A Proofs

A.1 Proof of Lemma 1 (special case of total variation distance)

We assume $\mathbb{P}(Z)$ and $\mathbb{P}(X)$ are continuous probability distributions, but this proof can be easily altered for other kinds of probability distributions and the result still holds. Let $\mathbb{P}(S = Q) = p$. By definition, the total variation information:

$$\mathrm{I}_{TV}(\mathbb{P}(Z);\mathbb{P}(S)) = d_{TV}(\mathbb{P}(Z,S)||\mathbb{P}(Z)\otimes\mathbb{P}(S)) = \textstyle\sum_{s\in\{Q,R\}}\mathbb{P}(S=s)\int_{\mathcal{Y}}\frac{1}{2}\left|\frac{f_{Z,S}(z,s)}{f_Z(z)\mathbb{P}(S=s)}-1\right| f_Z(z)\,\mathrm{d}z$$

By breaking up joint probabilities into conditional probabilities and factoring:

$$\begin{aligned}
I_{TV}(\mathbb{P}(Z);\mathbb{P}(S)) &= \sum_{s\in\{Q,R\}}\mathbb{P}(S=s)\int_{\mathcal{Y}}\frac{1}{2}\left|\frac{f_{Z|S=s}(z)\mathbb{P}(S=s)}{f_Z(z)\mathbb{P}(S=s)}-1\right| f_Z(z)\,\mathrm{d}z\\
&= \sum_{s\in\{Q,R\}}\mathbb{P}(S=s)\int_{\mathcal{Y}}\frac{1}{2}\left|f_{Z|S=s}(z)-f_Z(z)\right|\mathrm{d}z\\
&= \frac{p}{2}\int_{\mathcal{Y}}\left|f_{Z|S=Q}(z)-f_Z(z)\right|\mathrm{d}z+\int_{\mathcal{Y}}\frac{1-p}{2}\left|f_{Z|S=R}(z)-f_Z(z)\right|\mathrm{d}z\\
&= \frac{1}{2}\Big[p\int_{\mathcal{Y}}\left|f_{Z|S=Q}(z)-(pf_{Z|S=Q}(z)+(1-p)f_{Z|S=R}(z))\right|\mathrm{d}z\\
&\quad+(1-p)\int_{\mathcal{Y}}\left|f_{Z|S=R}(z)-(pf_{Z|S=Q}(z)+(1-p)f_{Z|S=R}(z))\right|\mathrm{d}z\Big]\\
&= \frac{1}{2}\Big[p\int_{\mathcal{Y}}\left|(1-p)f_{Z|S=Q}(z)-(1-p)f_{Z|S=R}(z)\right|\mathrm{d}z\\
&\quad+(1-p)\int_{\mathcal{Y}}\left|pf_{Z|S=R}(z)-pf_{Z|S=Q}(z)\right|\mathrm{d}z\Big]\\
&= \frac{p(1-p)}{2}\left[\int_{\mathcal{Y}}\left|f_{Z|S=Q}(z)-f_{Z|S=R}(z)\right|\mathrm{d}z+\int_{\mathcal{Y}}\left|f_{Z|S=R}(z)-f_{Z|S=Q}(z)\right|\mathrm{d}z\right]\\
&= p(1-p)\int_{\mathcal{Y}}\left|f_{Z|S=Q}(z)-f_{Z|S=R}(z)\right|\mathrm{d}z\\
&= 2p(1-p)d_{TV}(\mathbb{P}(Z|S=Q),\mathbb{P}(Z|S=R))
\end{aligned}$$

It can be similarly shown that $I_{TV}(\mathbb{P}(X);\mathbb{P}(S)) = 2p(1-p)d_{TV}(\mathbb{P}(X|S=Q),\mathbb{P}(X|S=R))$.

Now, because Z is conditionally independent of S given X, by the Data Processing Inequality, $I_{TV}(\mathbb{P}(Z);\mathbb{P}(S)) \le I_{TV}(\mathbb{P}(X);\mathbb{P}(S))$. Hence:

$$\begin{aligned}
d_{TV}(\mathbb{P}(Z|S=Q),\mathbb{P}(Z|S=R)) &= \frac{1}{2p(1-p)}I_{TV}(\mathbb{P}(Z);\mathbb{P}(S))\\
&\le \frac{1}{2p(1-p)}I_{TV}(\mathbb{P}(X);\mathbb{P}(S)) = d_{TV}(\mathbb{P}(X|S=Q),\mathbb{P}(X|S=R))
\end{aligned}$$

Similarly:

$$d_{TV}(\mathbb{P}(Z'|S=Q),\mathbb{P}(Z'|S=R)) \le d_{TV}(\mathbb{P}(X'|S=Q),\mathbb{P}(X'|S=R))$$

Note: Differences in the supports of $\mathbb{P}(Z)$ and $\mathbb{P}(X)$ should not influence one's interpretation of the inequality. $d_{TV}(\cdot,\cdot)$ only requires that its two arguments have the same support. Because d_{TV} outputs the largest possible difference between the probabilities that the two distributions can assign to the same event, the inequality can be viewed as a comparison of the differences in assigned probabilities.

A.2 Proof of Theorem 1

Leveraging the Bretagnolle–Huber (BH) bound[4], we can upper bound d_{TV} in terms of the KL-divergence d_{KL}:

[4]We use the BH bound rather than Pinsker's inequality because Pinkser's inequality becomes vacuous for KL-divergence > 2 [91].

$$d_{TV}(\mathbb{P}(X|S=Q),\mathbb{P}(X|S=R)) \leq \sqrt{1-e^{-d_{KL}(\mathbb{P}(X|S=Q)||\mathbb{P}(X|S=R))}}$$

By Section 9 from [92], $d_{KL}(\mathbb{P}(X|S=Q)||\mathbb{P}(X|S=R))$ admits a closed-form solution:

$$\begin{aligned} d_{KL}(\mathbb{P}(X|S=Q)||\mathbb{P}(X|S=R)) &= \frac{1}{2}\left(\log\frac{\det\Sigma_R}{\det\Sigma_Q} - d + tr(\Sigma_R^{-1}\Sigma_Q) + \|\mu_Q-\mu_R\|^2_{\Sigma_R^{-1}}\right) \\ &\leq \frac{1}{2}\left(\log\frac{\det\Sigma_R}{\det\Sigma_Q} - d + tr(\Sigma_R^{-1}\Sigma_Q) + \lambda_{max}(\Sigma_R^{-1})\|\mu_Q-\mu_R\|_2^2\right), \end{aligned}$$

where $\lambda_{max}(\Sigma_R^{-1})$ is the maximum eigenvalue of Σ_R^{-1}. We note that $\lambda_{max}(\Sigma_R^{-1}) = \frac{1}{\lambda_{min}(\Sigma_R)} > 0$ (where $\lambda_{min}(\Sigma_R)$ is the minimum eigenvalue of Σ_R) because Σ_R is positive semidefinite.

It is clear that $\|\mu_Q-\mu_R\|_\infty^2 = \max_{i\in[d]}|(\mu_Q)_i-(\mu_R)_i|^2 \leq \sum_{i\in[d]}|(\mu_Q)_i-(\mu_R)_i|^2 = \|\mu_Q-\mu_R\|_2^2$. Moreover, $\|\mu_Q-\mu_R\|_2^2 = \sum_{i\in[d]}|(\mu_Q)_i-(\mu_R)_i|^2 \leq \sum_{i\in[d]}\max_{j\in[d]}|(\mu_Q)_j-(\mu_R)_j|^2 = d\cdot\|\mu_Q-\mu_R\|_\infty^2$. Therefore, $\|\mu_Q-\mu_R\|_2^2 = C\cdot\|\mu_Q-\mu_R\|_\infty^2$, for $1\leq C\leq d$.

