Debiasing Graph Neural Networks using Biased Contrastive Learning

Ashutosh Tiwari
@ashutiwa

Prof. Sadamori Kojaku
@skojaku

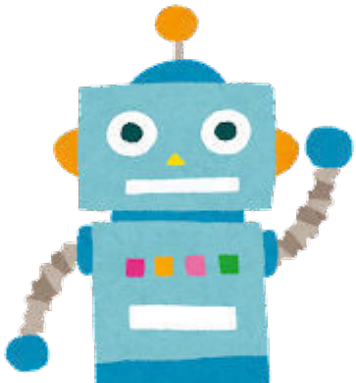
Prof. YY Ahn
@yyahn

Indiana University, Bloomington

April 5, 2024

Gender Bias in Google Translation

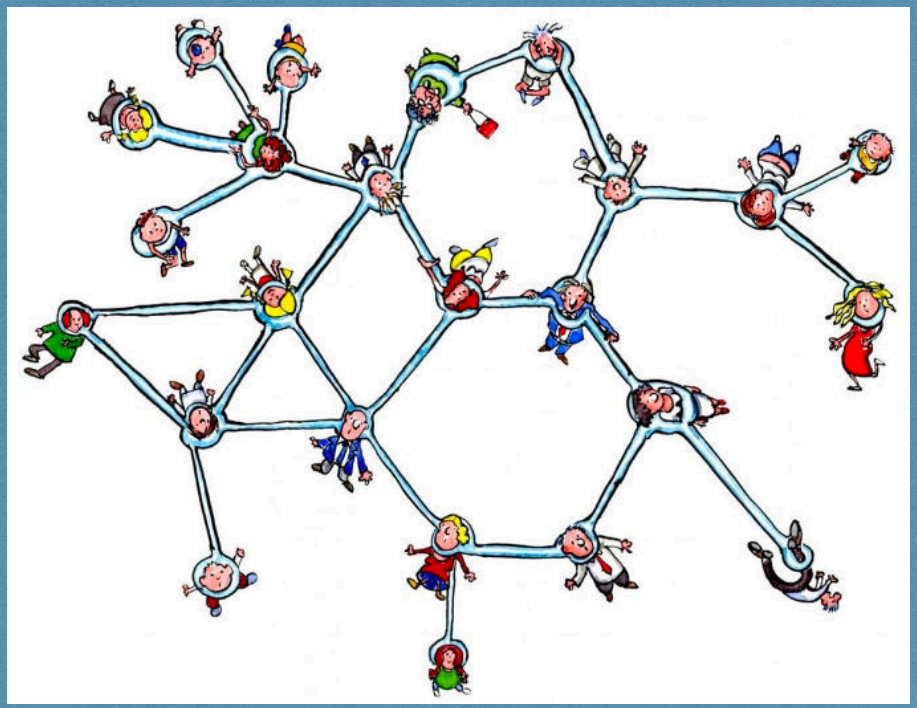
- E.g., when translating gender neutral Turkish sentences into English, Google associates he/she pronouns with stereotypically male/female dominated jobs, etc.



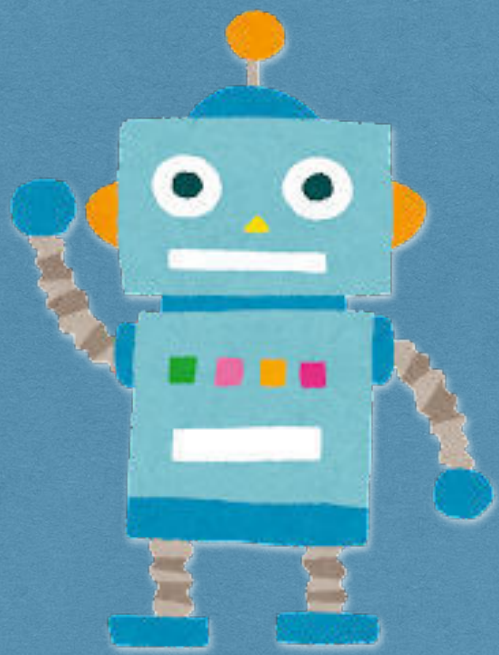
https://www.youtube.com/watch?v=fOlEKESnDv4

Networks & Bias in AI

Network data may contain sensitive node attributes



AIs may figure out sensitive information from network structure



Clever people have already developed debiasing methods for text and images.

