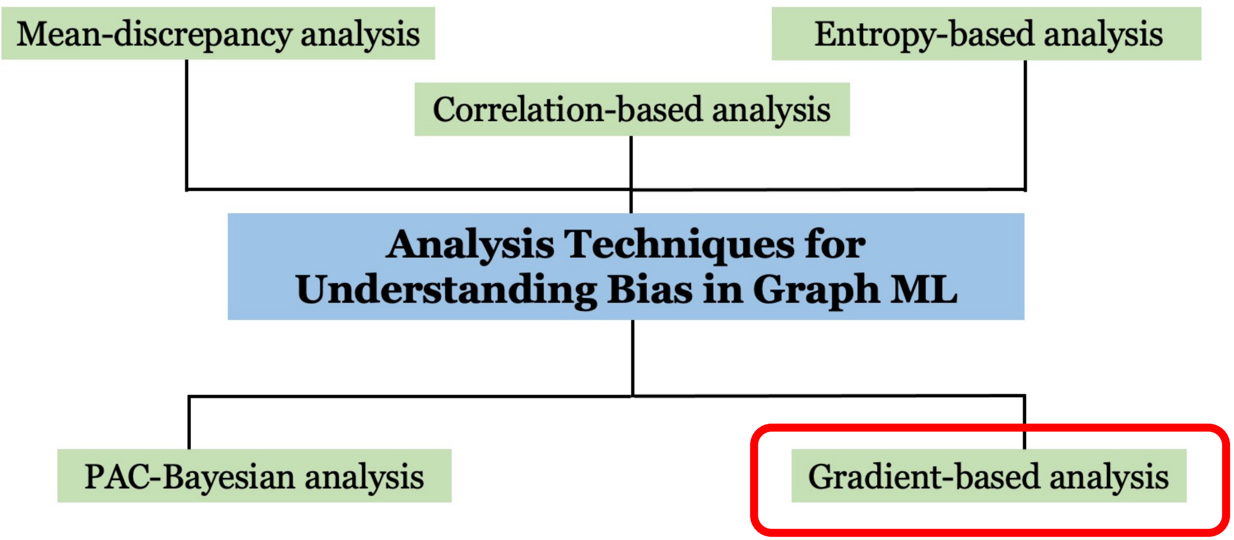
- For a tractable metric, upper bound mutual information