Combining the previous observations and recognizing that $\|\mu_Q-\mu_R\|_\infty^2 = \mathcal{R}_\mathcal{D}^2$:

$$d_{TV}(\mathbb{P}(X|S=Q),\mathbb{P}(X|S=R)) \leq \sqrt{1-\sqrt{\frac{\det\Sigma_Q}{\det\Sigma_R}\cdot e^{-\frac{C\cdot\mathcal{R}_\mathcal{D}^2}{\lambda_{min}(\Sigma_R)}-tr(\Sigma_R^{-1}\Sigma_Q)+d}}}$$

Now, by Lemma 2.7 from [93]:

$$\begin{aligned} \|\mu_Q-\mu_R\|_2^2 &\leq 4\cdot\max\{\lambda_{max}(\Sigma_Q),\lambda_{max}(\Sigma_R)\}\left(\frac{d_{TV}(\mathbb{P}(X|S=Q),\mathbb{P}(X|S=R))}{1-d_{TV}(\mathbb{P}(X|S=Q),\mathbb{P}(X|S=R))}\right) \\ &\leq \frac{4\cdot\max\{\lambda_{max}(\Sigma_Q),\lambda_{max}(\Sigma_R)\}}{\frac{1}{d_{TV}(\mathbb{P}(X|S=Q),\mathbb{P}(X|S=R))}-1} \end{aligned}$$

Using $\|\mu_Q-\mu_R\|_2^2 = C\cdot\|\mu_Q-\mu_R\|_\infty^2 = C\cdot\mathcal{R}_\mathcal{D}^2$, we can derive:

$$\frac{1}{\frac{4\cdot\max\{\lambda_{max}(\Sigma_Q),\lambda_{max}(\Sigma_R)\}}{C\cdot\mathcal{R}_\mathcal{D}^2}+1} \leq d_{TV}(\mathbb{P}(X|S=Q),\mathbb{P}(X|S=R))$$

Similarly:

$$\mathrm{d}_{TV}(\mathbb{P}(X'|S=Q),\mathbb{P}(X'|S=R)) \in \left[\frac{1}{\frac{4\cdot\max\{\lambda_{max}(\Sigma'_Q),\lambda_{max}(\Sigma'_R)\}}{C'\cdot\mathcal{R}_{\mathcal{D}'}^2}+1}, \sqrt{1-\sqrt{\frac{\det\Sigma'_Q}{\det\Sigma'_R}\cdot e^{-\frac{C'\cdot\mathcal{R}_{\mathcal{D}'}^2}{\lambda_{min}(\Sigma'R)}-tr(\Sigma'_R{}^{-1}\Sigma'_Q)+d}}}\right]$$

Then, the theorem is proved by application of Lemma 1.

A.3 Example mean aggregation imputation algorithms

Global Mean This method sets the unknown features to the uniform mean of all the known features. To achieve this, we can choose $M := I_N$, $T := \begin{bmatrix} I_{|K|} & 0 \\ \frac{1}{|K|}\mathbb{1}_{|U|\times|K|} & 0 \end{bmatrix}$ (where $\mathbb{1}$ is the all-ones matrix), $\beta := 0$, $X_K^{(0)} := X_K$, and $X_U^{(0)} := 0$. We only need to complete one iteration.

Neighbor Mean This method sets the unknown features to the degree-weighted mean of the known features for neighboring nodes. We can choose $M := I_N$, $T := D^{-1}A$, $\beta := 0$, $X_K^{(0)} := X_K$, and $X_U^{(0)} := 0$. We only need to complete one iteration.

Feature Propagation This method proposed by [16] predicts the unknown features to minimize the Dirichlet energy of the graph while preserving the known feature values. [16] shows that this is equivalent to iteratively computing until convergence:

$$X_K^{(t+1)} := X_K^{(t)}$$

$$X_U^{(t+1)} := (D_U^{-\frac{1}{2}} A_{UK} D_K^{-\frac{1}{2}}) X_K^{(t)} + (D_U^{-\frac{1}{2}} A_{UU} D_U^{-\frac{1}{2}}) X_U^{(t)}$$

Multiplying both sides by $D_U^{-\frac{1}{2}}$, we can re-express the second update rule as:

$$D_U^{-\frac{1}{2}} X_U^{(t+1)} = (D_U^{-1} A_{UK})(D_K^{-\frac{1}{2}} X_K^{(t)}) + (D_U^{-1} A_{UU})(D_U^{-\frac{1}{2}} X_U^{(t)})$$

Therefore, to achieve Feature Propagation, we can choose $M := D^{-\frac{1}{2}}$, $T := D^{-1}A$, $\beta := 0$, and $X_K^{(0)} := X_K$. Per [16], we can choose $X_U^{(0)}$ arbitrarily, and we need to iterate till convergence.
Graph Regularization This method inspired by [94] predicts the unknown features via a smoothness constraint and a fitting constraint for the known features. [94] shows that this is equivalent to iteratively computing until convergence:

$$X_K^{(t+1)} := \beta (D^{-\frac{1}{2}} A D^{-\frac{1}{2}} X^{(t)})_K + (1-\beta) X_K$$

$$X_U^{(t+1)} := (D^{-\frac{1}{2}} A D^{-\frac{1}{2}} X^{(t)})_U,$$

where the hyperparameter $\beta \in (0, 1]$. Therefore, similar to Feature Propagation, to achieve Graph Regularization, we can choose $M := D^{-\frac{1}{2}}$, $T := D^{-1}A$, and $X_K^{(0)} := X_K$. Per [94], we can choose $X_U^{(0)}$ arbitrarily, and we need to iterate till convergence.

A.4 Proof of Theorem 2

The following proof is partially inspired by the proof of Theorem 4.1 in [7]. Fix t to be an arbitrary iteration of feature imputation. Recall we use the following iterative update rule to impute features:

$$\tilde{X}^{(t+1)} := \begin{bmatrix} \beta I_{|K|} & 0 \\ 0 & I_{|U|} \end{bmatrix} T \tilde{X}^{(t)} + \begin{bmatrix} (1-\beta) I_{|K|} & 0 \\ 0 & 0 \end{bmatrix} \tilde{X}$$

For a node $q \in Q \cap U$, after one iteration of feature imputation:

$$\tilde{X}_q^{(t+1)} := \sum_{s \in Q} T_{qs} \tilde{X}_s^{(t)} + \sum_{s \in R} T_{qs} \tilde{X}_s^{(t)}$$

Similarly, for a node $r \in R \cap U$, after one iteration of feature imputation:

$$\tilde{X}_r^{(t+1)} := \sum_{s \in Q} T_{rs} \tilde{X}_s^{(t)} + \sum_{s \in R} T_{rs} \tilde{X}_s^{(t)}$$

In contrast, for a node $q \in Q \cap K$, after one iteration of feature imputation:

$$\tilde{X}_q^{(t+1)} := \beta \left(\sum_{s \in Q} T_{qs} \tilde{X}_s^{(t)} + \sum_{s \in R} T_{qs} \tilde{X}_s^{(t)} \right) + (1-\beta) \tilde{X}_q^{(t)}$$

Similarly, for a node $r \in R \cap K$, after one iteration of feature imputation:

$$\tilde{X}_r^{(t+1)} := \beta \left(\sum_{s \in Q} T_{rs} \tilde{X}_s^{(t)} + \sum_{s \in R} T_{rs} \tilde{X}_s^{(t)} \right) + (1-\beta) \tilde{X}_r^{(t)}$$

We say $v \in [\mu \pm \sigma] \iff \mu - \sigma \preceq v \preceq \mu + \sigma$. Then, for a node $q \in Q \cap U$, by the right-stochastic nature of T:

$$\begin{aligned} \tilde{X}_q^{(t+1)} &\in \left[\left(\sum_{s \in Q} T_{qs} \tilde{\mu}_Q^{(t)} + \sum_{s \in R} T_{qs} \tilde{\mu}_R^{(t)} \right) \pm \tilde{\sigma}^{(t)} \right] \\ &\in \left[\left(\tilde{\mu}_Q^{(t)} + \sum_{s \in R} T_{qs} \left(\tilde{\mu}_R^{(t)} - \tilde{\mu}_Q^{(t)} \right) \right) \pm \tilde{\sigma}^{(t)} \right] \end{aligned}$$

Similarly, for a node $r \in R \cap U$:

$$\tilde{X}_r^{(t+1)} \in \left[\left(\sum_{s \in Q} T_{rs} \tilde{\mu}_Q^{(t)} + \sum_{s \in R} T_{rs} \tilde{\mu}_R^{(t)} \right) \pm \tilde{\sigma}^{(t)} \right]$$
$$\in \left[\left(\tilde{\mu}_R^{(t)} + \sum_{s \in Q} T_{rs} \left(\tilde{\mu}_Q^{(t)} - \tilde{\mu}_R^{(t)} \right) \right) \pm \tilde{\sigma}^{(t)} \right]$$