Can we just apply them to graph data?

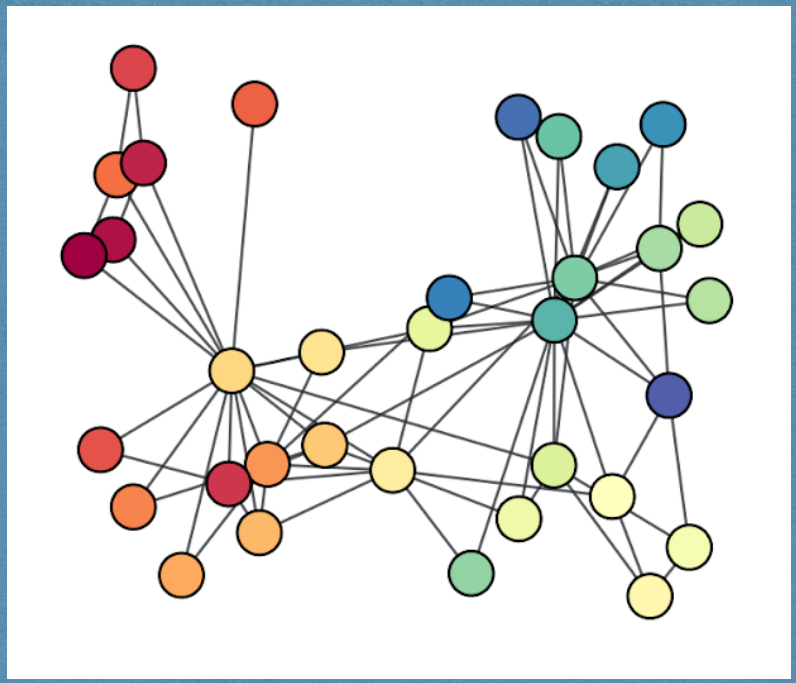
Debiasing: Output manipulation

T. Bolukbasi, et al. “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,” *NIPS*, 2016, vol. 29.

Debiasing: Adversarial Learning

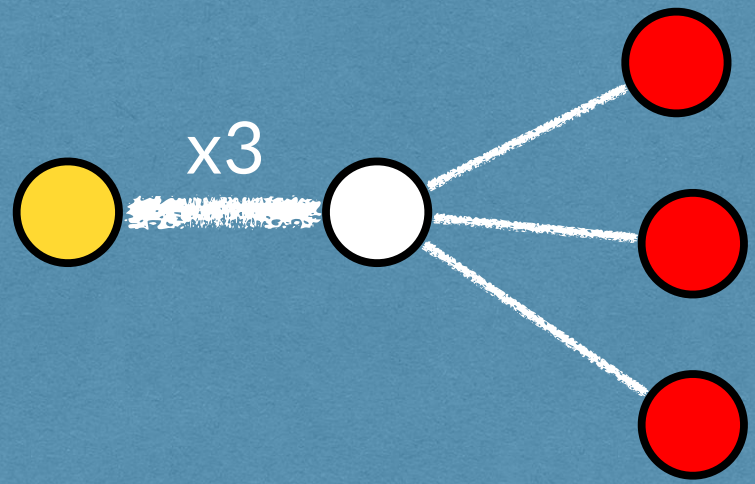
- In adversarial learning, one trains an embedding model and an adversarial model simultaneously.
- The adversarial model attempts to extract any biased features from the embedding model, while the embedding model generates a new embedding that is resistant to the adversary's attempts to extract the biased features.

