Topological Bias Mitigation via Graph Rewiring

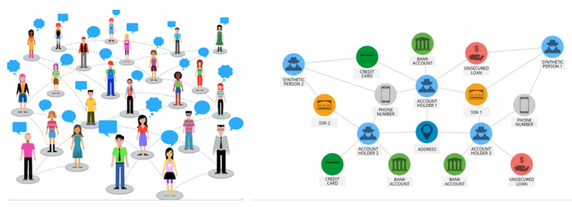
O. Deniz Kose[1], Gonzalo Mateos[2], and Yanning Shen[1]

[1]University of California, Irvine, Dept. of Electrical Engineering and Computer Science

[2]University of Rochester, Dept. of Electrical and Computer Engineering

Motivation

- Connectivity era: Growing amount of data describing interconnected systems



- Graphs are utilized to model such complex data
  - Graph nodes: users in social networks, accounts holding money
  - Graph edges: friendship between users, money transactions
  - Nodal features: education level of users, locations of accounts
- Processing & learning from graph data can provide significant advancements
  - Increasing attention towards graph signal processing & ML over graphs
  - Cross-pollination of GSP and ML over graphs provides new insights [2]
- ML algorithms propagate algorithmic bias
  - Impact of ethnicity in crime prediction
  - Impact of gender in ad recommendation

- Use of network connectivity in learning amplifies existing bias [3]
- Motivation: Consideration of bias is necessary for graph-based learning
- Limitation of current works: Task/algorithm-specific, no theoretical analysis
- Innovation: Leverage GSP tools to design a general-purpose bias mitigation strategy

Preliminaries & Problem Statement

- Focus on undirected graphs, $\mathcal{G} := (\mathcal{V}, \mathcal{E})$
- Connectivity information described via graph adjacency $\mathbf{A} \in \{0,1\}^{N\times N}$ and normalized Laplacian matrices $\mathbf{L} = \mathbf{I}_N - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$
- Sensitive attributes $\mathbf{s} \in \{-1,1\}^N$, nodal features $\mathbf{X} \in \mathbb{R}^{N\times F}$ and labels $\mathbf{y} \in \{-1,1\}^N$ for node classification
- Graph Fourier Transform of signal $\mathbf{z} \in \mathbb{R}^N$ is $\tilde{\mathbf{z}} = \mathbf{V}^\top\mathbf{z}$, where $\mathbf{L} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$
- Filtering graph signal $\mathbf{z} \in \mathbb{R}^N$ via a filter with frequency response $\tilde{\mathbf{h}} := [\tilde{h}_1, \ldots, \tilde{h}_N]^\top$ yields the output signal $\mathbf{z}_{out} = \mathbf{V}\,\mathrm{diag}(\tilde{h}_1, \ldots, \tilde{h}_N)\tilde{\mathbf{z}}$.

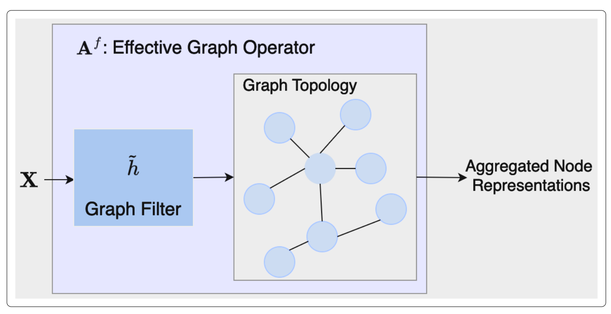
Problem Statement

Given $\mathcal{G}$ and $\mathbf{s}$, design of graph filters with frequency response $\tilde{\mathbf{h}} \in \mathbb{R}^N$, so that algorithmic bias sourced from graph topology can be attenuated with the application of such filters.

