A biased contrastive learning debiases graph neural networks

Ashutosh Tiwari, Yong-Yeol Ahn, *Sadamori Kojaku

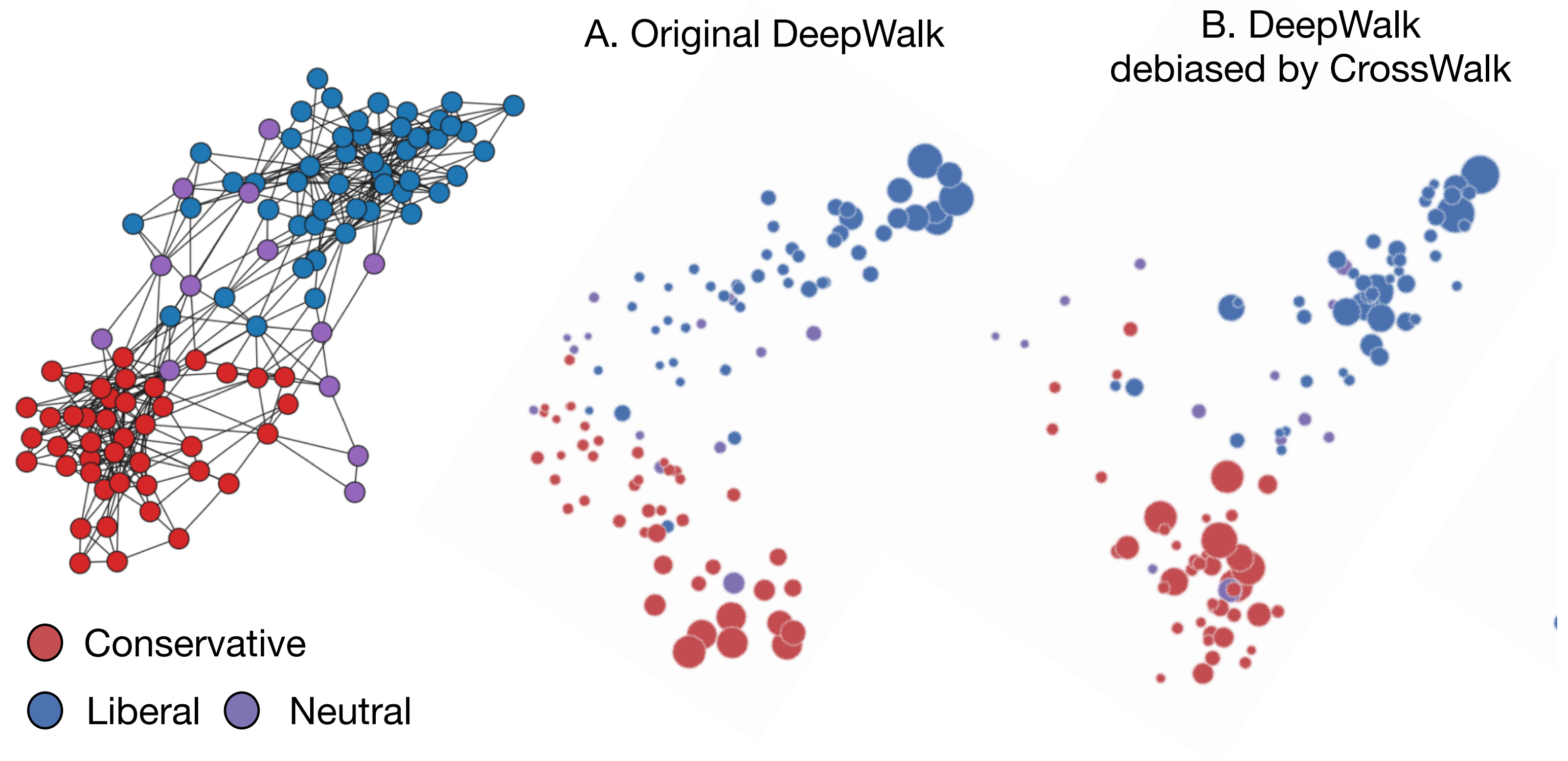
skojaku@iu.edu

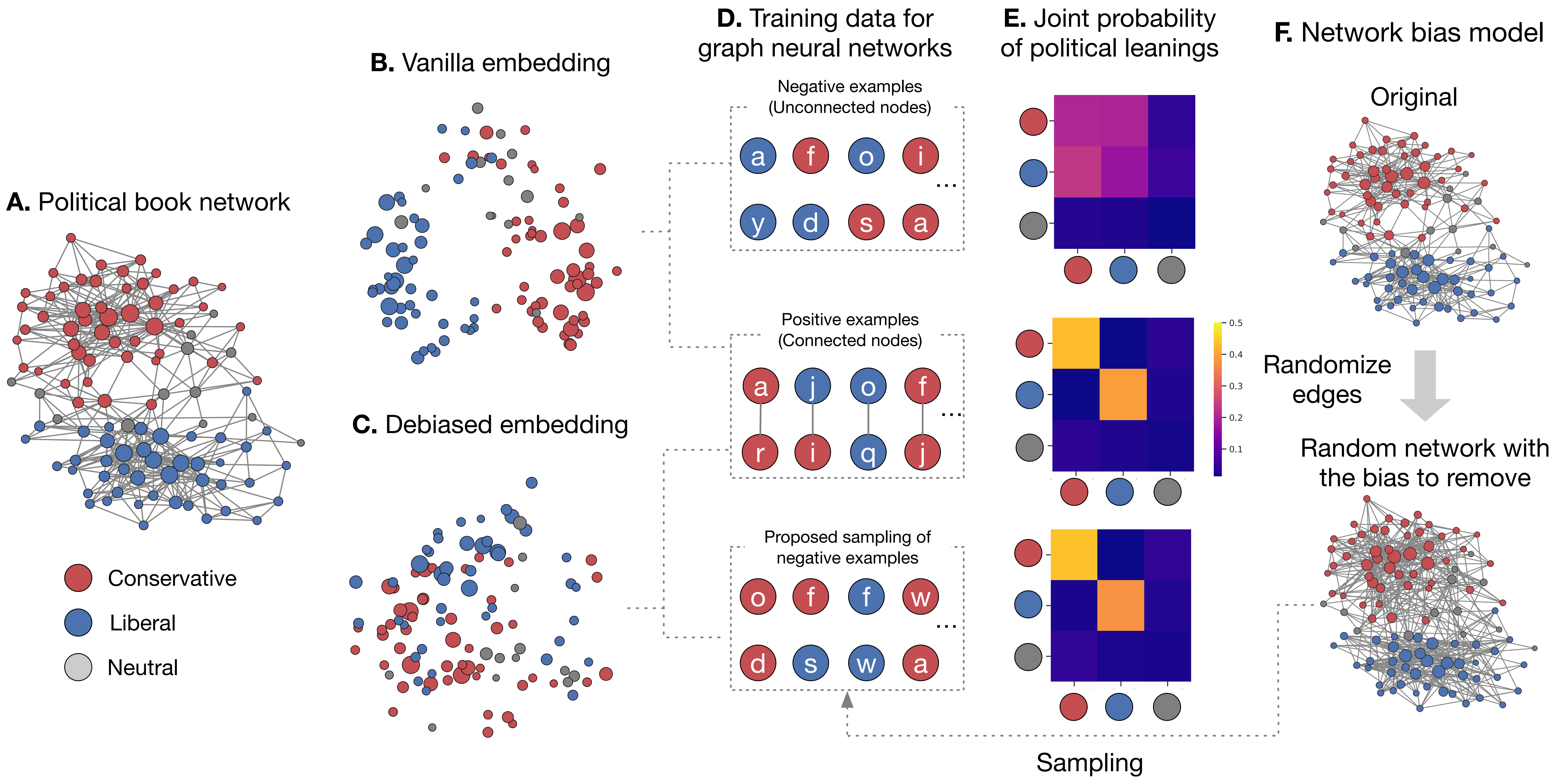
Center for Complex Networks and Systems Research, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, US.

Our research is supported by the Air Force Office of Scientific Research under award numbers FA9550-21-1-0446.