Debiasing: Input manipulation



FairWalk

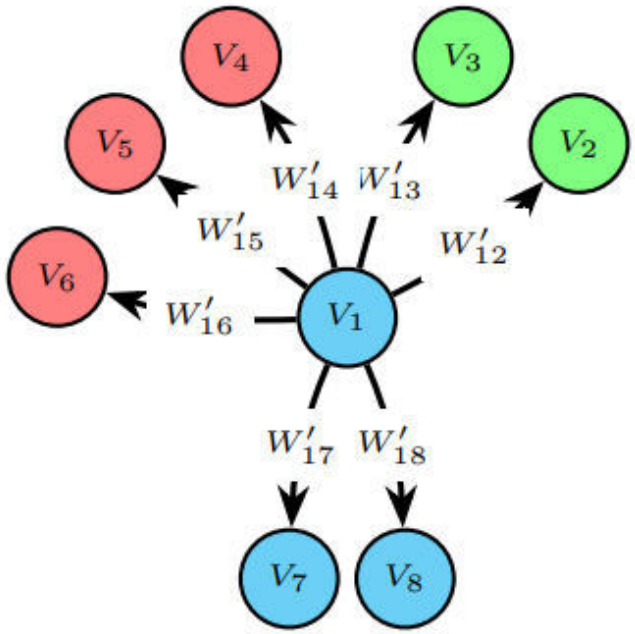
(Tahleen Rahman et al. IJCAI)



CrossWalk

(Ahmad Khajehnejad et al. AAAI)

Crosswalk Illustration



$$w'_{12} + w'_{13} = \frac{\alpha}{2}$$

$$w'_{14} + w'_{15} + w'_{16} = \frac{\alpha}{2}$$

$$w'_{17} + w'_{18} = 1 - \alpha$$

$$\frac{w'_{14}}{w'_{15}} = \frac{w_{14}\ m(v_4)^p}{w_{15}\ m(v_5)^p}$$

$$\frac{w'_{17}}{w'_{18}} = \frac{w_{17}\ m(v_7)^p}{w_{18}\ m(v_8)^p}$$

But non of them are satisfactory ...

Input/Output manipulation

- May not remove bias or even introduce new bias

Adversarial learning

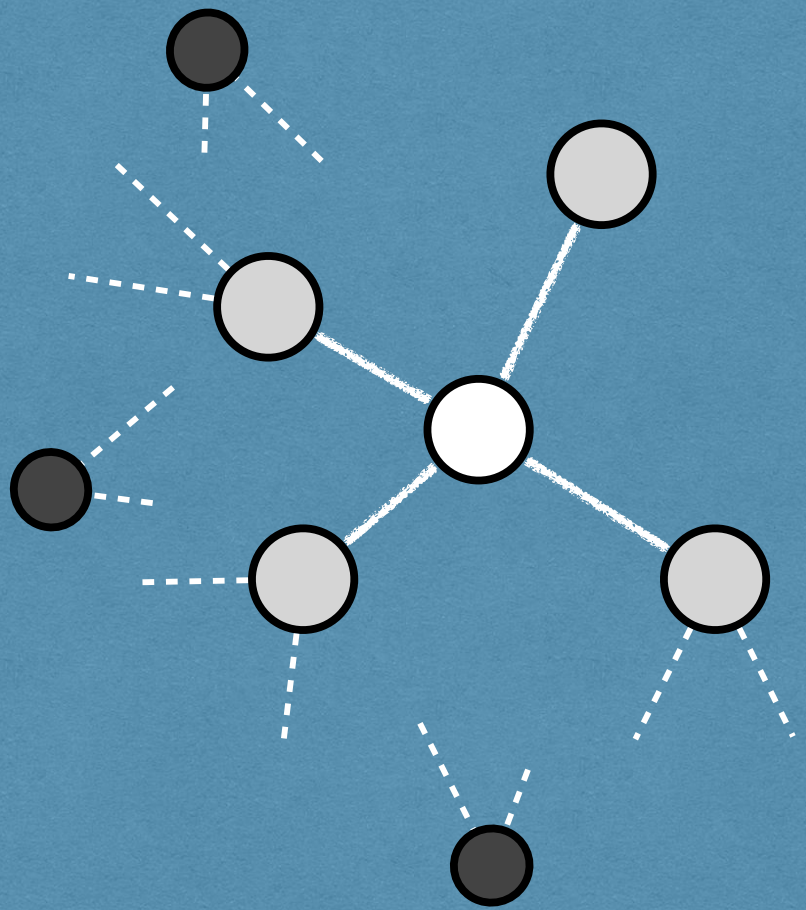
- Non-trivial hyperparameter tuning

- Unstable (because adversarial learning has two learning modules that keep adapting to the other)

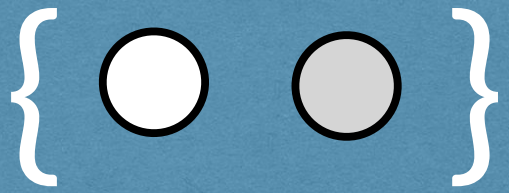
We provide a simple yet powerful alternative that utilizes an organic debiasing capacity of contrastive learning.