PAC-Bayesian Analysis



- Generalization ability of trained GNN on different sensitive groups

PAC-Bayesian Analysis

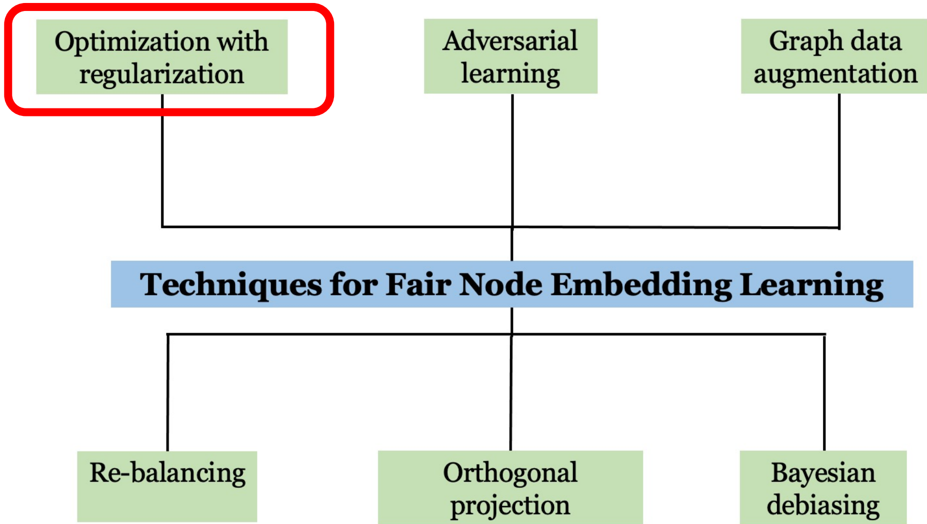


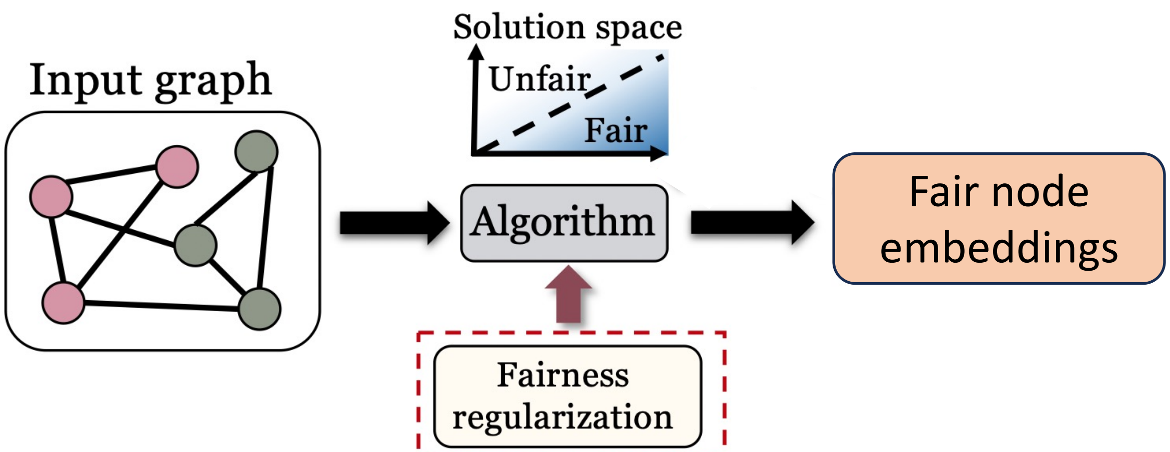
- Gradients of loss wrt weight matrices in GNN layers
  - key component in training

$$\frac{\partial J}{\partial \mathbf{W}^{(l)}} = \sum_{j=1}^{n} \deg(j)\mathbf{I}_j^{(row)} = \sum_{i=1}^{n} \deg(i)\mathbf{I}_i^{(col)}$$