In contrast, for a node $q \in Q \cap K$:

$$\tilde{X}_q^{(t+1)} \in \left[\left(\beta \left(\sum_{s \in Q} T_{qs} \tilde{\mu}_Q^{(t)} + \sum_{s \in R} T_{qs} \tilde{\mu}_R^{(t)} \right) + (1-\beta) \tilde{\mu}_Q^{(t)} \right) \pm \tilde{\sigma}^{(t)} \right]$$
$$\in \left[\left(\tilde{\mu}_Q^{(t)} + \beta \sum_{s \in R} T_{qs} \left(\tilde{\mu}_R^{(t)} - \tilde{\mu}_Q^{(t)} \right) \right) \pm \tilde{\sigma}^{(t)} \right]$$

Similarly, for a node $r \in R \cap K$:

$$\tilde{X}_r^{(t+1)} \in \left[\left(\beta \left(\sum_{s \in Q} T_{rs} \tilde{\mu}_Q^{(t)} + \sum_{s \in R} T_{rs} \tilde{\mu}_R^{(t)} \right) + (1-\beta) \tilde{\mu}_R^{(t)} \right) \pm \tilde{\sigma}^{(t)} \right]$$
$$\in \left[\left(\tilde{\mu}_R^{(t)} + \beta \sum_{s \in Q} T_{rs} \left(\tilde{\mu}_Q^{(t)} - \tilde{\mu}_R^{(t)} \right) \right) \pm \tilde{\sigma}^{(t)} \right]$$

By the Law of Total Expectation:

$$\mathbb{E}_{q \sim Q}[\tilde{X}_q^{(t+1)}] = \mathbb{P}(q \in U | q \in Q) \mathbb{E}_{q \sim Q \cap U}[\tilde{X}_q^{(t+1)}] + \mathbb{P}(q \in K | q \in Q) \mathbb{E}_{q \sim Q \cap K}[\tilde{X}_q^{(t+1)}]$$
$$\in \Big[\left(\frac{1}{|Q|} \Big(\sum_{q \in Q \cap U} \tilde{\mu}_Q^{(t)} + \sum_{s \in R} T_{qs} \left(\tilde{\mu}_R^{(t)} - \tilde{\mu}_Q^{(t)} \right) \right)$$
$$+ \frac{1}{|Q|} \left(\sum_{q \in Q \cap K} \tilde{\mu}_Q^{(t)} + \beta \sum_{s \in R} T_{qs} \left(\tilde{\mu}_R^{(t)} - \tilde{\mu}_Q^{(t)} \right) \right) \Big) \pm \tilde{\sigma}^{(t)} \Big]$$
$$\in \Big[\Big(\tilde{\mu}_Q^{(t)} + \frac{1}{|Q|} \sum_{q \in Q \cap U} \sum_{s \in R} T_{qs} \left(\tilde{\mu}_R^{(t)} - \tilde{\mu}_Q^{(t)} \right)$$
$$+ \frac{\beta}{|Q|} \sum_{q \in Q \cap K} \sum_{s \in R} T_{qs} \left(\tilde{\mu}_R^{(t)} - \tilde{\mu}_Q^{(t)} \right) \Big) \pm \tilde{\sigma}^{(t)} \Big]$$

Similarly, $\mathbb{E}_{r \sim R}[\tilde{X}_r^{(t+1)}]$:

$$\in \left[\left(\tilde{\mu}_R^{(t)} + \frac{1}{|R|} \sum_{r \in R \cap U} \sum_{s \in Q} T_{rs} \left(\tilde{\mu}_Q^{(t)} - \tilde{\mu}_R^{(t)} \right) + \frac{\beta}{|R|} \sum_{r \in R \cap K} \sum_{s \in Q} T_{rs} \left(\tilde{\mu}_Q^{(t)} - \tilde{\mu}_R^{(t)} \right) \right) \pm \tilde{\sigma}^{(t)} \right]$$

Thus, the gap in expectation of the features of the nodes in Q and R after one iteration of feature imputation is:

$$\mathbb{E}_{q \sim Q}[\tilde{X}_q^{(t+1)}] - \mathbb{E}_{r \sim R}[\tilde{X}_r^{(t+1)}] \in \left[\left(1 - \left(\frac{1}{|Q|} \sum_{q \in Q \cap U} \sum_{s \in R} T_{qs} + \frac{1}{|R|} \sum_{r \in R \cap U} \sum_{s \in Q} T_{qs} \right) \right.$$
$$\left. - \ \beta \left(\frac{1}{|Q|} \sum_{q \in Q \cap K} \sum_{s \in R} T_{qs} + \frac{1}{|R|} \sum_{r \in R \cap K} \sum_{s \in Q} T_{qs} \right) \right) \cdot \left(\tilde{\mu}_Q^{(t)} - \tilde{\mu}_R^{(t)} \right) \Bigg) \pm 2\tilde{\sigma}^{(t)} \Bigg]$$

Define the contraction coefficient:

$$\alpha := \left|1 - \frac{T_{R\to Q\cap U} + \beta T_{R\to Q\cap K}}{|Q|} - \frac{T_{Q\to R\cap U} + \beta T_{Q\to R\cap K}}{|R|}\right|$$

Because $0 \leq \frac{T_{R\to Q\cap U}+\beta T_{R\to Q\cap K}}{|Q|} \leq \frac{T_{R\to Q\cap U}+T_{R\to Q\cap K}}{|Q|} = \frac{T_{R\to Q}}{|Q|} < \frac{T_{V\to Q}}{|Q|} = 1$, and similarly, $0 \leq \frac{T_{Q\to R\cap U}+\beta T_{Q\to R\cap K}}{|R|} < 1$, it must be that $0 \leq \alpha \leq 1$.

Then:

$$\max\{\alpha|\mathbb{E}_{q\sim Q}[\tilde{X}_q^{(t)}] - \mathbb{E}_{r\sim R}[\tilde{X}_r^{(t)}]| - 2\tilde{\sigma}^{(t)}, 0\} \leq |\mathbb{E}_{q\sim Q}[\tilde{X}_q^{(t+1)}] - \mathbb{E}_{r\sim R}[\tilde{X}_r^{(t+1)}]|$$
$$|\mathbb{E}_{q\sim Q}[\tilde{X}_q^{(t+1)}] - \mathbb{E}_{r\sim R}[\tilde{X}_r^{(t+1)}]| \leq \alpha|\mathbb{E}_{q\sim Q}[\tilde{X}_q^{(t)}] - \mathbb{E}_{r\sim R}[\tilde{X}_r^{(t)}]| + 2\tilde{\sigma}^{(t)}$$
$$\max\{\alpha\tilde{\mathcal{R}}^{(t)} - 2\tilde{\sigma}^{(t)}, 0\} \leq \tilde{\mathcal{R}}^{(t+1)} \leq \alpha\tilde{\mathcal{R}}^{(t)} + 2\tilde{\sigma}^{(t)}$$

Inductively, the discrimination risk $\tilde{\mathcal{R}}^{(t)}$ after t iterations of feature imputation is bounded by:

$$\max\left\{\alpha^t\tilde{\mathcal{R}}^{(0)} - 2\left(\sum_{j=0}^{t-1}\alpha^j\tilde{\sigma}^{(j)}\right), 0\right\} \leq \tilde{\mathcal{R}}^{(t)} \leq \alpha^t\tilde{\mathcal{R}}^{(0)} + 2\left(\sum_{j=0}^{t-1}\alpha^j\tilde{\sigma}^{(j)}\right)$$

$\forall v \in V$, $\tilde{X}_v^{(t+1)}$ is a convex combination of $\bigcup_{u\in V}\{\tilde{X}_u^{(t)}\}$. This is because each row of T and $\beta T + (1-\beta)I_{|K|}$ contains nonnegative entries that sum to 1. Therefore, $\tilde{X}_v^{(t+1)}$ must be in the (closed) convex hull of $\bigcup_{u\in V}\{\tilde{X}_u^{(t)}\}$. Thus, $\bigcup_{u\in V}\{\tilde{X}_u^{(t)}\}$ inductively must be contained within the (closed) convex hull of $\bigcup_{u\in V}\{\tilde{X}_u^{(0)}\}$, which has extreme points $\subseteq \bigcup_{u\in V}\{\tilde{X}_u^{(0)}\}$. Consequently, $\forall t \in [0,\infty), \tilde{\sigma}^{(t)} \leq \tilde{\sigma}^{(0)}$.