- No manipulation
- No hyperparameters
- Ensured to learn unbiased representation

Contrastive learning

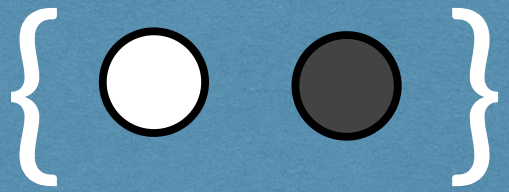


Adjacent



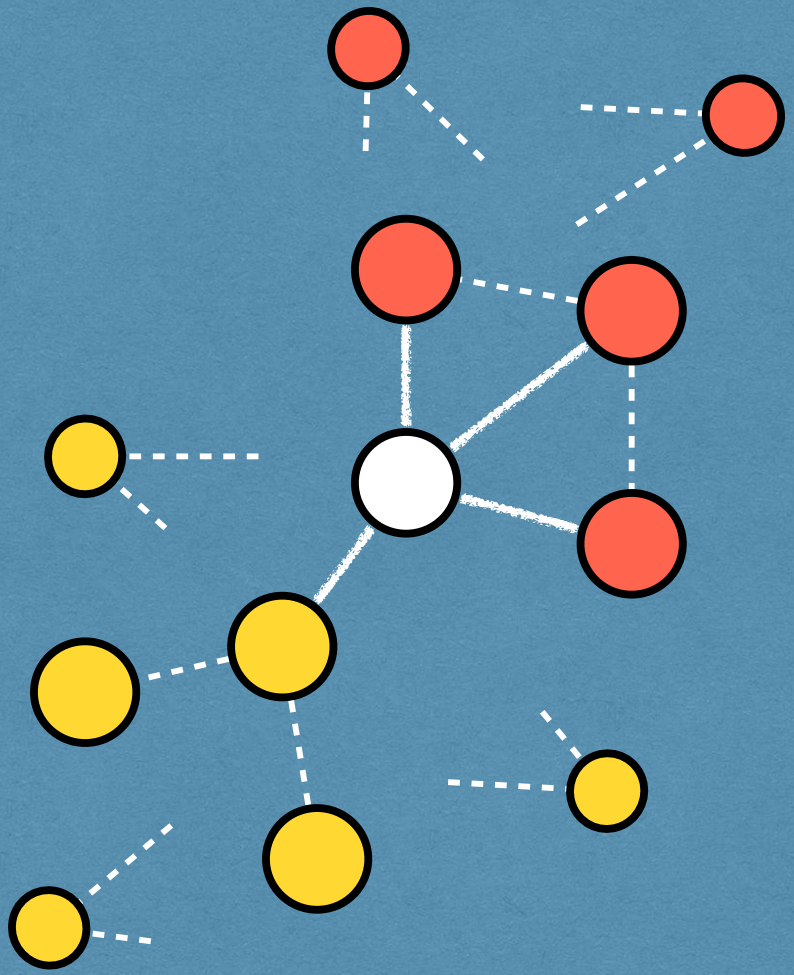
Make them closer

Non-adjacent

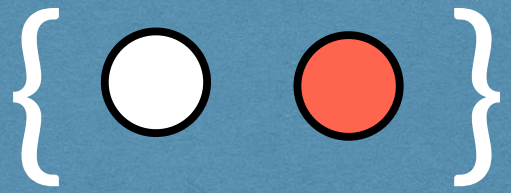