- Explainability for degree-related bias

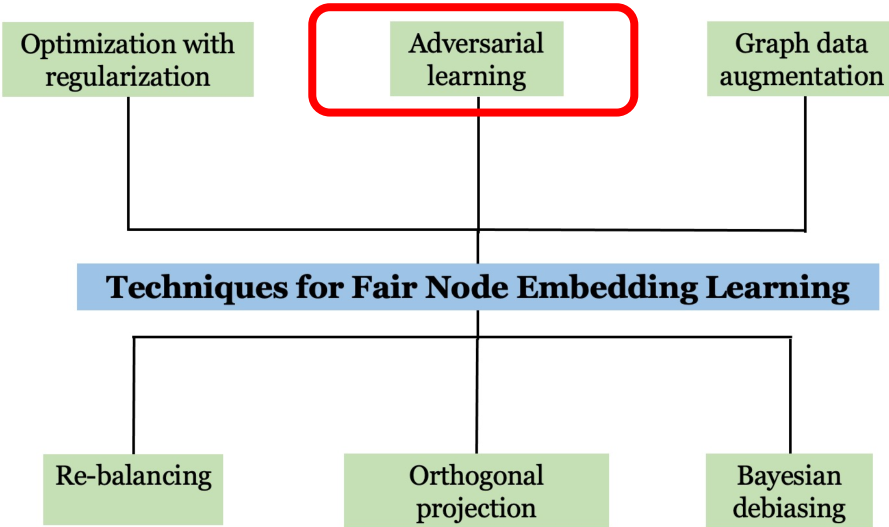
Optimization with Regularization

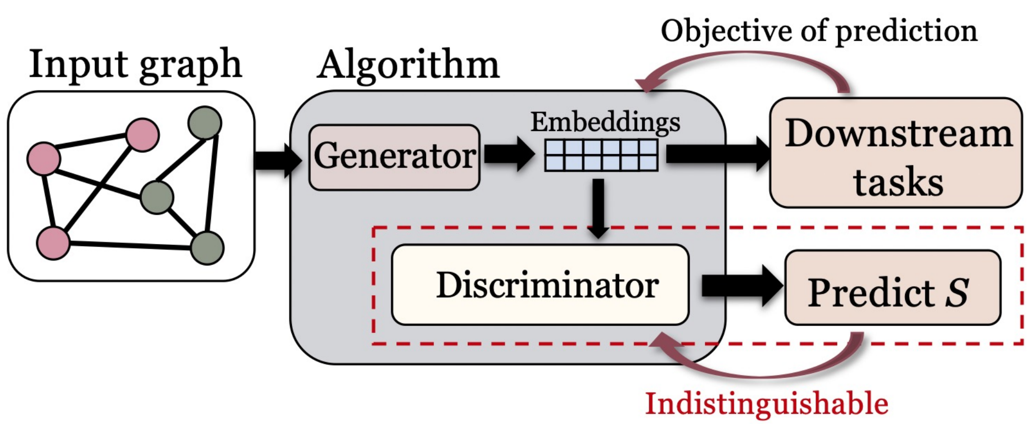




- Design fairness-aware regularizers to learn fair node embeddings

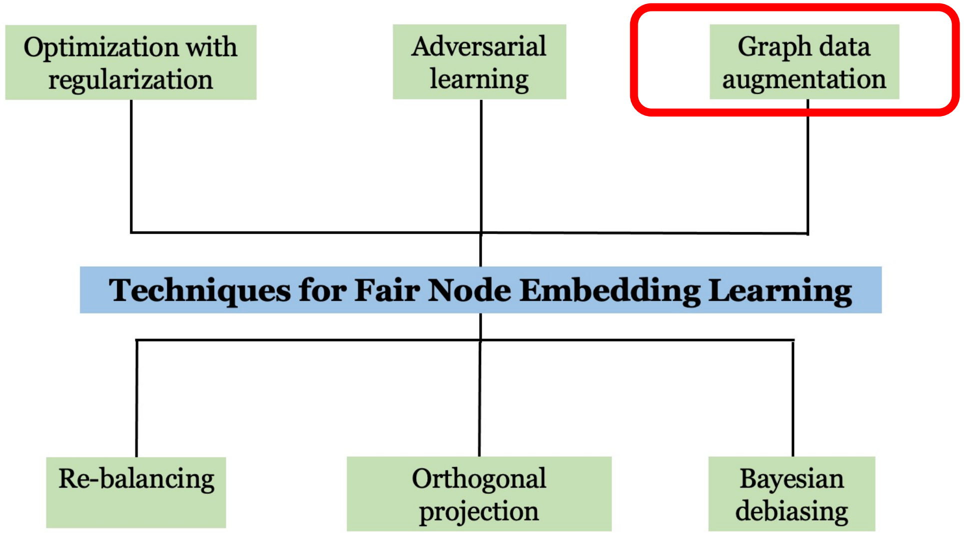
Adversarial Learning