Persistent bias in network embedding

- Graph data are ubiquitous data type as text and image data. Yet, little is known about how it may perpetuate social inequalities when combined with AIs.
- For example, professional recommendation systems trained on biased social networks can lead to biased recommendations, reinforcing racial and gender homophily, amplifying echo chambers, and restricting job opportunities for marginalized groups.
- Many AI systems utilize graph data through graph embedding, a vector representation of networks. Despite efforts to debias graph embedding and achieve unbiased AI, the bias is still persistent (right figure), underscoring the need for a comprehensive theoretical framework to effectively create debiased representations.



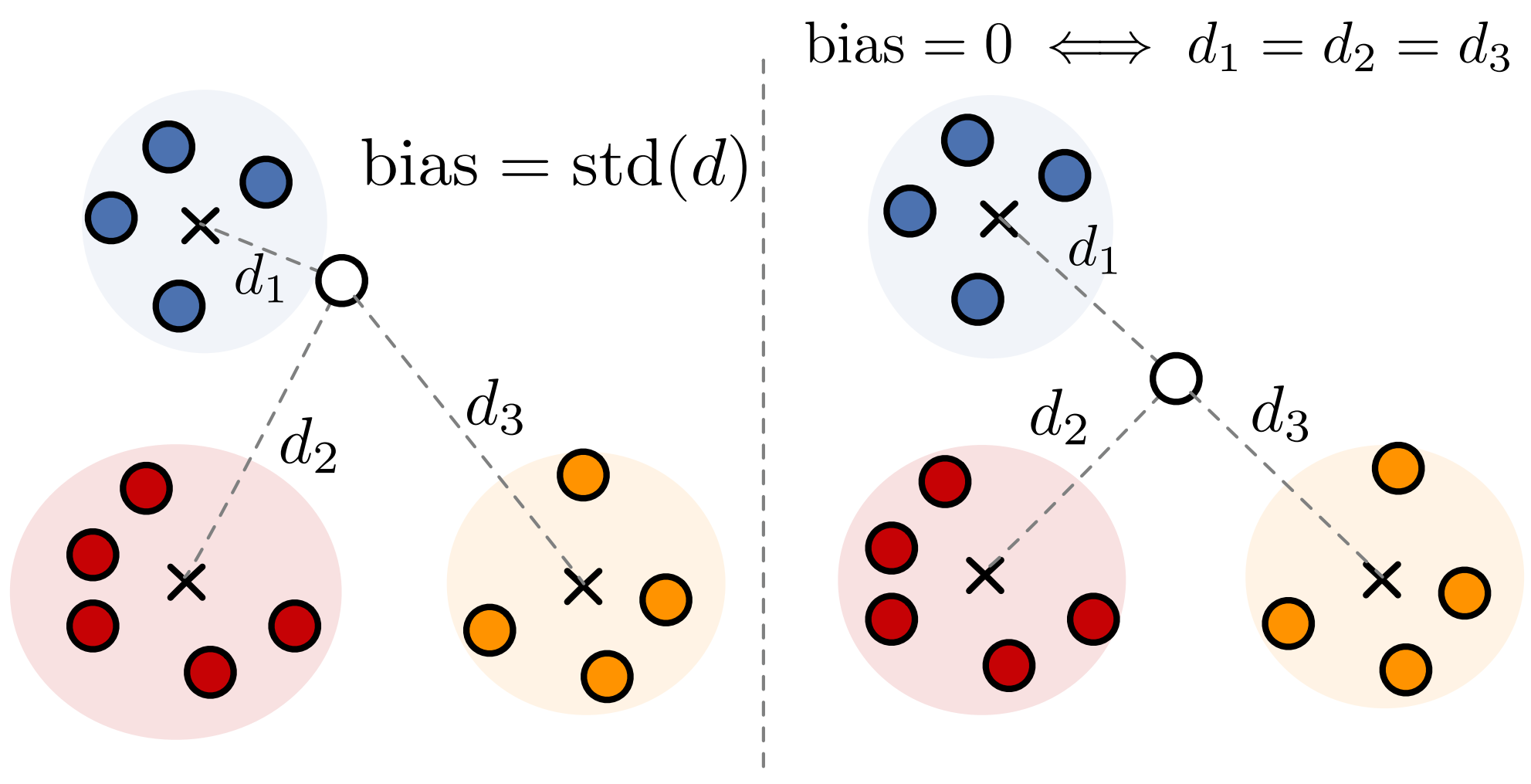

A. Political book network
B. Vanilla embedding
C. Debiased embedding
Conservative
Liberal
Neutral
D. Training data for graph neural networks
Negative examples (Unconnected nodes)
a f o i y d s a
Positive examples (Connected nodes)
a j o f r i q j
Proposed sampling of negative examples
o f f w d s w a
E. Joint probability of political leanings
0.5
0.4
0.3
0.2
0.1
F. Network bias model
Original
Randomize edges
Random network with the bias to remove
Sampling