Make them distant

Contrastive learning

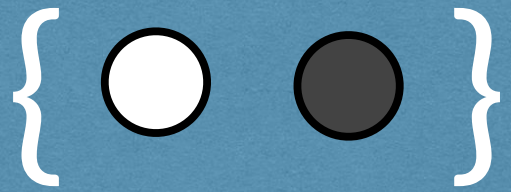


Adjacent



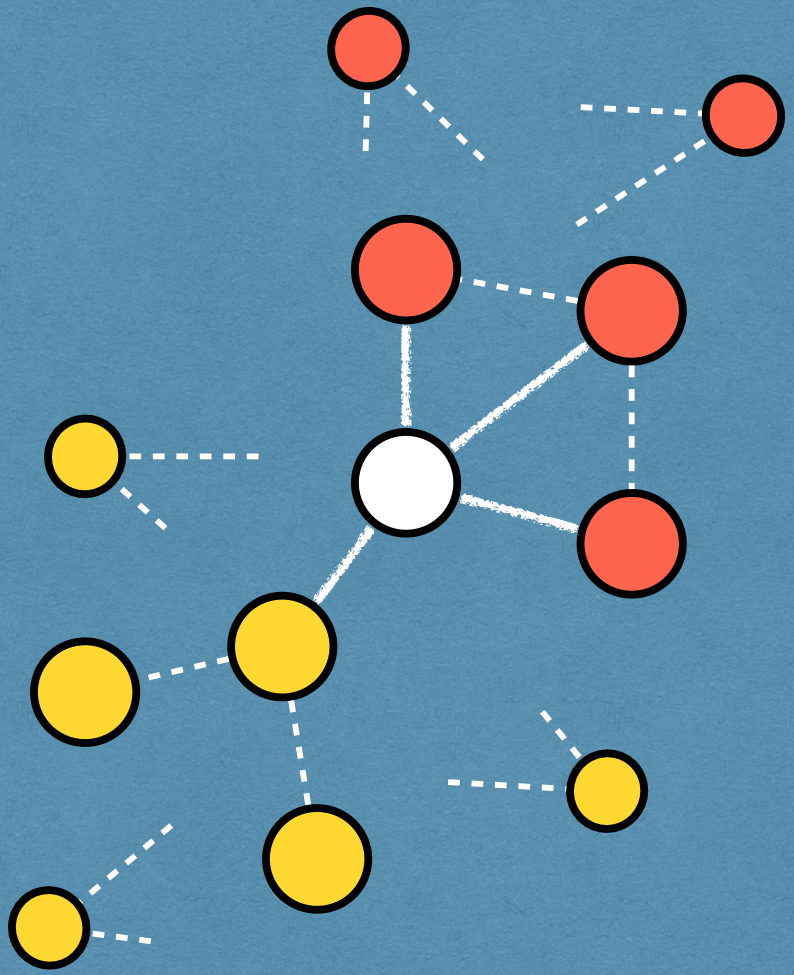
Make them closer

Non-adjacent

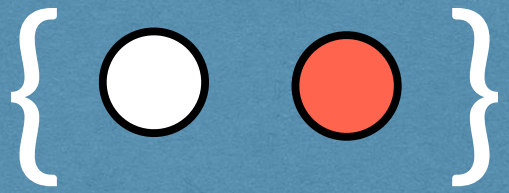


Make them distant

Biased Contrastive learning

