Appendix for "FairGLite: Fair Graph Representation Learning with Limited Demographics"

1 Theoretical analysis

This section provides additional theoretical analysis demonstrating the performance guarantees of FairGLite: Specifically, for a binary classification task, statistical par-ity is defined as: $\Delta_{DP} = |P(\hat{y} = 1|s = 0) - P(\hat{y} = 1|s = 1)|$. We consider the binary classification task and examine the properties of the Softmax function in this context. Let P_1 and P_2 represent the probabilities of class 1 (c_1) and class 2 (c_2), respectively. The function $Softmax(\cdot)$ is Lipschitz continuous with a Lipschitz constant L. Due to this Lipschitz continuity, the difference in output probabilities can be bounded by the difference in input vectors:

$$\begin{aligned} \|f(\mathbf{h_i}) - f(\mathbf{h_j})\| &= |P_1 - P_2| + |(1 - P_1) - (1 - P_2)| \\ &= 2|P_1 - P_2| \leq L\|\mathbf{h_i} - \mathbf{h_j}\| \end{aligned} \tag{1}$$

where $\mathbf{h_i}$ is the node representation for $\forall v_i \in S_d$ and $\mathbf{h_j}$ is the node representation for $\forall v_j \in S_f$.

Building on this, we can rewrite the statistical parity as follows:

$$\Delta_{\text{DP}} = \left| \frac{1}{N_d} \sum_{i \in \mathcal{S}_d} f\big(\mathbf{z}_i\big)_1 \; - \; \frac{1}{N_f} \sum_{j \in \mathcal{S}_f} f\big(\mathbf{z}_j\big)_1 \right| \tag{2}$$

where $\mathbf{z}_i = W^l \mathbf{h}_i^{(l)}$, and $W^{(l)}$ is the weight matrix at layer l.

Hence, the upper bound of disparity of DP depends on the node representation disparity. Given that the Graph Attention Networks (GAT) adopt the message passing by assigning different weights to neighbor nodes. Hence, we can write the node aggregation process as follows:

$$\mathbf{h}_u^{(l)} \;=\; \xi\, \mathbf{h}_u^{(l-1)} + \sum_{k \in \mathcal{N}(u)} \alpha_{u,k}^{(l)}\, \mathrm{ReLU}\Big(\mathbf{W}^{(l)}\, \mathbf{h}_k^{(l-1)}\Big), \quad \alpha_{u,k}^{(l)} \;=\; \frac{\exp\big(e_{u,k}^{(l)}\big)}{\sum_{k \in \mathcal{N}(u)} \exp\big(e_{u,k}^{(l)}\big)} \tag{3}$$

Building on this, the node representation disparity arising from the node aggregation process can be measured. Specifically, for a node v_i, the neighbor information during the aggregation process is:

$$\begin{aligned}\mathbf{h}_i^{(l)} = \xi\mathbf{h}_i^{(l-1)} &+ \sum_{u\in N(i)\cap\mathcal{S}_d}\alpha_{i,u}^{(l)}\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}) + \sum_{u\in N(i)\cap\mathcal{S}_f}\alpha_{i,u}^{(l)}\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}) \\ &- \frac{1}{N_d^2}\sum_{u\in\mathcal{S}_d} k(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}),\sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)}))\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}) \\ &+ \frac{1}{N_dN_f}\sum_{u\in\mathcal{S}_f} k(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}),\sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)}))\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}) \\ &+ \left[\frac{1}{N_d^2}\sum_{u\in\mathcal{S}_d} k(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}),\sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)})) \right. \\ &\left. - \frac{1}{N_dN_f}\sum_{u\in\mathcal{S}_f} k(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}),\sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)}))\right]\sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)})\end{aligned} \tag{4}$$