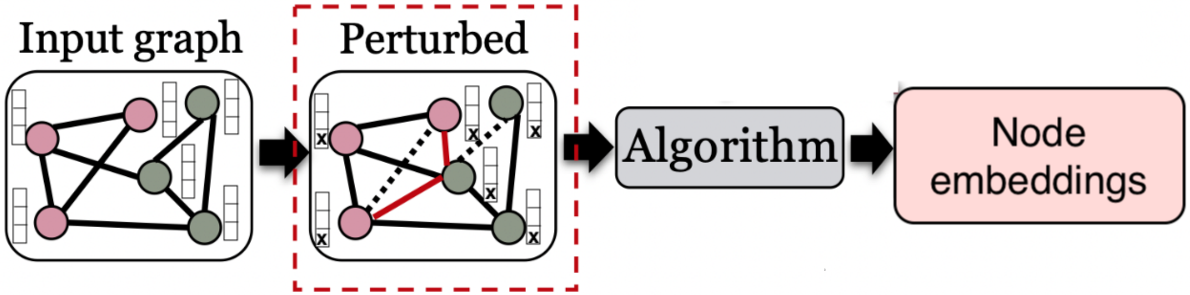




- Node embeddings whose sensitive attributes cannot be inferred by discriminator

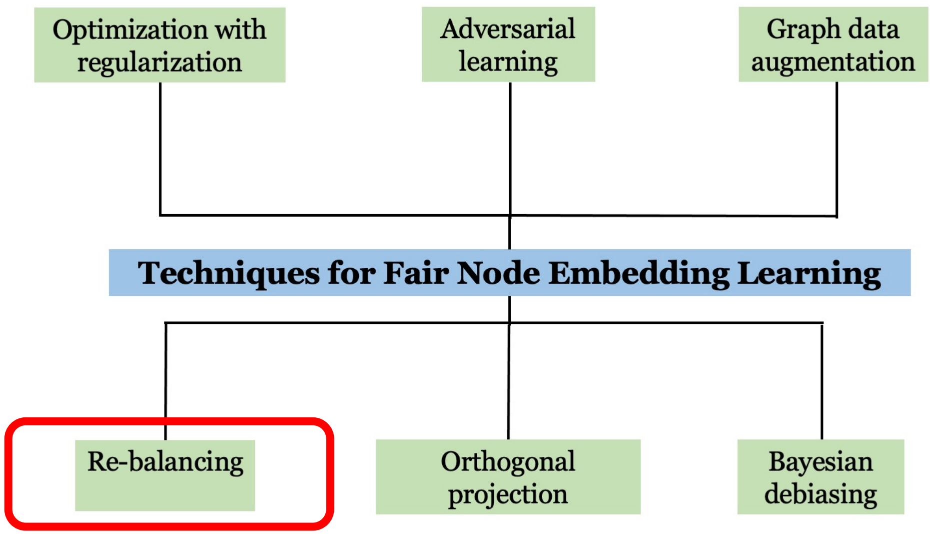
Graph Data Augmentation

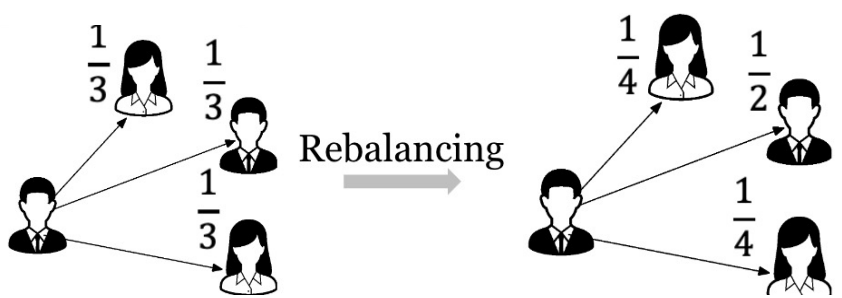




- Eliminate bias amplifying factors in graph structure and nodal features via augmentation design