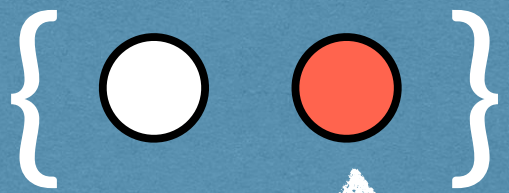


Adjacent



Make them closer

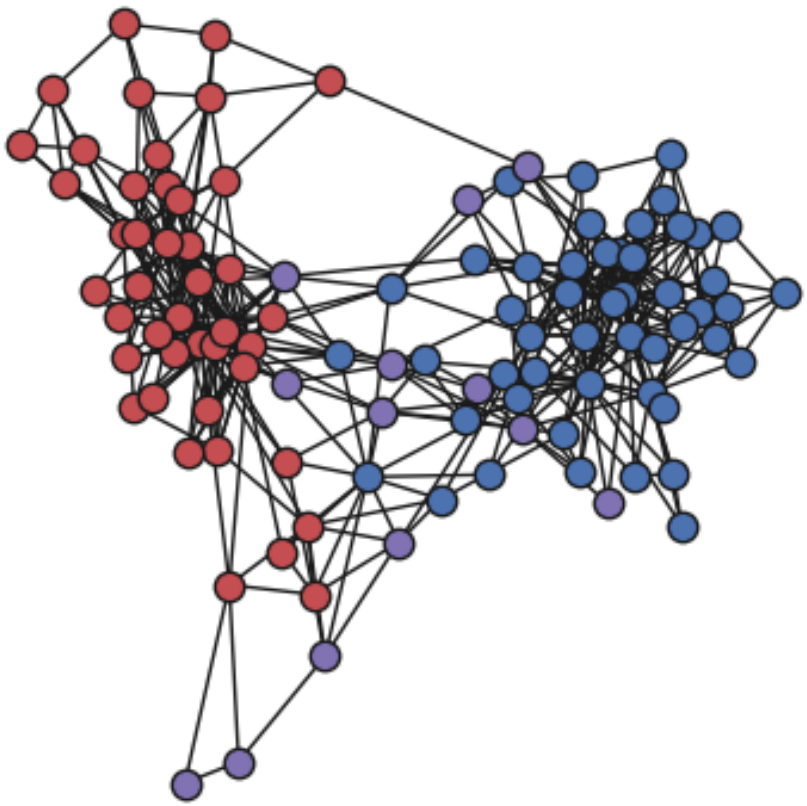
Non-adjacent



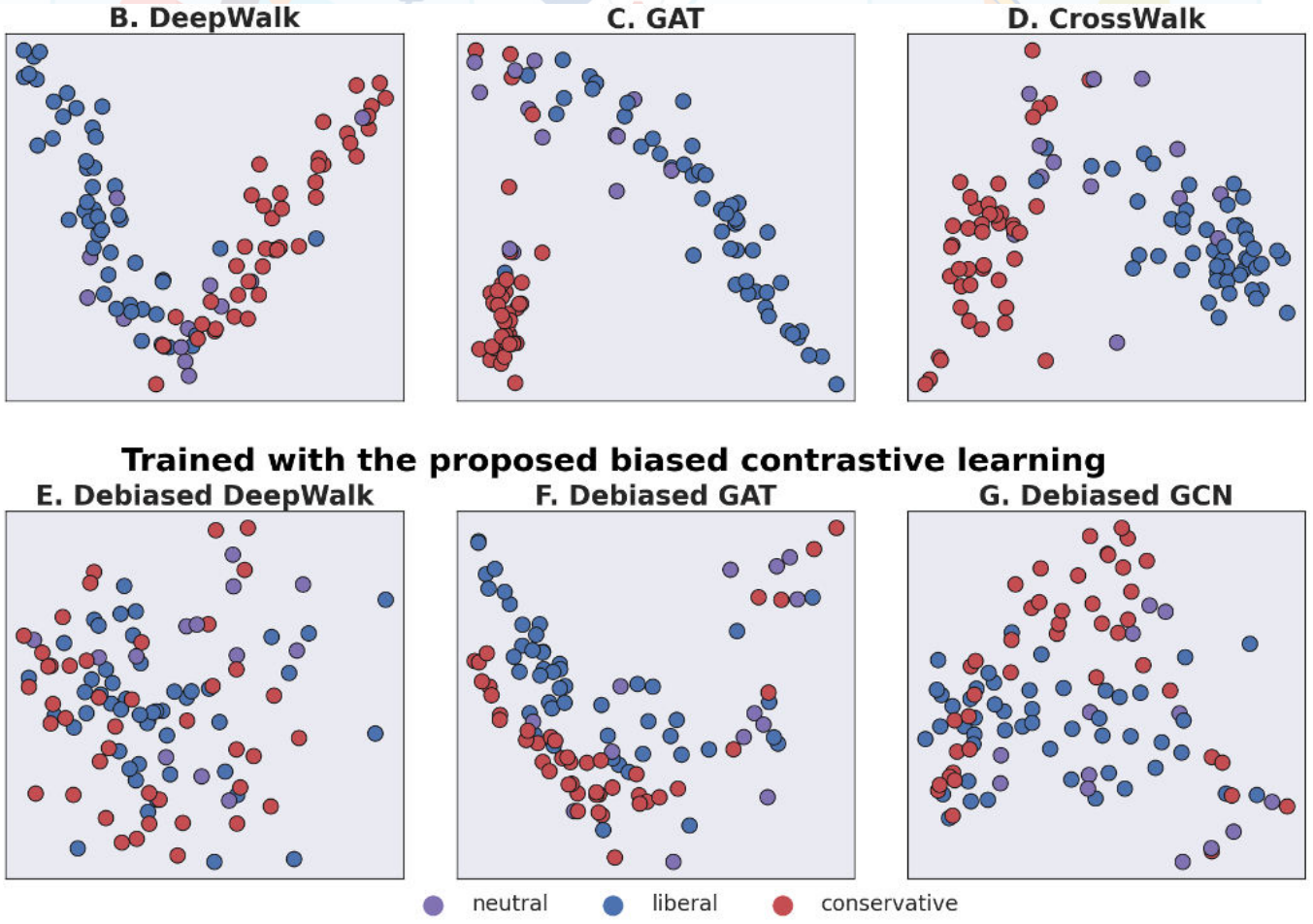
Make them distant

Sample x3 times more likely than

Our Method: Demonstration (Political Blogs Dataset)