Hence:

$$\max\left\{\alpha^t\tilde{\mathcal{R}}^{(0)} - 2\left(\sum_{j=0}^{t-1}\alpha^j\right)\tilde{\sigma}^{(0)}, 0\right\} \leq \tilde{\mathcal{R}}^{(t)} \leq \alpha^t\tilde{\mathcal{R}}^{(0)} + 2\left(\sum_{j=0}^{t-1}\alpha^j\right)\tilde{\sigma}^{(0)}$$

If $\alpha < 1$:

$$\max\left\{\alpha^t\tilde{\mathcal{R}}^{(0)} - 2\left(\frac{1-\alpha^t}{1-\alpha}\right)\tilde{\sigma}^{(0)}, 0\right\} \leq \tilde{\mathcal{R}}^{(t)} \leq \alpha^t\tilde{\mathcal{R}}^{(0)} + 2\left(\frac{1-\alpha^t}{1-\alpha}\right)\tilde{\sigma}^{(0)}$$

Moreover, upon convergence:

$$0 \leq \lim_{t\to\infty}\tilde{\mathcal{R}}^{(t)} \leq \frac{2\tilde{\sigma}^{(0)}}{1-\alpha}$$

Note: While it appears that a large initial maximal deviation in feature values within a group may harm fairness, a large initial deviation does not necessarily entail diversity. For example, suppose a few nodes in a group have a low initial feature value but many more nodes have a much higher initial feature value (i.e., large initial difference without diversity). Then, after mean aggregation, the feature values for all the nodes in the group may be higher on average than they were initially, and more distinct on average from the node feature values in the other group. This would contribute to a higher discrimination risk.

A.5 Extending Theorem 2

We can extend Theorem 2 to the case the number of features $d > 1$. By Theorem 1, the modified discrimination risk at iteration t (including all features) is:

$$\max\left\{\min_{i\in[d]}\alpha_i^t\tilde{\mathcal{R}}_i^{(0)} - 2\left(\sum_{j=0}^{t-1}\alpha_i^j\right)\tilde{\sigma}_i^{(0)}, 0\right\} \leq \max_{i\in[d]}\tilde{\mathcal{R}}_i^{(t)} \leq \max_{i\in[d]}\alpha_i^t\tilde{\mathcal{R}}_i^{(0)} + 2\left(\sum_{j=0}^{t-1}\alpha_i^j\right)\tilde{\sigma}_i^{(0)}$$

Moreover, assuming $\forall i \in [d], \alpha_i < 1$, upon convergence, the discrimination risk is: $\max_{i\in[d]}\lim_{t\to\infty}\tilde{\mathcal{R}}_i^{(t)} \leq \max_{i\in[d]}\frac{2\tilde{\sigma}_i^{(0)}}{1-\alpha_i}$.

A.6 Proof of Theorem 3

We want to constrain the discrimination risk of mean aggregation feature imputation to be at most ϵ. To this end, we can modify mean aggregation feature imputation to update $X_U^{(t+1)} := P_W Z_U^{(t)} + P_B$ such that $X^{(t+1)}$ has a discrimination risk of at most ϵ for all t. $|\mathbb{E}_{q\sim Q}[X_q^{(t+1)}] - \mathbb{E}_{r\sim R}[X_r^{(t+1)}]| = |\frac{1}{|Q|}\sum_{q\in Q\cap K} X_q + \frac{1}{|Q|}\sum_{q\in Q\cap U} Z_q^{(t)} - (\frac{1}{|R|}\sum_{r\in R\cap K} X_r + \frac{1}{|R|}\sum_{r\in R\cap U} Z_r^{(t)})|$. Hence, we have a closed convex polytope wherein unknown feature values yield discrimination risk of at most ϵ:

$$\mathcal{R}_K - \epsilon \leq \frac{1}{|Q|}\sum_{q\in Q\cap U} Z_q^{(t)} - \frac{1}{|R|}\sum_{r\in R\cap U} Z_r^{(t)} = c^T Z_U^{(t)} \leq \mathcal{R}_K + \epsilon$$

If $\mathcal{R}_K - \epsilon \leq c^T Z_U^{(t)} \leq \mathcal{R}_K + \epsilon$, then $P_W = I_{|U|}$ and $P_B = 0$. Otherwise, we must project onto the closer of the two boundaries of the polytope. In this case, $P_W = I_{|U|} - \frac{cc^T}{c^T c}$ and $P_B = \frac{cc^T}{c^T c}\begin{cases}\mathcal{R}_K - \epsilon, & c^T Z_U^{(t)} < \mathcal{R}_K - \epsilon \\ \mathcal{R}_K + \epsilon, & c^T Z_U^{(t)} > \mathcal{R}_K + \epsilon\end{cases}$.

The affine projection we perform at each step is closed and convex. Furthermore, ℓ is $\lambda_{max}(\Delta_{UU})$-smooth for the Euclidean norm (where λ_{max} is the maximum eigenvalue) because for $x_1, x_2 \in \mathbb{R}^{|U|}$:

$$\begin{aligned}\|\nabla\ell(x_1) - \nabla\ell(x_2)\|_2 &= \|(\Delta_{UU}x_1 + \Delta_{UK}X_K) - (\Delta_{UU}x_2 + \Delta_{UK}X_K)\|_2 \\ &= \sqrt{(x_1 - x_2)^T\Delta_{UU}^2(x_1 - x_2)} \\ &\leq \sqrt{\lambda_{max}^2(\Delta_{UU})\|x_1 - x_2\|_2^2} \\ &= \lambda_{max}(\Delta_{UU})\|x_1 - x_2\|_2\end{aligned}$$

In the case of Feature Propagation, $\lambda_{max}(\Delta_{UU}) < 1$ due to properties of the symmetric normalized Laplacian [16].

Additionally, for $m \geq 0$, when $m \leq \lambda_{min}(\Delta_{UU})$, $\ell(x) - \frac{m}{2}x^T x$ is convex because:

$$\begin{aligned}\ell(x) - \frac{m}{2}x^T x &= \frac{1}{2}x^T\Delta_{UU}x + X_K^T\Delta_{KU}x + \frac{1}{2}X_K^T\Delta_{KK}X_K - \frac{m}{2}x^T x \\ &= \frac{1}{2}x^T(\Delta_{UU} - mI)x + X_K^T\Delta_{KU}x + \frac{1}{2}X_K^T\Delta_{KK}X_K\end{aligned}$$

This expression is convex if and only if its Hessian $\Delta_{UU} - mI$ has nonnegative eigenvalues. Therefore, m can be at most $\lambda_{min}(\Delta_{UU})$.

Then, by [95] and [96]:

1. a unique optimal (with respect to ℓ) feasible solution X_U^* exists
2. for fixed step size $\gamma = \frac{1}{\lambda_{max}(\Delta_{UU})}$, ϵ-fair imputation converges as $\|X_U^{(t)} - X_U^*\|_2^2 \leq \left(1 - \frac{\lambda_{min}(\Delta_{UU})}{\lambda_{max}(\Delta_{UU})}\right)^t \|X_U^{(0)} - X_U^*\|_2^2$
3. for fixed step size $\gamma \leq \frac{1}{\lambda_{max}(\Delta_{UU})}$, ϵ-fair imputation converges to X_U^*

A.7 Theorem 4

We have a solution when $\beta > 0$ (i.e., when the known node feature values do not remain fixed). We can view the update of $X^{(t+1)} := \begin{bmatrix}\beta I_{|K|} & 0 \\ 0 & I_{|U|}\end{bmatrix} M^{-1}TMX^{(t)} + \begin{bmatrix}(1-\beta)I_{|K|} & 0 \\ 0 & 0\end{bmatrix} X$ as an iteration of gradient descent (with step size $\gamma = 1$) for the objective function $\ell(x) = \frac{1}{2}x^T\Delta x + \frac{1}{2}(\frac{1-\beta}{\beta})\|x_K - X_K\|_2^2$ [94].