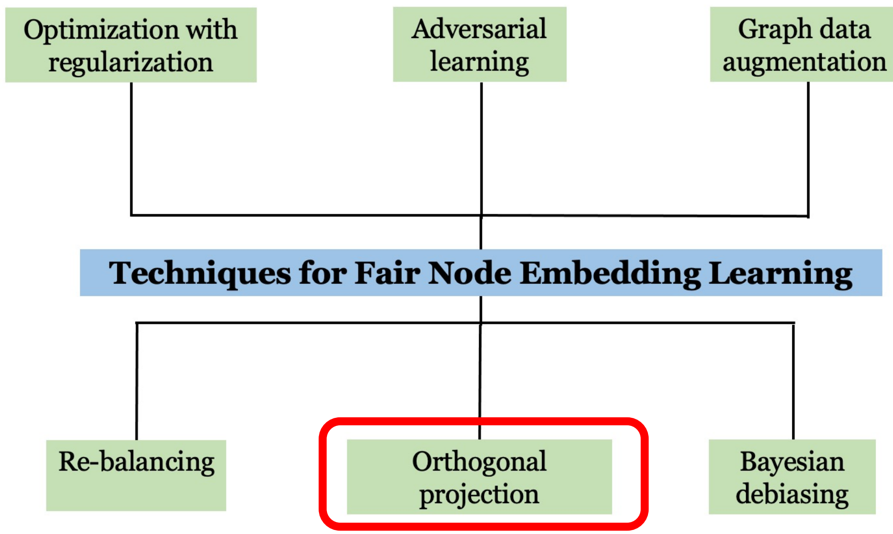
Re-balancing

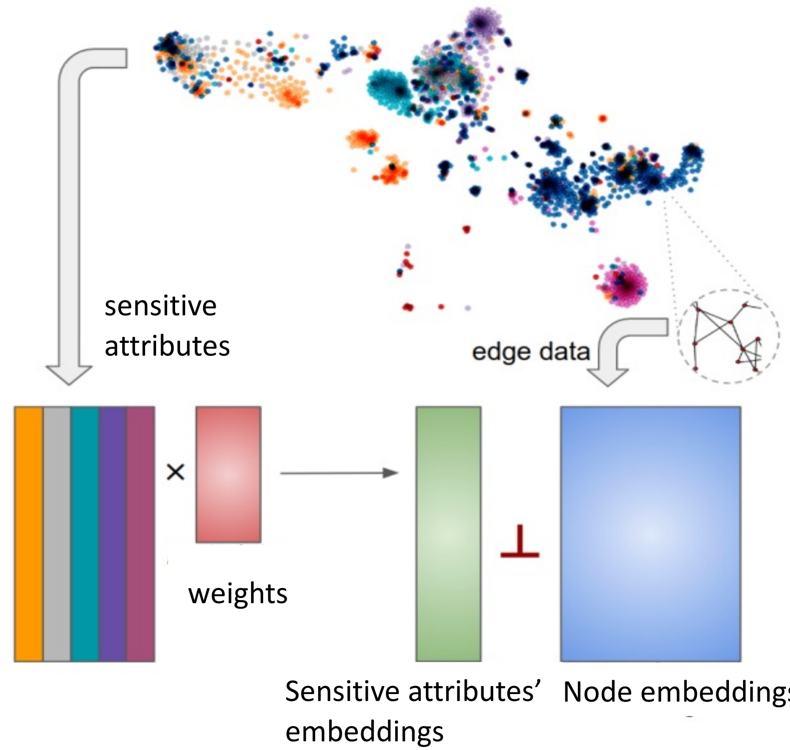




- Re-balance weights of edges without topology change
  - Balance inter and intra edges