Let each node v_i have a node representation $h_i^{(l)}$ subject to $\mu^{(d)} - \Delta^l \preceq h_i^{(l)} \preceq \mu^{(d)} + \Delta^l$, where the parameter Δ^l serves as a tolerance per layer indicating the allowed deviation of the representation from group mean ($\mu^{(d)}$) along each coordinate. Hence, we can re-write the Equation 4 as follow:

$$\begin{aligned}\mathbf{h}_i^{(l)} = \xi\mathbf{h}_i^{(l-1)} &+ \sum_{u\in N(i)\cap\mathcal{S}_d}\alpha_{i,u}^{(l)}\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}) + \sum_{u\in N(i)\cap\mathcal{S}_f}\alpha_{i,u}^{(l)}\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}) \\ &- \frac{1}{N_d^2}\sum_{u\in\mathcal{S}_d} k(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}),\sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)}))\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}) \\ &+ \frac{1}{N_dN_f}\sum_{u\in\mathcal{S}_f} k(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}),\sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)}))\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}) \\ &+ \left[\frac{1}{N_d^2}\sum_{u\in\mathcal{S}_d} k(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}),\sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)})) \right. \\ &\left. - \frac{1}{N_dN_f}\sum_{u\in\mathcal{S}_f} k(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}),\sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)}))\right]\sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)}) \\ &\pm C^{(l)}\cdot\frac{1}{4}\left\|W^{(l)}\right\|_2\left(\left(1+\frac{2}{N_d}\right)\sqrt{d_h}\Delta^{l-1} + \Delta_{\text{base}}^{(l)}\right)\end{aligned} \tag{5}$$

Building on this, for nodes $v_i \in S_d$, we have:

$$\frac{1}{N_d}\sum_{i\in\mathcal{S}_d}\mathbf{h}_i^{(l)} \in \Bigg[\mu^{(d)} + \left(\mu_{l-1}^{(d)} + \frac{1}{N_d}\sum_{i\in\mathcal{S}_d}\left(\sum_{u\in\mathcal{N}(i)\cap\mathcal{S}_f}\alpha_{i,u}^{(l)}\right)\left(\mu_{l-1}^{(f)} - \mu_{l-1}^{(d)}\right)\right) \\ + \frac{1}{N_d^2 N_f}\sum_{i\in\mathcal{S}_d}\sum_{u\in\mathcal{S}_f} k\left(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}), \sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)})\right)\left(\mu_{l-1}^{(f)} - \mu_{l-1}^{(d)}\right)\Bigg] \\ \pm \left[C^{(l)}\cdot\frac{1}{4}\left\|W^{(l)}\right\|_2\left(\left(1+\frac{2}{N_d}\right)\sqrt{d_h}\Delta^{l-1} + \Delta_{\text{base}}^{(l)}\right)\right] \tag{6}$$

A similar way can be applied for S_f as follows:

$$\frac{1}{N_f}\sum_{i\in\mathcal{S}_f}\mathbf{h}_i^{(l)} \in \Bigg[\mu^{(f)} + \left(\mu_{l-1}^{(f)} + \frac{1}{N_f}\sum_{i\in\mathcal{S}_f}\left(\sum_{u\in\mathcal{N}(i)\cap\mathcal{S}_d}\alpha_{i,u}^{(l)}\right)\left(\mu_{l-1}^{(d)} - \mu_{l-1}^{(f)}\right)\right) \\ + \frac{1}{N_f^2 N_d}\sum_{i\in\mathcal{S}_f}\sum_{u\in\mathcal{S}_d} k\left(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}), \sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)})\right)\left(\mu_{l-1}^{(d)} - \mu_{l-1}^{(f)}\right)\Bigg] \\ \pm \left[C^{(l)}\cdot\frac{1}{4}\left\|W^{(l)}\right\|_2\left(\left(1+\frac{2}{N_f}\right)\sqrt{d_h}\Delta^{l-1} + \Delta_{\text{base}}^{(l)}\right)\right] \tag{7}$$