Theorem 4 (ϵ-Fair Imputation, $\beta > 0$) Vanilla mean aggregation feature imputation updates $X^{(t+1)} := \begin{bmatrix}\beta I_{|K|} & 0 \\ 0 & I_{|U|}\end{bmatrix}(I_N - \Delta)X^{(t)} + \begin{bmatrix}(1-\beta)I_{|K|} & 0 \\ 0 & 0\end{bmatrix} X = Z^{(t)}$. Let ϵ-fair mean aggregation

feature imputation instead update $X^{(t+1)} := P_W Z^{(t)} + P_B$, where:

$$P_W = \begin{cases} I_N, & -\epsilon \leq c^T Z^{(t)} \leq \epsilon \\ I_N - \frac{cc^T}{c^T c}, & \text{otherwise} \end{cases}, P_B = \frac{cc^T}{c^T c} \begin{cases} -\epsilon, & c^T Z^{(t)} < -\epsilon \\ \epsilon, & c^T Z^{(t)} > \epsilon \\ 0, & \text{otherwise} \end{cases}$$

$$c \in \mathbb{R}^N, c^T Z^{(t)} = \frac{1}{|Q|} \sum_{q \in Q} Z_q^{(t)} - \frac{1}{|R|} \sum_{r \in R} Z_r^{(t)}$$

Then, assuming $0 \leq \lambda_{min}(\Delta) + \frac{1-\beta}{\beta} \leq \lambda_{max}(\Delta) + \frac{1-\beta}{\beta} < 1$: 1) a unique optimal (with respect to ℓ) feasible solution X^* exists; 2) for fixed step size $\gamma = \frac{1}{\lambda_{max}(\Delta)+\frac{1-\beta}{\beta}}$, ϵ-fair imputation converges as $\|X^{(t)} - X^*\|_2^2 \leq \left(1 - \frac{\lambda_{min}(\Delta)+\frac{1-\beta}{\beta}}{\lambda_{max}(\Delta)+\frac{1-\beta}{\beta}}\right)^t \|X^{(0)} - X^*\|_2^2$; 3) for fixed step size $\gamma \leq \frac{1}{\lambda_{max}(\Delta)+\frac{1-\beta}{\beta}}$, ϵ-fair imputation converges to X^*.

Proof of Theorem 4 We want to constrain the discrimination risk of mean aggregation feature imputation to be at most ϵ. To this end, we can modify mean aggregation feature imputation to update $X^{(t+1)} := P_W Z^{(t)} + P_B$ such that $X^{(t+1)}$ has a discrimination risk of at most ϵ for all t.

$|\mathbb{E}_{q \sim Q}[X_q^{(t+1)}] - \mathbb{E}_{r \sim R}[X_r^{(t+1)}]| = \frac{1}{|Q|} \sum_{q \in Q} Z_q^{(t)} - \frac{1}{|R|} \sum_{r \in R} Z_r^{(t)}$. Hence, we have a closed convex polytope wherein feature values have discrimination risk of at most ϵ:

$$-\epsilon \leq \frac{1}{|Q|} \sum_{q \in Q} Z_q^{(t)} - \frac{1}{|R|} \sum_{r \in R} Z_r^{(t)} = c^T Z^{(t)} \leq \epsilon$$

If $-\epsilon \leq c^T Z^{(t)} \leq \epsilon$, then $P_W = I_N$ and $P_B = 0$. Otherwise, we must project onto the closer of the two boundaries of the polytope. In this case, $P_W = I_N - \frac{cc^T}{c^T c}$ and $P_B = \frac{cc^T}{c^T c} \begin{cases} -\epsilon, & c^T Z^{(t)} < -\epsilon \\ \epsilon, & c^T Z^{(t)} > \epsilon \end{cases}$.

The affine projection we perform at each step is closed and convex. Furthermore, ℓ is $\left(\lambda_{max}(\Delta) + \frac{1-\beta}{\beta}\right)$-smooth for the Euclidean norm because for $x_1, x_2 \in \mathbb{R}^N$:

$$\begin{aligned} \|\nabla \ell(x_1) - \nabla \ell(x_2)\|_2 &= \left\| \left(\Delta x_1 + \frac{1-\beta}{\beta} ((x_1)_K - X_K) \right) - \left(\Delta x_2 + \frac{1-\beta}{\beta} ((x_2)_K - X_K) \right) \right\|_2 \\ &\leq \sqrt{(x_1 - x_2)^T \Delta^2 (x_1 - x_2)} + \frac{1-\beta}{\beta} \sqrt{(x_1 - x_2)^T \left(\begin{bmatrix} I_{|K|} & 0 \\ 0 & 0 \end{bmatrix} \right)^2 (x_1 - x_2)} \\ &\leq \left(\lambda_{max}(\Delta) + \frac{1-\beta}{\beta} \right) \|x_1 - x_2\|_2 \end{aligned}$$

Additionally, for $m \geq 0$, when $m \leq \lambda_{min}(\Delta) + \frac{1-\beta}{\beta}$, $\ell(x) - \frac{m}{2} x^T x$ is convex because:

$$\begin{aligned} \ell(x) - \frac{m}{2} x^T x &= \frac{1}{2} x^T \Delta x + \frac{1}{2} \left(\frac{1-\beta}{\beta} \right) \|x_K - X_K\|_2^2 - \frac{m}{2} x^T x \\ &= \frac{1}{2} x^T (\Delta - mI) x + \frac{1}{2} \left(\frac{1-\beta}{\beta} \right) \|x_K - X_K\|_2^2 \end{aligned}$$

This expression is convex if and only if its Hessian $\Delta - mI + \frac{1-\beta}{\beta} \begin{bmatrix} I_{|K|} & 0 \\ 0 & 0 \end{bmatrix}$ has nonnegative eigenvalues. Therefore, m can be at most $\lambda_{min}(\Delta) + \frac{1-\beta}{\beta}$.

Then, by [95] and [96]:

1. a unique optimal (with respect to ℓ) feasible solution X^* exists
2. for fixed step size $\gamma = \frac{1}{\lambda_{max}(\Delta)+\frac{1-\beta}{\beta}}$, ϵ-fair imputation converges as $\|X^{(t)} - X^*\|_2^2 \leq \left(1 - \frac{\lambda_{min}(\Delta)+\frac{1-\beta}{\beta}}{\lambda_{max}(\Delta)+\frac{1-\beta}{\beta}}\right)^t \|X^{(0)} - X^*\|_2^2$
3. for fixed step size $\gamma \leq \frac{1}{\lambda_{max}(\Delta)+\frac{1-\beta}{\beta}}$, ϵ-fair imputation converges to X^*

B Additional experimental results

B.1 Datasets

SBM synthetic datasets Each network has 500 train nodes, 250 validation nodes, and 250 test nodes, split uniformly at random. Each node has a 10-dimensional feature vector sampled as described in the documentation[5]. All edges have a weight of 1. PyTorch Geometric is used in accordance with its MIT license.

Real-world datasets In the `Credit defaulter` dataset, each node has 13 features (e.g., education, credit history, etc.), with an average degree of 95.79 ± 85.88 [29]. In the `German credit` dataset, each node has 27 features (e.g., loan amount, account-related features, etc.), with an average degree of 44.48 ± 26.51. For both datasets, we use a 50/25/25 train/validation/test split, with each split comprising an equal portion of each label, and we do not include group membership as a feature. To the best of our knowledge (via manual sampling and inspection), neither dataset contains personally identifiable information or offensive content. We use [29]'s data and data loading code[6] in accordance with the MIT license.

B.2 Imputation algorithms

We run GM and NM for 1 iteration each, and FP and GR for 40 iterations. We adapted the code for data utilities, Feature Propagation, and model training from [16][7] in accordance with its Apache-2.0 license. We state all changes in this paper. We implement all algorithms using PyTorch, in accordance with its license [97].