Bias Metric

- Aggregation over graph $\mathcal{G}$

$$\mathbf{R}^f = \hat{\mathbf{A}}\mathbf{V}\mathrm{diag}(\tilde{\mathbf{h}})\mathbf{V}^\top\mathbf{X} = \mathbf{A}^f\mathbf{X},$$
$$\mathbf{A}^f := \mathbf{V}(\mathbf{I}_N - \mathbf{\Lambda})\mathrm{diag}(\tilde{\mathbf{h}})\mathbf{V}^\top$$



- Novel unsupervised bias measure: $\rho := \|\mathbf{s}^\top\mathbf{A}^f\|_2$, can be manipulated via filter design

Linear Programming-based Optimal Filter Design

- Proposition: Bias metric ρ, can be upper bounded by:

$$\rho \leq \sqrt{N}\sum_{i=1}^{N}|\tilde{s}_i||(1-\lambda_i)||\tilde{h}_i|.$$

- Define $\mathbf{m} \in \mathbb{R}^N$, where $m_i := |\tilde{s}_i||(1-\lambda_i)|, \forall i = 1, \ldots, N$. Let $\boldsymbol{\alpha} := \mathrm{argsort}(-\mathbf{m})$.
- Optimal fairness-aware filter design:

$$\tilde{\mathbf{h}}^{\text{fair}} := \underset{\tilde{\mathbf{h}}}{\mathrm{argmin}}\ \mathbf{m}^\top\tilde{\mathbf{h}}$$
$$\text{s. to } \sum_{i=1}^{N}\tilde{h}_i \geq N\tau,\ 0 \leq \tilde{h}_i \leq 1, \forall i \in \{1, \ldots, N\}.$$

- Closed-form solution: $(\tilde{h}^{\text{fair}})_{\alpha_i} = \left[1 - \left[N(1-\tau) - \sum_{j=1}^{i-1}\left(1 - (\tilde{h}^{\text{fair}})_{\alpha_j}\right)\right]_+\right]_+$.

Fairness-aware Filtering as Graph Rewiring

- Re-formulate fairness-aware filter design as a network inference problem
- Direct optimization of ρ:

$$\mathbf{A}^f := \underset{\bar{\mathbf{A}}}{\mathrm{argmin}}\quad \left\{\|\mathbf{s}^\top\bar{\mathbf{A}}\|_1 + \beta r(\bar{\mathbf{A}}, \hat{\mathbf{A}})\right\},$$
$$\text{s.t.}\quad \bar{\mathbf{A}}\hat{\mathbf{A}} = \hat{\mathbf{A}}\bar{\mathbf{A}},\quad \bar{\mathbf{A}} \in \mathcal{S}.$$

- Utility consideration

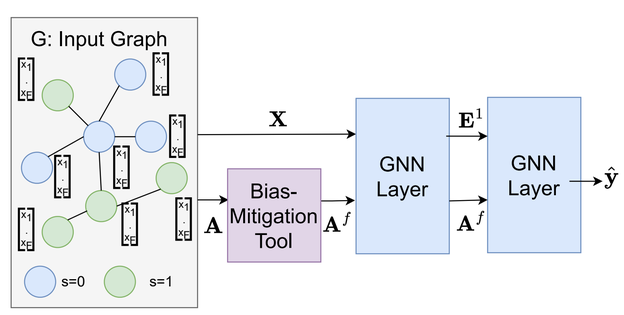
$$r(\mathbf{A}, \mathbf{A}) := \|\mathbf{A} - \mathbf{A}\|_{1,1} \quad (1)$$

- Design of $\mathcal{S}$ provides flexibility

$$\mathcal{S} := \left\{\bar{\mathbf{A}} \mid \bar{A}_{ij} \geq 0, \bar{\mathbf{A}} \in \mathcal{M}^N, \|\bar{\mathbf{A}}\|_{1,1} = \|\hat{\mathbf{A}}\|_{1,1}\right\}.$$

- $\bar{\mathbf{A}}\hat{\mathbf{A}} = \hat{\mathbf{A}}\bar{\mathbf{A}} \rightarrow \hat{\mathbf{A}}$ and $\bar{\mathbf{A}}$ share the same set of eigenvectors
  - Filtering shapes the eigenvalues of the effective graph operator, same eigenvectors
  - Implicitly optimize $\tilde{\mathbf{h}}$ to minimize ρ without requiring a spectral decomposition
- A flexible design that can be **pre-computed once** for different learning algorithms, and can be used at different stages of learning (i.e., pre-processing, post-processing)