Building on this, the upper bound of the consequent representation discrepancy on node representations between two demographic groups is defined as follows:

$$
\begin{aligned}
\mathbf{h}_D^{(l)} &= \left\| \frac{1}{N_d} \sum_{i \in \mathcal{S}_d} \mathbf{h}_i^{(l)} - \frac{1}{N_f} \sum_{j \in \mathcal{S}_f} \mathbf{h}_j^{(l)} \right\| \qquad (8) \\
&\leq \Bigg| 1 - \left(\frac{1}{N_d} \sum_{i \in \mathcal{S}_d} \sum_{u \in \mathcal{N}(i) \cap \mathcal{S}_f} \alpha_{i,u}^{(l)} + \frac{1}{N_f} \sum_{j \in \mathcal{S}_f} \sum_{u \in \mathcal{N}(j) \cap \mathcal{S}_d} \alpha_{j,u}^{(l)} \right) \\
&\quad - \frac{1}{N_d^2 N_f} \sum_{i \in \mathcal{S}_d} \sum_{u \in \mathcal{S}_f} k\left(\sigma(\mathbf{W}^{(l)} \mathbf{h}_u^{(l-1)}), \sigma(\mathbf{W}^{(l)} \mathbf{h}_i^{(l-1)}) \right) \\
&\quad - \frac{1}{N_f^2 N_d} \sum_{j \in \mathcal{S}_f} \sum_{u \in \mathcal{S}_d} k\left(\sigma(\mathbf{W}^{(l)} \mathbf{h}_u^{(l-1)}), \sigma(\mathbf{W}^{(l)} \mathbf{h}_j^{(l-1)}) \right) \Bigg| \left\| \mu_{l-1}^{(d)} - \mu_{l-1}^{(f)} \right\| \\
&\quad + \left\| \mu^{(d)} - \mu^{(f)} \right\| + \varepsilon_d^{(l)} + \varepsilon_f^{(l)} \\
&\leq \Bigg(\frac{1}{N_d} \sum_{i \in \mathcal{S}_d} \sum_{u \in \mathcal{N}(i) \cap \mathcal{S}_f} \alpha_{i,u}^{(l)} + \frac{1}{N_f} \sum_{j \in \mathcal{S}_f} \sum_{u \in \mathcal{N}(j) \cap \mathcal{S}_d} \alpha_{j,u}^{(l)} + 1 \\
&\quad - \frac{1}{N_d^2 N_f} \sum_{i \in \mathcal{S}_d} \sum_{u \in \mathcal{S}_f} k\left(\sigma(\mathbf{W}^{(l)} \mathbf{h}_u^{(l-1)}), \sigma(\mathbf{W}^{(l)} \mathbf{h}_i^{(l-1)}) \right) \\
&\quad - \frac{1}{N_f^2 N_d} \sum_{j \in \mathcal{S}_f} \sum_{u \in \mathcal{S}_d} k\left(\sigma(\mathbf{W}^{(l)} \mathbf{h}_u^{(l-1)}), \sigma(\mathbf{W}^{(l)} \mathbf{h}_j^{(l-1)}) \right) \Bigg) \left\| \mu_{l-1}^{(d)} - \mu_{l-1}^{(f)} \right\| \\
&\quad + \left\| \mu^{(d)} - \mu^{(f)} \right\| + \varepsilon_d^{(l)} + \varepsilon_f^{(l)} \\
&\leq \Bigg(3 - \frac{1}{N_d^2 N_f} \sum_{i \in \mathcal{S}_d} \sum_{u \in \mathcal{S}_f} k\left(\sigma(\mathbf{W}^{(l)} \mathbf{h}_u^{(l-1)}), \sigma(\mathbf{W}^{(l)} \mathbf{h}_i^{(l-1)}) \right) \qquad (9) \\
&\quad - \frac{1}{N_f^2 N_d} \sum_{j \in \mathcal{S}_f} \sum_{u \in \mathcal{S}_d} k\left(\sigma(\mathbf{W}^{(l)} \mathbf{h}_u^{(l-1)}), \sigma(\mathbf{W}^{(l)} \mathbf{h}_j^{(l-1)}) \right) \Bigg) \left\| \mu_{l-1}^{(d)} - \mu_{l-1}^{(f)} \right\| \\
&\quad + \left\| \mu^{(d)} - \mu^{(f)} \right\| \pm \left[C^{(l)} \cdot \frac{1}{4} \left\| W^{(l)} \right\|_2 \left(\left(1 + \frac{2}{N_f} \right) \sqrt{d_h} \Delta^{l-1} + \Delta_{\text{base}}^{(l)} \right) \right]
\end{aligned}
$$