B.3 Models and training

For mlp and gcn, we use a hidden dimension of 64. We train all models full-batch using the Adam optimizer with a learning rate of 0.005 and Dropout rate of 0.5 [98, 99]. We also use early stopping with a patience of 200 epochs, i.e., we stop training when the best validation accuracy has not changed for 200 epochs, and train for a maximum of 10000 epochs. We do not do any hyperparameter tuning. We implement and train all models using PyTorch and PyTorch Geometric [97, 83]. We train all models on a single `tesla v100-sxm2-16gb` GPU on an internal cluster.

B.4 Performance Evaluation

To evaluate imputed features for `SBM`, since we don't have labels, we employ relative reconstruction error $\mathbf{RE}$ (calculated as $\|X_{true} - X_{pred}\|_2 / \|X_{true}\|_2$, where X_{true} and X_{pred} are the ground-truth and imputed features, respectively [16]. A lower reconstruction error is better, and we would like regular mean aggregation imputation and its ϵ-fair counterparts to have comparable reconstruction errors. To measure performance on the real-world datasets, we consider the test accuracy ($\mathbf{Acc}$) of models applied to the imputed data. A higher test accuracy is preferable, and we again would like comparable accuracies for regular and ϵ-fair imputation.

To evaluate group fairness, we compute the discrimination risk ($\mathbf{DR}$) of the imputed data. A lower discrimination risk is preferable. For the `SBM` synthetic datasets, we also measure how much information the imputed features contain about group membership. We do this by training the models on the imputed data to predict group membership and calculate the test accuracy of the models on identifying group membership (which we refer to as $\mathbf{MI}$) [86, 21]. (We note that this setting may violate our theoretical assumptions in 3 that the association of group membership with model predictions can be fully explained by the node features.) The models may be conceptualized as adversaries attempting to recover group membership from the imputed features. Thus, we would like $\mathbf{MI}$ to be closer to 0.5 (i.e., the imputed features contain no information about group membership). We do not compute $\mathbf{MI}$ for the real-world datasets, as inferring group membership or identity from real-world data is invasive, invalid, and can be weaponized against marginalized communities (e.g., to find

[5] `https://pytorch-geometric.readthedocs.io/en/latest/modules/datasets.html#torch_geometric.datasets.StochasticBlockModelDataset`

[6] `https://github.com/chirag126/nifty`

[7] `https://github.com/twitter-research/feature-propagation`

Table 3: Reconstruction error ($\mathbf{RE}$), discrimination risk ($\mathbf{DR}$), and test group membership identification accuracy ($\mathbf{MI}$) of all models averaged over relative sizes of group Q of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ in SBM. We use 0.5 unknown feature rates for both groups and 0.5 inter- and intra-link rates.

Method	$\mathbf{RE}\downarrow$	$\mathbf{DR}\downarrow$	$\mathbf{MI}_{linear}\downarrow$	$\mathbf{MI}_{mlp}\downarrow$	$\mathbf{MI}_{gcn}\downarrow$
0-Fair GM	1.054 ± 0.004	$\mathbf{0 \pm 0}$	$\mathbf{0.758 \pm 0.025}$	$\mathbf{0.78 \pm 0.014}$	0.742 ± 0.018
0.025-Fair GM	1.051 ± 0.004	0.022 ± 0.004	0.79 ± 0.012	0.787 ± 0.012	$\mathbf{0.74 \pm 0.014}$
0.05-Fair GM	1.048 ± 0.003	0.032 ± 0.003	0.791 ± 0.018	0.794 ± 0.013	0.747 ± 0.019
Regular GM	$\mathbf{1 \pm 0}$	0.041 ± 0.008	0.835 ± 0.01	0.845 ± 0.011	0.771 ± 0.027
0-Fair NM	1.015 ± 0.003	$\mathbf{0 \pm 0}$	$\mathbf{0.757 \pm 0.023}$	$\mathbf{0.792 \pm 0.015}$	$\mathbf{0.738 \pm 0.016}$
0.025-Fair NM	1.012 ± 0.003	0.019 ± 0.003	0.791 ± 0.022	0.8 ± 0.014	0.744 ± 0.011
0.05-Fair NM	1.009 ± 0.003	0.029 ± 0.006	0.787 ± 0.017	0.807 ± 0.011	0.746 ± 0.017
Regular NM	$\mathbf{0.959 \pm 0.003}$	0.038 ± 0.013	0.835 ± 0.011	0.843 ± 0.015	0.763 ± 0.024
0-Fair FP	1.003 ± 0.005	$\mathbf{0 \pm 0}$	$\mathbf{0.757 \pm 0.019}$	$\mathbf{0.799 \pm 0.014}$	$\mathbf{0.736 \pm 0.013}$
0.025-Fair FP	1 ± 0.005	0.021 ± 0.002	0.785 ± 0.021	0.801 ± 0.014	0.754 ± 0.017
0.05-Fair FP	0.997 ± 0.005	0.033 ± 0.005	0.789 ± 0.016	0.806 ± 0.012	0.738 ± 0.016
Regular FP	$\mathbf{0.947 \pm 0.005}$	0.051 ± 0.017	0.829 ± 0.006	0.841 ± 0.02	0.760 ± 0.022
0-Fair GR	0.962 ± 0.005	$\mathbf{0 \pm 0}$	$\mathbf{0.752 \pm 0.024}$	$\mathbf{0.788 \pm 0.019}$	0.742 ± 0.013
0.025-Fair GR	0.961 ± 0.005	0.023 ± 0.003	0.797 ± 0.015	0.797 ± 0.02	0.752 ± 0.016
0.05-Fair GR	0.96 ± 0.005	0.036 ± 0.005	0.799 ± 0.009	0.805 ± 0.014	$\mathbf{0.739 \pm 0.017}$
Regular GR	$\mathbf{0.945 \pm 0.006}$	0.036 ± 0.012	0.821 ± 0.015	0.82 ± 0.014	0.759 ± 0.021

Table 4: Reconstruction error ($\mathbf{RE}$), discrimination risk ($\mathbf{DR}$), and test group membership identification accuracy ($\mathbf{MI}$) of all models averaged over all 25 combinations of inter- and intra-link rates of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ in SBM. We use 0.5 relative group sizes and 0.5 unknown feature rates for both groups.

Method	$\mathbf{RE}\downarrow$	$\mathbf{DR}\downarrow$	$\mathbf{MI}_{linear}\downarrow$	$\mathbf{MI}_{mlp}\downarrow$	$\mathbf{MI}_{gcn}\downarrow$
0-Fair NM	1.028 ± 0.009	$\mathbf{0 \pm 0}$	$\mathbf{0.609 \pm 0.102}$	$\mathbf{0.729 \pm 0.046}$	$\mathbf{0.905 \pm 0.003}$
0.025-Fair NM	1.023 ± 0.008	0.014 ± 0.011	0.724 ± 0.09	0.749 ± 0.035	0.911 ± 0.008
0.05-Fair NM	1.019 ± 0.008	0.02 ± 0.02	0.74 ± 0.046	0.768 ± 0.036	0.911 ± 0.01
Regular NM	$\mathbf{0.931 \pm 0.003}$	0.022 ± 0.024	0.845 ± 0.026	0.866 ± 0.027	0.924 ± 0.008
0-Fair FP	1.022 ± 0.012	$\mathbf{0 \pm 0}$	$\mathbf{0.64 \pm 0.064}$	$\mathbf{0.742 \pm 0.038}$	$\mathbf{0.905 \pm 0.003}$
0.025-Fair FP	1.017 ± 0.012	0.014 ± 0.012	0.697 ± 0.095	0.753 ± 0.039	0.912 ± 0.007
0.05-Fair FP	1.013 ± 0.012	0.023 ± 0.022	0.740 ± 0.040	0.762 ± 0.04	0.909 ± 0.008
Regular FP	$\mathbf{0.918 \pm 0.004}$	0.034 ± 0.043	0.844 ± 0.025	0.853 ± 0.035	0.922 ± 0.009
0-Fair GR	0.948 ± 0.043	$\mathbf{0 \pm 0}$	$\mathbf{0.578 \pm 0.105}$	$\mathbf{0.773 \pm 0.038}$	$\mathbf{0.905 \pm 0.004}$
0.025-Fair GR	0.946 ± 0.004	0.016 ± 0.013	0.779 ± 0.036	0.793 ± 0.038	0.915 ± 0.005
0.05-Fair GR	0.945 ± 0.004	0.02 ± 0.02	0.769 ± 0.018	0.797 ± 0.0366	0.912 ± 0.009
Regular GR	$\mathbf{0.916 \pm 0.005}$	0.023 ± 0.032	0.846 ± 0.023	0.864 ± 0.023	0.921 ± 0.009

and incarcerate LGBTQIA+ individuals) [100]. To evaluate group fairness for the real-world datasets, we use the test statistical parity ($\mathbf{SP}$) of the models, defined as $|\mathbb{P}(Z = 1|S = Q) - \mathbb{P}(Z = 1|S = R)|$ (disparity in positive outcome rate for the groups) [78], and test equalized odds ($\mathbf{EO}$), defined as $|\mathbb{P}(Z = 1|S = Q, Y = 1) - \mathbb{P}(Z = 1|S = R, Y = 1)|$ (disparity in accuracy of predicting positive outcome for the groups) [87].