Experimental Settings & Results



- Datasets: Real social networks, region is sensitive attribute & job is label
- Task: Node classification, classification accuracy is reported
- Fairness metrics (lower values are desired):
  - $\Delta_{SP} = |P(\hat{y} = 1 \mid s = 0) - P(\hat{y} = 1 \mid s = 1)|$
  - $\Delta_{EO} = |P(\hat{y} = 1 \mid y = 1, s = 0) - P(\hat{y} = 1 \mid y = 1, s = 1)|$
- Comparative results

	NBA			SPokec-n			SPokec-z		
	Accuracy (%)	Δ_{SP} (%)	Δ_{EO} (%)	Accuracy (%)	Δ_{SP} (%)	Δ_{EO} (%)	Accuracy (%)	Δ_{SP} (%)	Δ_{EO} (%)
GNN	61.72 ± 6.2	9.24 ± 5.2	10.66 ± 11.1	76.57 ± 2.4	10.63 ± 6.2	6.64 ± 4.1	72.79 ± 3.8	3.91 ± 4.6	10.77 ± 5.6
Adversarial	61.94 ± 4.0	4.93 ± 2.3	7.20 ± 7.7	76.57 ± 2.7	8.95 ± 5.8	5.11 ± 2.0	70.63 ± 2.8	9.56 ± 5.6	7.96 ± 7.9
EDITS	65.38 ± 2.8	6.49 ± 4.6	10.31 ± 6.4	75.37 ± 3.0	$\mathbf{3.70} \pm 3.9$	$\mathbf{3.10} \pm 1.9$	71.53 ± 4.8	6.04 ± 4.2	7.63 ± 4.7
$\tilde{\mathbf{h}}^{\text{fair}}$+ GNN [1]	60.43 ± 6.8	5.89 ± 4.4	10.74 ± 10.5	75.67 ± 1.7	5.43 ± 2.6	4.91 ± 1.8	72.07 ± 3.4	4.14 ± 4.1	7.23 ± 7.8
$\mathbf{A}^f$	$\mathbf{68.39} \pm 6.9$	$\mathbf{3.28} \pm 1.6$	$\mathbf{6.61} \pm 4.8$	$\mathbf{78.51} \pm 3.0$	6.27 ± 4.1	$\mathbf{3.09} \pm 2.6$	$\mathbf{72.97} \pm 1.1$	$\mathbf{3.04} \pm 4.5$	$\mathbf{2.49} \pm 3.8$

- Similar utility performance compared to fairness-agnostic GNN model
- Better fairness, utility compared to SOTA fairness-aware baselines
- Direct ρ optimization provides better fairness enhancement
- A possible explanation for effective bias mitigation:

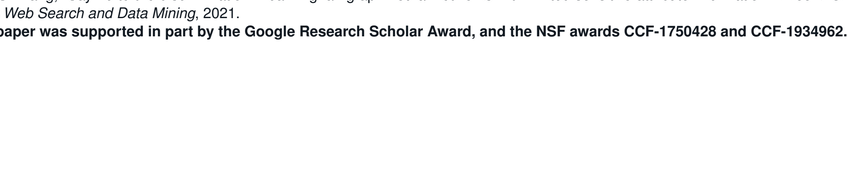


Figure 1: Distribution of intra- (green) and inter-(red)edges in the effective network topology before (left)/ after (right) applying $\tilde{\mathbf{h}}^{\text{fair}}$.

Conclusions

- A novel, unsupervised bias measure
- Two alternative optimal graph filter designs:
  - Theory-based surrogate loss with a closed-form solution, allowing efficient graph filter design
  - Filtering as graph rewiring:Implicit filter design via direct minimization of ρ, without spectral decomposition
- Versatile use and pre-trained computation
- Verified effectiveness on real-world social networks
- Future work: Non-linear relations between $\mathbf{s}$ and graph topology as a bias measure

[1] O. D. Kose, G. Mateos and Y. Shen, "Fairness-aware optimal graph filter design," *IEEE Journal of Selected Topics in Signal Processing*, 2024.
[2] F. Gama et al., "Stability properties of graph neural networks", *IEEE Transactions on Signal Processing*, 2020.
[3] E. Dai and S. Wang, "Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information." *Proc. ACM International Conference on Web Search and Data Mining*, 2021.

Work in this paper was supported in part by the Google Research Scholar Award, and the NSF awards CCF-1750428 and CCF-1934962.