Armed with this node representation disparity, the effect on statistical parity can then be measured. Given that:

$$
\Delta_{\text{DP}} = \left| \frac{1}{N_d} \sum_{i \in \mathcal{S}_d} f\big(\mathbf{z}_i\big)_1 \; - \; \frac{1}{N_f} \sum_{j \in \mathcal{S}_f} f\big(\mathbf{z}_j\big)_1 \right| \qquad (10)
$$

As discussed previously, given that $\|f(\mathbf{h_i}) - f(\mathbf{h_j})\| = 2|P_1 - P_2| \leq L\|\mathbf{h_i} - \mathbf{h_j}\|$. Hence, we can rewrite it as:

$$
2 \left| f(\mathbf{z}_i)_1 - f(\mathbf{z}_{\mu^{(d)}})_1 \right| \;\; \leq \;\; L \left\| \mathbf{z}_i - \mathbf{z}_{\mu^{(d)}} \right\| \qquad (11)
$$

Therefore, the following inequality holds:

$$f\big(\mathbf{z}_{\mu^{(d)}}\big)_1 \;-\; \tfrac{L}{2}\,\|\mathbf{z}_i - \mathbf{z}_{\mu^{(d)}}\| \;\;\leq\;\; f\big(\mathbf{z}_i\big)_1 \;\;\leq\;\; f\big(\mathbf{z}_{\mu^{(d)}}\big)_1 \;+\; \tfrac{L}{2}\,\|\mathbf{z}_i - \mathbf{z}_{\mu^{(d)}}\| \tag{12}$$

Let $\mathbf{z}_i = \mathbf{W}^{(l)}\mathbf{h}_i^{(l)}$ for node i, and $\mathbf{z}_{\mu^{(d)}} = \mathbf{W}^{(l)}\mu_l^{(d)}$, $\mathbf{z}_{\mu^{(f)}} = \mathbf{W}^{(l)}\mu_l^{(f)}$ be the group means in logits space.

$$f(\mathbf{z}_{\mu^{(d)}})_1 - f(\mathbf{z}_{\mu^{(f)}})_1 - \tfrac{1}{N_d}\sum_{i=1}^{N_d}\tfrac{L}{2}\|\mathbf{z}_i - \mathbf{z}_{\mu^{(d)}}\| - \tfrac{1}{N_f}\sum_{j=1}^{N_f}\tfrac{L}{2}\|\mathbf{z}_j - \mathbf{z}_{\mu^{(f)}}\| \tag{13}$$

$$\leq \tfrac{1}{N_d}\sum_{i\in\mathcal{S}_d} f(\mathbf{z}_i)_1 - \tfrac{1}{N_f}\sum_{j\in\mathcal{S}_f} f(\mathbf{z}_j)_1 \leq f(\mathbf{z}_{\mu^{(d)}})_1 - f(\mathbf{z}_{\mu^{(f)}})_1 + \tfrac{1}{N_d}\sum_{i=1}^{N_d}\tfrac{L}{2}\|\mathbf{z}_i - \mathbf{z}_{\mu^{(d)}}\| + \tfrac{1}{N_f}\sum_{j=1}^{N_f}\tfrac{L}{2}\|\mathbf{z}_j - \mathbf{z}_{\mu^{(f)}}\|$$