B.5 Contraction coefficient

As we analyzed, Figures 2 to 12 show that, for SBM: 1) a low unknown feature rate for both groups or disparate unknown feature rates across the groups increases α and the discrimination risk (Figures 2, 5, 8, 11); 2) group size alone does not affect α or the discrimination risk (Figures 3, 6, 9, 12); 3) a lower inter-link to intra-link ratio increases α and the discrimination risk (Figures 4, 7, 10).

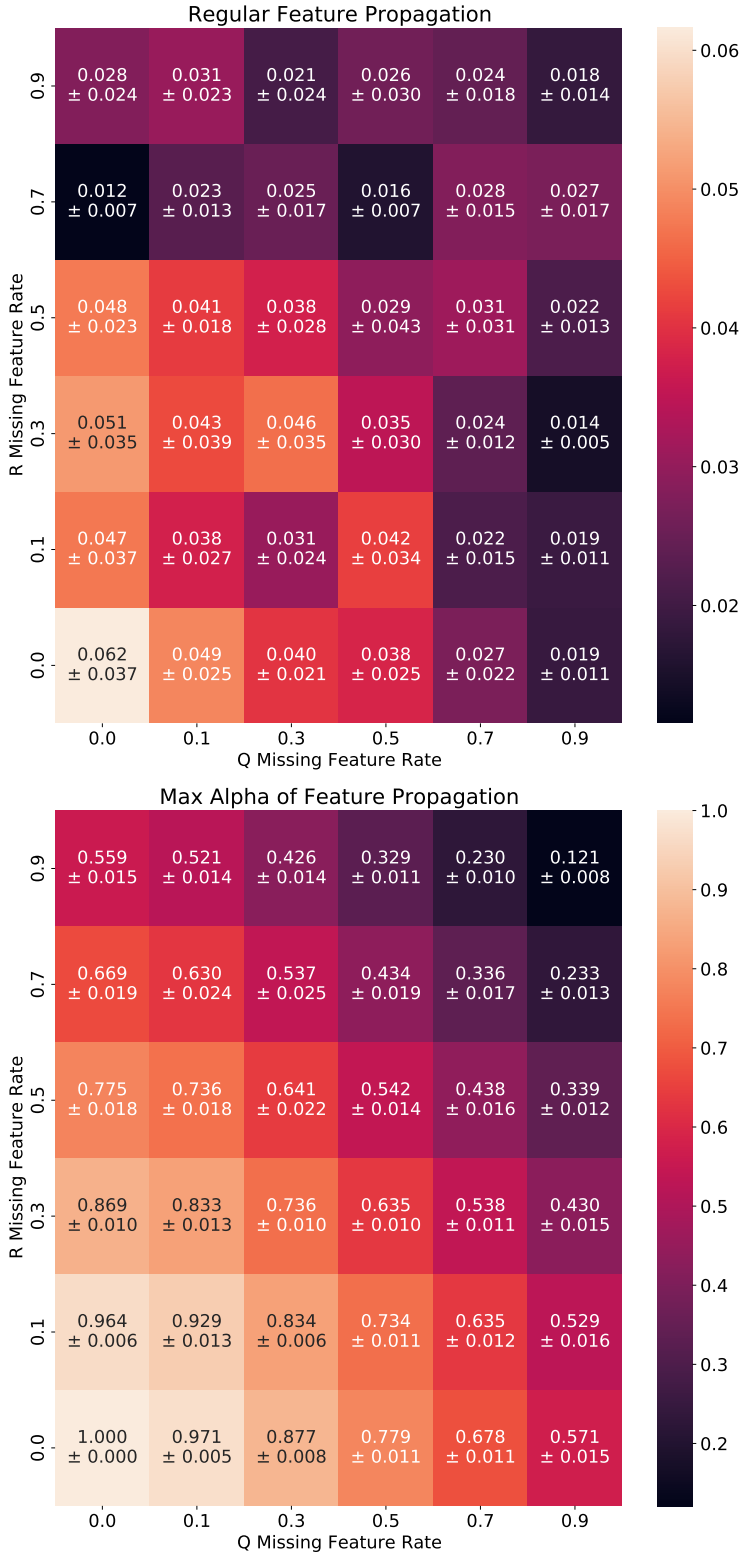

Figure 2: Heatmap of discrimination risks and maximum α (over all channels) of Feature Propagation for 36 combinations of unknown feature rates for each group in SBM. We use 0.5 relative group sizes and 0.5 inter- and intra-link rates.

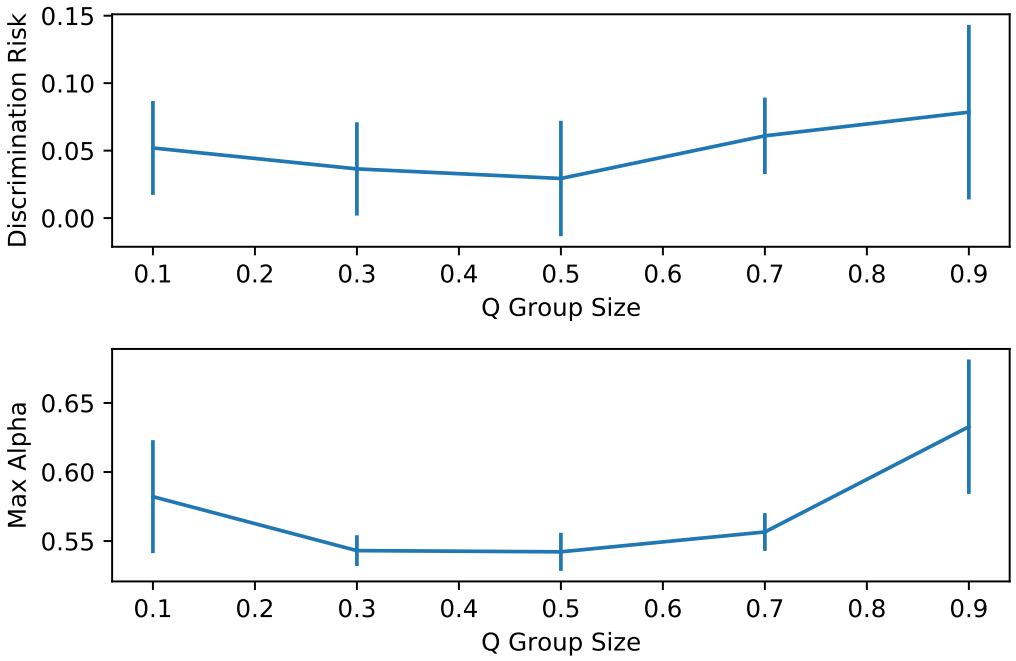

Figure 3: Plots of discrimination risk and maximum α (over all channels) of Feature Propagation vs. relative size of group Q in `SBM`. We use 0.5 unknown feature rates for both groups and 0.5 inter- and intra-link rates.

Table 5: equalized odds (**EO**) averaged over all 25 combinations of unknown feature rates of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ for each group in `German credit`.