Our Method: Demonstration(Political Blogs Dataset)

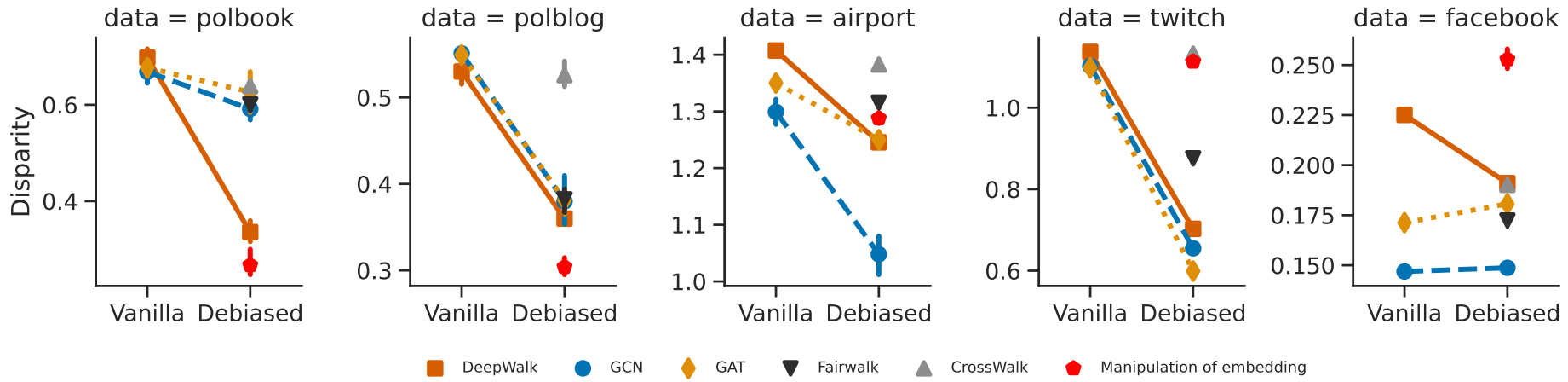


Biased contrastive learning is a training framework.

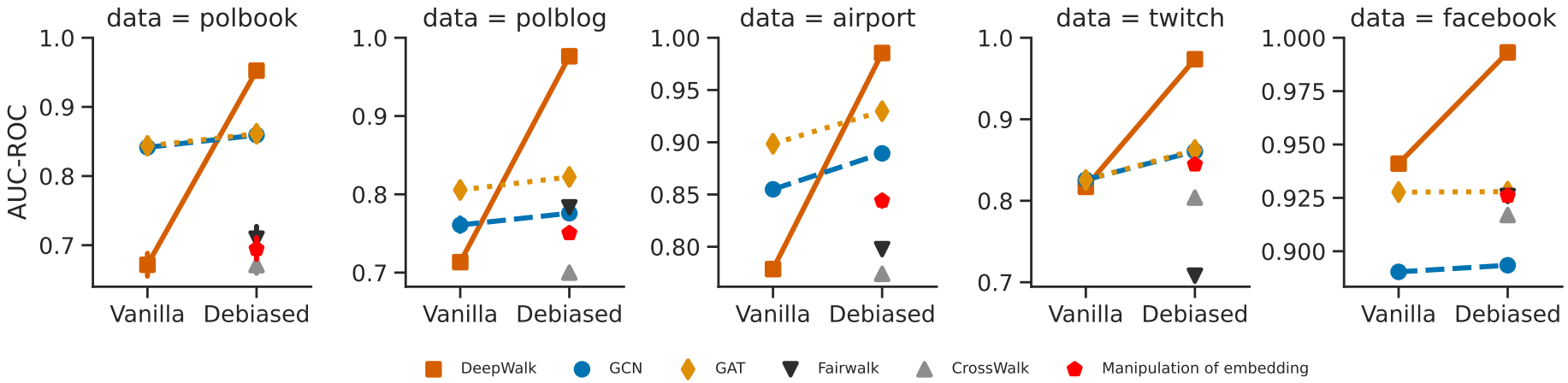
Agnostic to model architecture

Applicable to DeepWalk, node2vec, GCN, GAT, GraphSAGE etc.