Considering the obtained h_D, the following inequality holds:

$$\begin{aligned}
\big\|\mathbf{z}_i - \mathbf{z}_{\mu^{(d)}}\big\| &= \Big\|\mathbf{W}^{(l)}\left(\mathbf{h}_i^{(l)} - \mu_l^{(d)}\right)\Big\| \leq \Big\|\mathbf{W}^{(l)}\Big\|_2 \Big\|\mathbf{h}_i^{(l)} - \mu_l^{(d)}\Big\| \qquad (14)\\
&\leq \Big\|\mathbf{W}^{(l)}\Big\|_2 \Bigg[\sqrt{d_h}\Delta^l + C^{(l)}\Bigg(\Bigg(3 - \frac{1}{N_d^2 N_f}\sum_{p\in\mathcal{S}_d}\sum_{q\in\mathcal{S}_f} k\left(\sigma(\mathbf{W}^{(l)}\mathbf{h}_q^{(l-1)}), \sigma(\mathbf{W}^{(l)}\mathbf{h}_p^{(l-1)})\right)\\
&\quad - \frac{1}{N_f^2 N_d}\sum_{j\in\mathcal{S}_f}\sum_{u\in\mathcal{S}_d} k\left(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}), \sigma(\mathbf{W}^{(l)}\mathbf{h}_j^{(l-1)})\right)\Bigg)\Big\|\mu_{l-1}^{(d)} - \mu_{l-1}^{(f)}\Big\|\\
&\quad + \Big\|\mu^{(d)} - \mu^{(f)}\Big\| + \Bigg[C^{(l)}\cdot\frac{1}{4}\Big\|W^{(l)}\Big\|_2\left(\left(1 + \frac{2}{N_f}\right)\sqrt{d_h}\Delta^{l-1} + \Delta_{\text{base}}^{(l)}\right)\Bigg]\Bigg)\Bigg]
\end{aligned}$$

A similar expression can be derived for $|\mathbf{z}i - \mathbf{z}\mu^{(f)}|$.

Hence, we can write the upper bound of the DP as:

$$\Delta_{\text{DP}} \leq \big|f(\mathbf{z}_{\mu^{(d)}})_1 - f(\mathbf{z}_{\mu^{(f)}})_1\big| + \frac{L}{2}\left(\frac{1}{N_d}\sum_i \big\|\mathbf{z}_i - \mathbf{z}_{\mu^{(d)}}\big\| + \frac{1}{N_f}\sum_j \big\|\mathbf{z}_j - \mathbf{z}_{\mu^{(f)}}\big\|\right) \tag{15}$$

$$\begin{aligned}
&\leq \frac{L}{2}\Big\|\mathbf{W}^{(l)}\Big\|_2\Bigg[\Bigg(3 - \frac{1}{N_d^2 N_f}\sum_{i\in\mathcal{S}_d}\sum_{u\in\mathcal{S}_f} k\left(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}), \sigma(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)})\right)\\
&\quad - \frac{1}{N_f^2 N_d}\sum_{j\in\mathcal{S}_f}\sum_{u\in\mathcal{S}_d} k\left(\sigma(\mathbf{W}^{(l)}\mathbf{h}_u^{(l-1)}), \sigma(\mathbf{W}^{(l)}\mathbf{h}_j^{(l-1)})\right)\Bigg)\Big\|\mu_{l-1}^{(d)} - \mu_{l-1}^{(f)}\Big\|\\
&\quad + \Big\|\mu^{(d)} - \mu^{(f)}\Big\| + \varepsilon_d^{(l)} + \varepsilon_f^{(l)} + 2\sqrt{d_h}\Delta^l\Bigg]
\end{aligned}$$

This completes the proof.

2 Additional Experimental Results

We further investigated the sensitivity of FairGLite to hyperparameters a and b on the NBA and Pokec-n datasets, as shown in Figure 1. Similar to the observations in Credit and Pokec-z datasets, increasing a consistently enhanced both fairness and prediction performance until reaching a threshold, beyond which improvements stabilized. For hyperparameter b, the results demonstrated three distinct phases: initially, at low values, fairness constraints had negligible effects, reflecting minimal regularization. As b increased to moderate levels, there was a clear improvement in fairness metrics, accompanied by a gradual decline in predictive accuracy, reflecting a stronger regularization impact. Eventually, beyond a certain threshold (approximately e^1 for NBA and e^3 for Pokec-n), further increasing b resulted in stabilized or slightly deteriorated fairness performance.