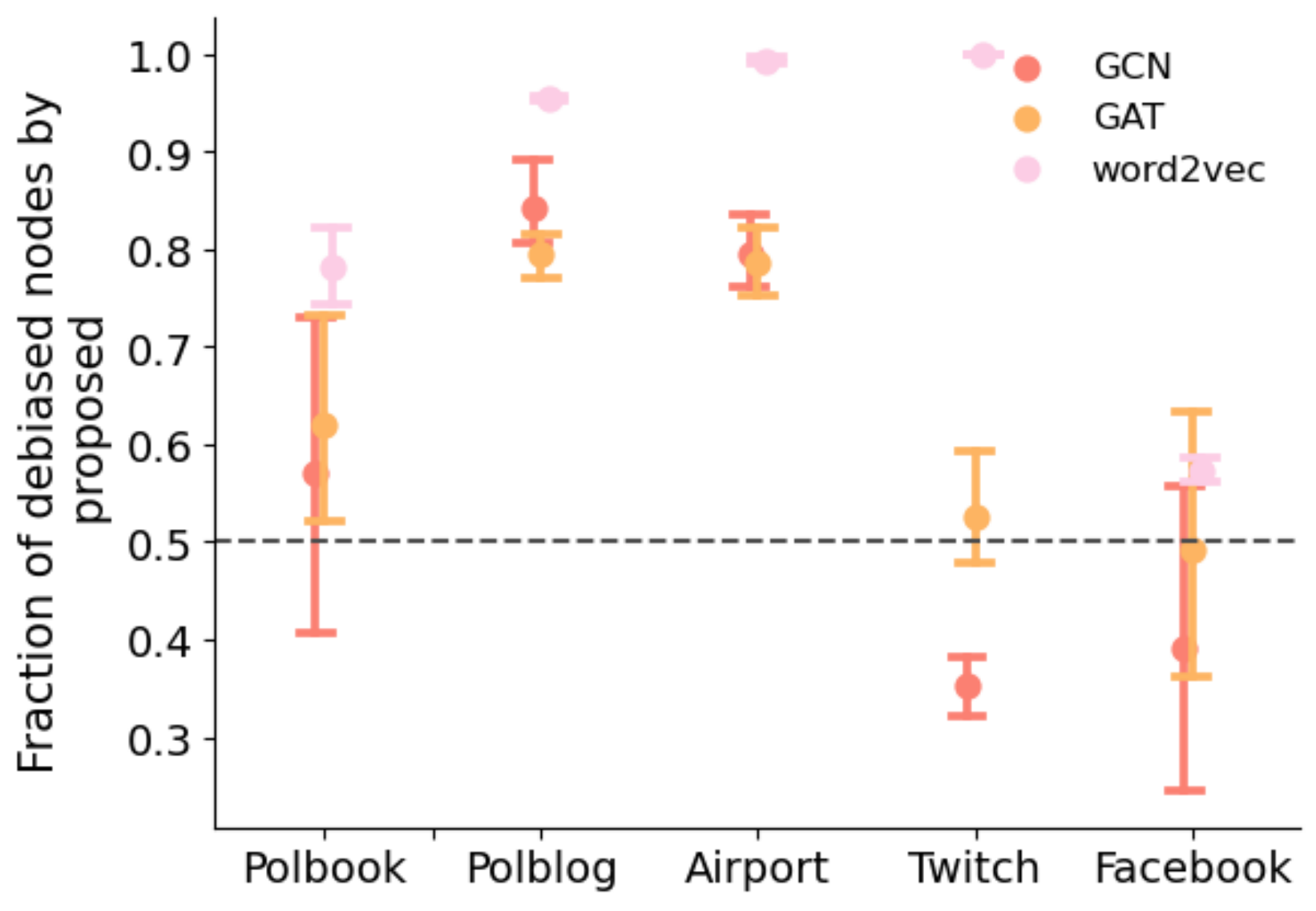
Biasing for debiasing

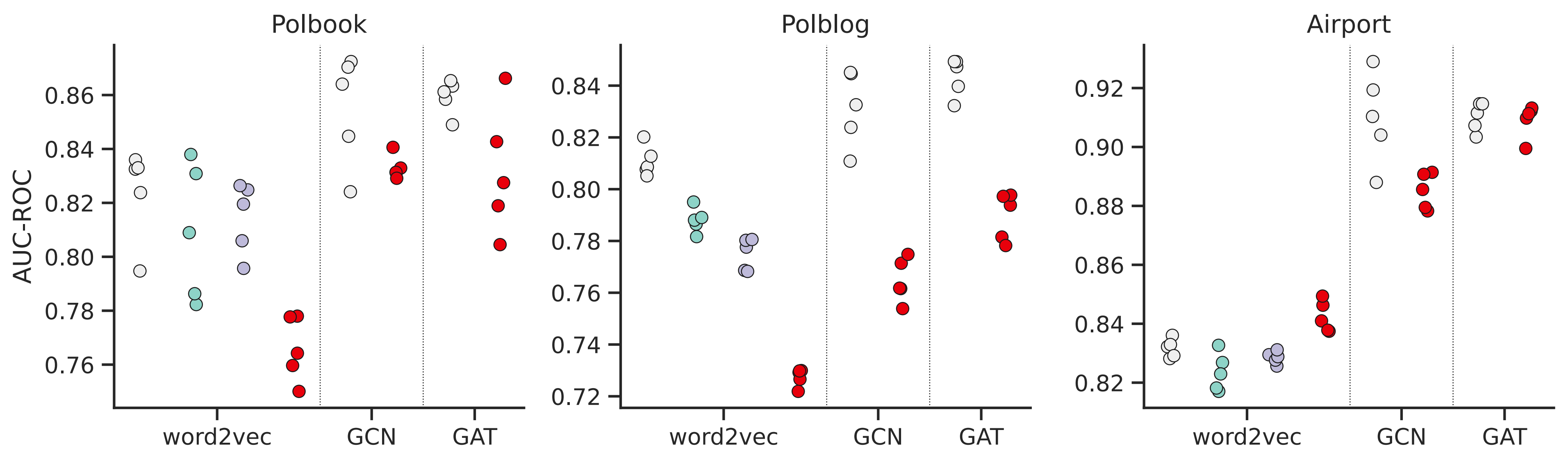
- We focused on contrastive learning, a widespread framework to train neural networks.
- With contrastive learning, the model learns the unique characteristics of the given data (positive examples) by contrasting it with a random data, i.e., negative examples.
- In graph embedding, the positive examples are the edges in the network, and the negative examples are usually randomly sampled node pairs (Fig. D).
- We bias the negative examples to have the same bias in the given data, making the positive and negative examples indistinguishable by the focal bias (e.g., political leaning in Fig. E).
- We do so by sampling the negative examples from a random network with the same group homophily and degree heterogeneity (dcSBM; Fig. F).