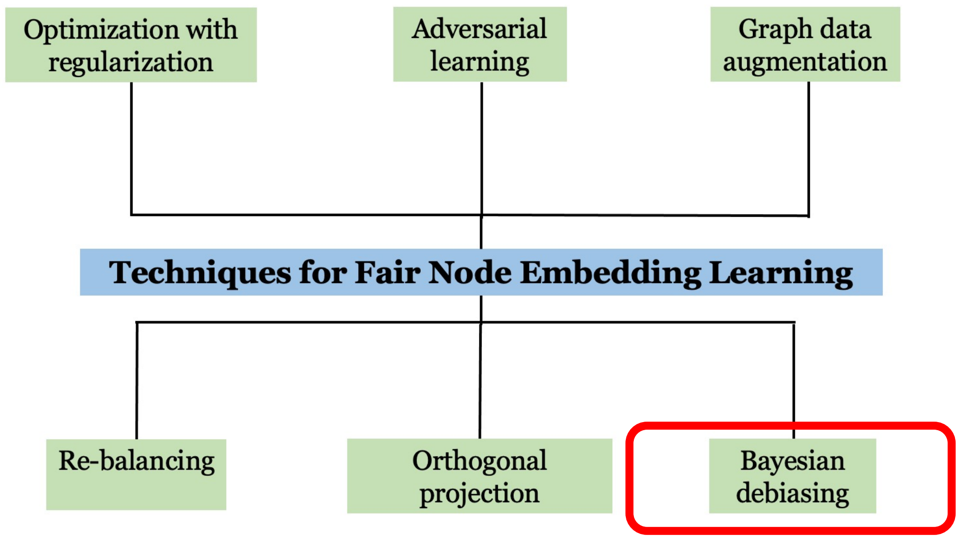
Orthogonal Projection

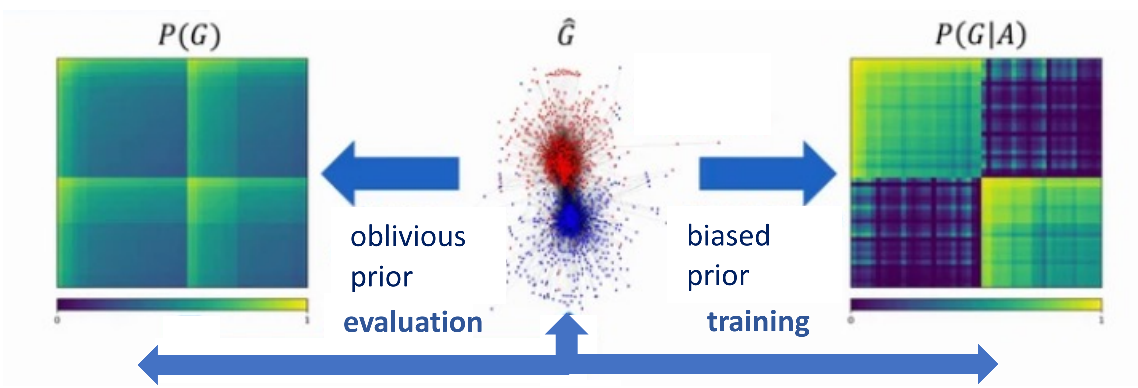




- Ensure linear independence between node embeddings and sensitive attributes' embeddings

Bayesian Debiasing



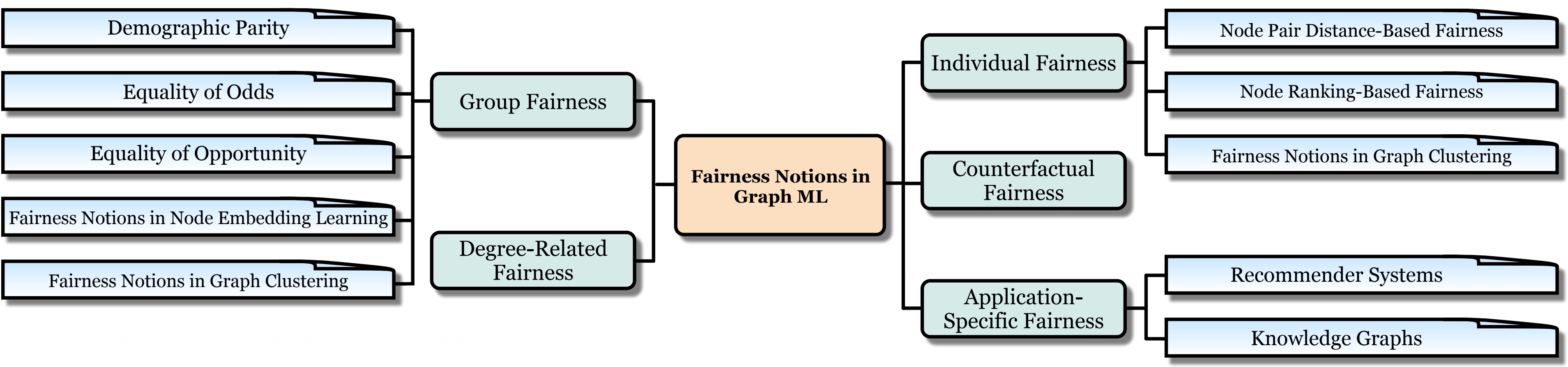


- Model sensitive information in prior distribution of graph
  - Node embeddings no longer need to represent sensitive information

Challenge 1: Insufficient Fairness Notions

- The Insufficiency of fairness notions

Can existing fairness notions help to avoid **all cases where people may feel unfair**?



Challenge 2: Multiple Fairness Notions

- The insufficiency of fairness notions
- Fulfilling multiple fairness notions

How to achieve **multiple types of fairness**?

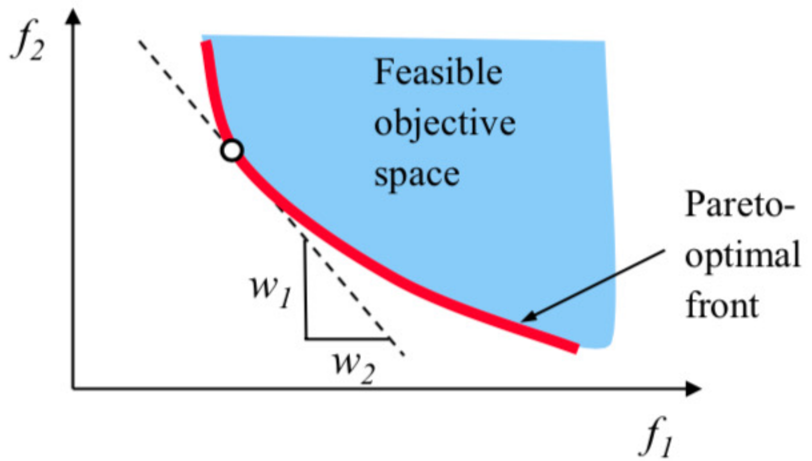
Are some of the existing fairness notions in **conflict** with each other?