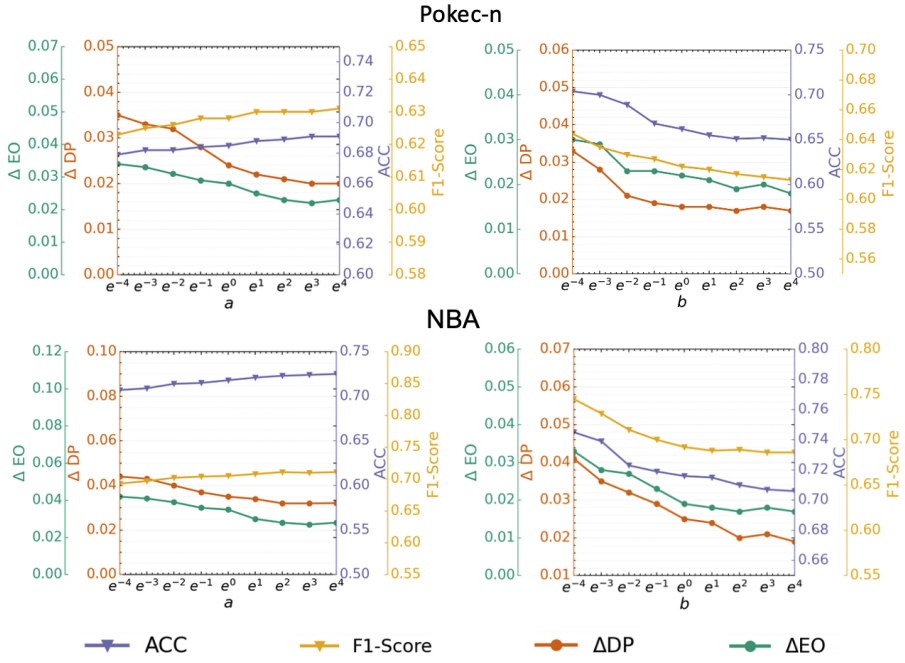

Fig. 1: Study on Hyper-parameters sensitivity analysis.

We further conducted ablation studies to assess the contributions of each FairGLite module on the NBA and Pokec-n datasets, as shown in Figure 2. Consistent with the Credit and Pokec-z datasets observations, removing the Fairness Constraint (FairGLite-NF) reduced model fairness due to biases propagating directly into predictions. The variant without the Graph Reconstruction Constraint (FairGLite-NG) exhibited a less severe drop in fairness but demonstrated reduced predictive accuracy, underscoring the

importance of structural information. Lastly, removing the Adaptivity Confidence Strategy Module (FairGLite-NA) led to decreased overall performance. This reduction illustrates the module's critical function in dynamically weighting fairness constraints based on prediction confidence, effectively balancing predictive accuracy and fairness enforcement.

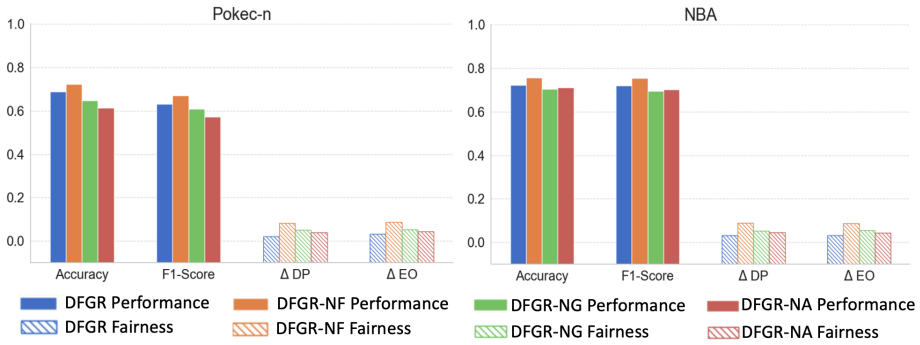

Fig. 2: Ablation study results for FairGLite, FairGLite-NF, FairGLite-NG, and FairGLite-NA.