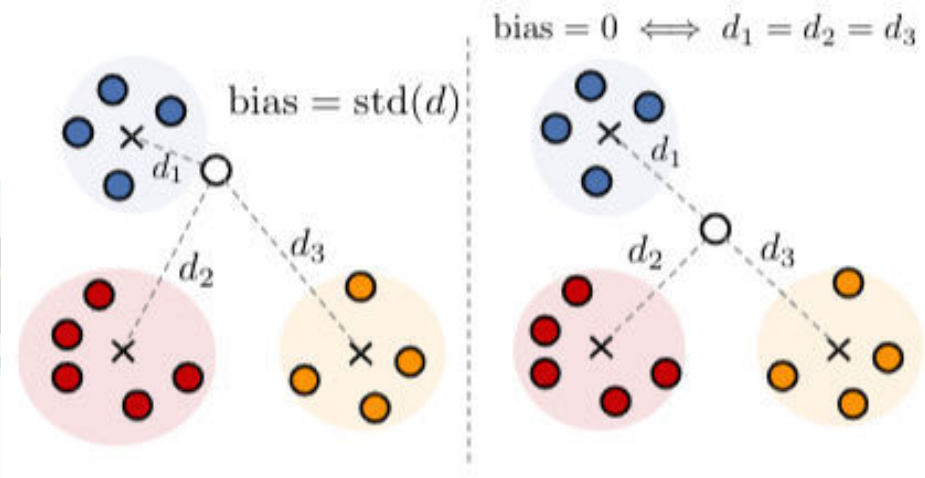
Disparity

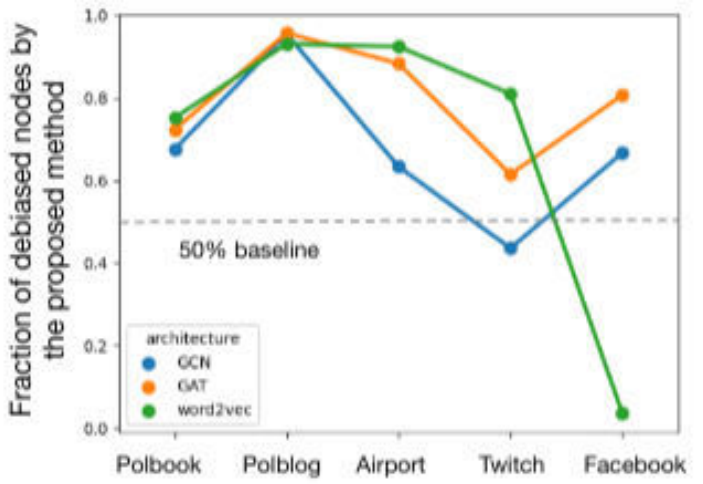

data = polbook
data = polblog
data = airport
data = twitch
data = facebook
Disparity
0.6
0.4
0.5
0.4
0.3
1.4
1.3
1.2
1.1
1.0
1.0
0.8
0.6
0.250
0.225
0.200
0.175
0.150
Vanilla
Debiased
DeepWalk
GCN
GAT
Fairwalk
CrossWalk
Manipulation of embedding

Prediction Accuracy



Bias per node





- Bias here is defined here as std dev of distances from centroids of all groups.
- Plot shows the ratio of # of nodes for which this disparity decreases.

Summary

- Introduced *biased* contrastive learning for debiasing
- Demonstrated the effectiveness with link prediction task on four different networks

Applications

- Mitigating AI biases in recommendation, search engine, and ranking algorithms
- Control or remove inconsequential noises and biases in scholarly embedding

Previous Work: Indiana University, Bloomington

- was trying to find out if there is non-linearity to the bias structure.
- Plan was to propose an alternative to WEAT (Word Embedding Association Test) which is a linear test.