If we could achieve multiple types of fairness, will people get a **stronger sense of fairness**? If not, what will be beneficial for social good?

Challenge 3: Fairness and Utility Tradeoff

- The insufficiency of fairness notions
- Fulfilling multiple fairness notions
- Balance fairness and model utility

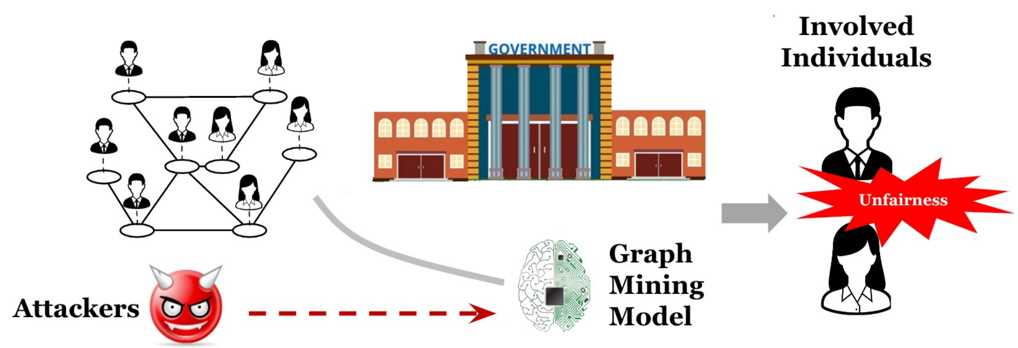
How to achieve fairness **at low or no cost of utility?**



Challenge 4: Robustness of Fairness

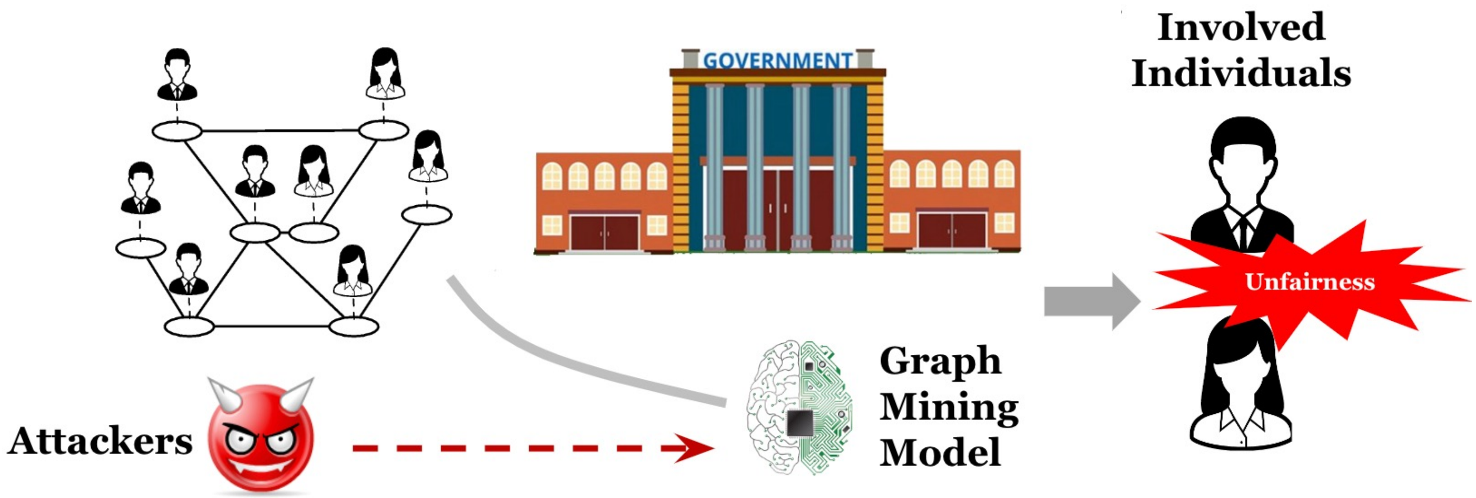
- The insufficiency of fairness notions
- Fulfilling multiple fairness notions
- Balance fairness and model utility
- Enhance robustness of fairness

Example: there are malicious attackers **whose goal is to induce bias** in the decisions made by the government



Challenge 4: Robustness of Fairness

- Enhance robustness of fairness
- Example: there are malicious attackers **whose goal is to induce bias** in the decisions made by the government



How would existing fairness-aware algorithms perform in terms of bias mitigation **under malicious attack**?