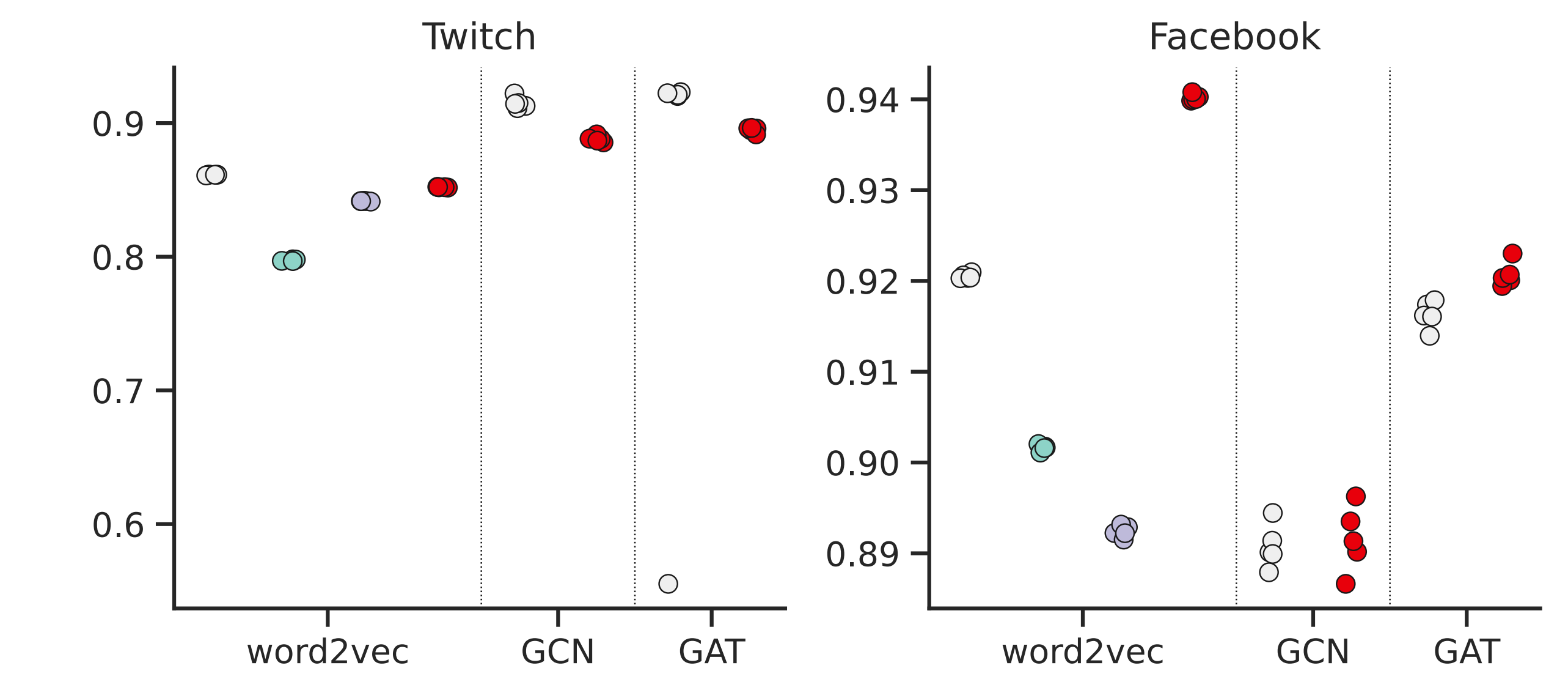
Bias reduction

- We quantify the bias in graph embedding by going through each node and quantify the bias as the variance of the distances to each protected group (e.g., political leaning and gender).
- Our method reduces the variance for the majority of nodes for the 12 out of 15 combinations of the neural network types and data, demonstrating the effectiveness across data and neural network models.









Network	Nodes	Edges	Protected category (# of categories)	Information Modularity
Political Books	105	441	Political leaning (3)	0.061
Political Blogs	1224	16,715	Political leaning (3)	0.0925
Airport	2898	15,564	Geographical regions (5)	0.112
Twitch	168,114	6,797,557	Language (20)	0.05
Facebook	41,554	1,362,229	Gender (3)	−0.004

Link prediction benchmark

- We test if the debiased embedding maintains features useful for link prediction, which is a basis of many applications, e.g. recommendations, and fraud detection.
- We compare our method with the previous methods, e.g., Fairwalk and Crosswalk, which are based on the word2vec neural network.
- The proposed methods yield a comparable or higher link predictability for three out of five networks than Fairwalk and CrossWalk.
- Furthermore, the link predictability increases by combining with a more powerful neural network models such as GCN and GAT.
- Overall, our debiasing method offers an effective and more flexible framework for debiasing graph embedding.