How to achieve **better robustness** in terms of fairness?

Challenge 5: Unavailable/Missing Sensitive Information

- The insufficiency of fairness notions
- Fulfilling multiple fairness notions
- Balance fairness and model utility
- Enhance robustness of fairness
- Bias mitigation strategy design without sensitive information

How to design fairness-aware algorithms with **missing sensitive information**?

How to define fairness when sensitive information is not fully available?

Challenge 6: Privacy

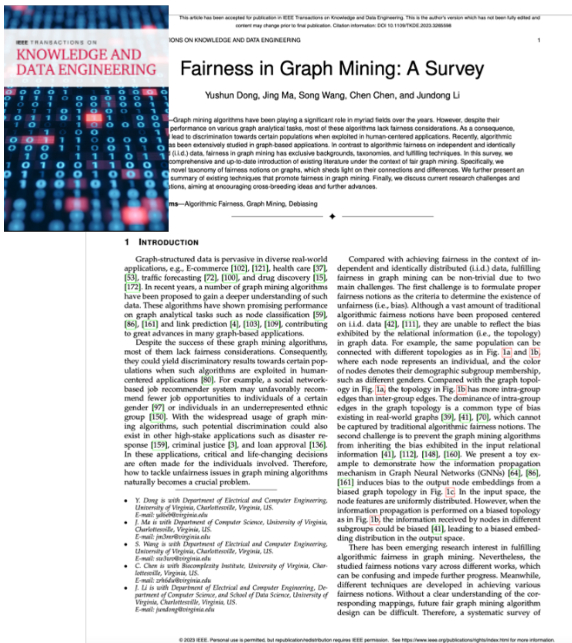
- The insufficiency of fairness notions
- Fulfilling multiple fairness notions
- Balance fairness and model utility
- Enhance robustness of fairness
- Bias mitigation strategy design without sensitive information
- Prevent sensitive information leakage

How much sensitive information can be retrieved in existing fairness-aware training strategies by different adversaries?

How to mitigate **sensitive information leakage** while training fair models?

Q&A

Related Materials:



This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering.

KNOWLEDGE AND DATA ENGINEERING

Fairness in Graph Mining: A Survey

Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li

1 INTRODUCTION

Our survey paper has been released on arxiv.

PyGDebias: 10+ popular algorithms and 20+ graph datasets.

Collected Algorithms

13 different methods in total are implemented in this library. We provide an overview of their characteristics as follows.

Methods	Debiasing Technique	Fairness Notions	Paper & Code
FairGNN [2]	Adversarial Learning	Group Fairness	[Paper] [Code]
EDITS [3]	Edge Rewiring	Group Fairness	[Paper] [Code]
FairWalk [4]	Rebalancing	Group Fairness	[Paper] [Code]
CrossWalk [5]	Rebalancing	Group Fairness	[Paper] [Code]
UGE [6]	Edge Rewiring	Group Fairness	[Paper] [Code]
FairVGNN [7]	Adversarial Learning	Group Fairness	[Paper] [Code]
FairEdit [8]	Edge Rewiring	Group Fairness	[Paper] [Code]
NIFTY [9]	Optimization with Regularization	Group/Counterfactual Fairness	[Paper] [Code]
GEAR [10]	Edge Rewiring	Group/Counterfactual Fairness	[Paper] [Code]
InFoRM [11]	Optimization with Regularization	Individual Fairness	[Paper] [Code]
REDRESS [12]	Optimization with Regularization	Individual Fairness	[Paper] [Code]
GUIDE [13]	Optimization with Regularization	Individual Fairness	[Paper] [Code]
RawlsGCN [14]	Rebalancing	Degree-Related Fairness	[Paper] [Code]

Website of our tutorial

UNIVERSITY of VIRGINIA

UCIRVINE UNIVERSITY of CALIFORNIA · IRVINE

Thanks for listening!