Method	$\mathbf{EO}_{linear} \downarrow$	$\mathbf{EO}_{mlp} \downarrow$	$\mathbf{EO}_{gcn} \downarrow$
0.0-Fair GM	0.037 ± 0.01	0.029 ± 0.004	0.009 ± 0.008
0.025-Fair GM	0.031 ± 0.007	0.029 ± 0.008	0.018 ± 0.021
0.05-Fair GM	$\mathbf{0.026 \pm 0.003}$	0.03 ± 0.003	0.013 ± 0.005
Regular GM	0.033 ± 0.008	$\mathbf{0.023 \pm 0.003}$	$\mathbf{0.006 \pm 0.005}$
0.0-Fair NM	0.038 ± 0.009	0.037 ± 0.006	$\mathbf{0.007 \pm 0.006}$
0.025-Fair NM	$\mathbf{0.035 \pm 0.008}$	0.038 ± 0.007	0.013 ± 0.012
0.05-Fair NM	0.04 ± 0.006	0.035 ± 0.006	0.009 ± 0.003
Regular NM	0.038 ± 0.012	$\mathbf{0.032 \pm 0.006}$	0.012 ± 0.006
0.0-Fair FP	$\mathbf{0.01 \pm 0.011}$	0.034 ± 0.018	0.024 ± 0.041
0.025-Fair FP	0.028 ± 0.031	0.031 ± 0.018	0.023 ± 0.051
0.05-Fair FP	0.043 ± 0.07	$\mathbf{0.029 \pm 0.028}$	$\mathbf{0 \pm 0}$
Regular FP	0.042 ± 0.046	0.038 ± 0.02	0.004 ± 0.006
0.0-Fair GR	0.029 ± 0.012	$\mathbf{0.022 \pm 0.003}$	0.005 ± 0.006
0.025-Fair GR	0.031 ± 0.011	0.024 ± 0.005	0.007 ± 0.007
0.05-Fair GR	$\mathbf{0.027 \pm 0.007}$	0.024 ± 0.006	$\mathbf{0.004 \pm 0.004}$
Regular GR	0.032 ± 0.01	0.025 ± 0.007	0.009 ± 0.01

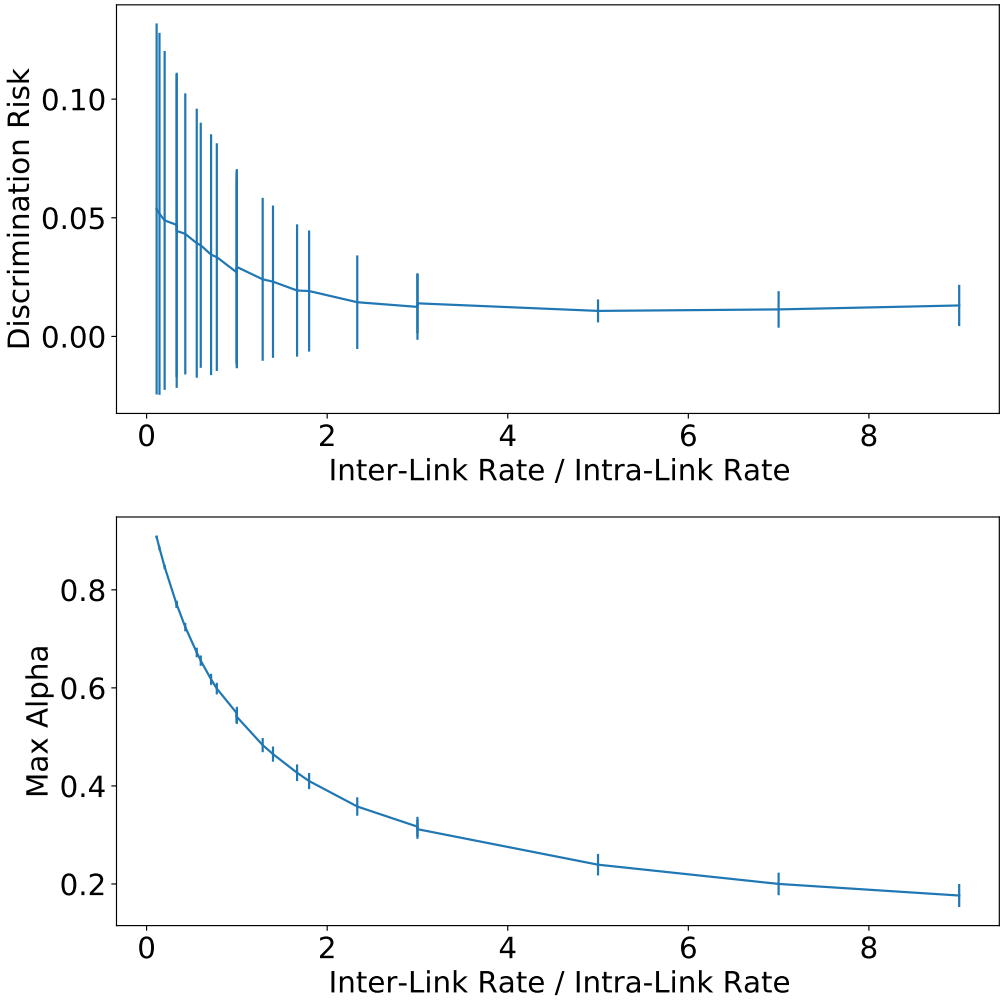

Figure 4: Plots of discrimination risk and maximum α (over all channels) of Feature Propagation vs. ratio of inter-link rate to intra-link rate in `SBM`. We use 0.5 relative group sizes and 0.5 unknown feature rates for both groups.

Table 6: Test accuracy (**Acc**) and statistical parity (**SP**) of all models averaged over all 25 combinations of unknown feature rates of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ for each group in `Credit defaulter`.

Method	$\mathbf{Acc}_{linear} \uparrow$	$\mathbf{Acc}_{mlp} \uparrow$	$\mathbf{Acc}_{gcn} \uparrow$	$\mathbf{SP}_{linear} \downarrow$	$\mathbf{SP}_{mlp} \downarrow$	$\mathbf{SP}_{gcn} \downarrow$
0.0-Fair GM	0.781 ± 0.002	$\mathbf{0.764 \pm 0.012}$	0.771 ± 0.007	0.063 ± 0.015	0.08 ± 0.024	0.016 ± 0.009
0.025-Fair GM	0.78 ± 0.006	0.757 ± 0.008	0.774 ± 0.002	0.059 ± 0.021	0.083 ± 0.012	0.015 ± 0.004
0.05-Fair GM	0.78 ± 0.003	0.759 ± 0.013	$\mathbf{0.775 \pm 0.002}$	0.076 ± 0.015	0.08 ± 0.017	0.015 ± 0.003
Regular GM	$\mathbf{0.782 \pm 0.006}$	0.76 ± 0.018	0.775 ± 0.006	$\mathbf{0.055 \pm 0.031}$	$\mathbf{0.056 \pm 0.012}$	$\mathbf{0.005 \pm 0.006}$
0.0-Fair NM	$\mathbf{0.781 \pm 0.002}$	0.765 ± 0.006	0.771 ± 0.007	0.063 ± 0.017	0.085 ± 0.013	0.015 ± 0.013
0.025-Fair NM	0.78 ± 0.005	$\mathbf{0.766 \pm 0.005}$	0.774 ± 0.002	0.057 ± 0.025	0.088 ± 0.008	0.015 ± 0.005
0.05-Fair GM	0.781 ± 0.003	0.769 ± 0.01	$\mathbf{0.775 \pm 0.002}$	0.082 ± 0.011	0.082 ± 0.018	0.016 ± 0.003
Regular GM	0.781 ± 0.007	0.762 ± 0.014	0.773 ± 0.008	$\mathbf{0.054 \pm 0.031}$	$\mathbf{0.061 \pm 0.011}$	$\mathbf{0.005 \pm 0.007}$
0.0-Fair FP	0.779 ± 0.005	0.757 ± 0.022	$\mathbf{0.77 \pm 0.008}$	0.06 ± 0.022	0.085 ± 0.014	0.016 ± 0.014
0.025-Fair FP	0.78 ± 0.002	0.764 ± 0.004	0.774 ± 0.002	0.056 ± 0.027	0.092 ± 0.008	0.016 ± 0.005
0.05-Fair FP	0.78 ± 0.001	$\mathbf{0.768 \pm 0.008}$	0.774 ± 0.002	0.076 ± 0.005	0.084 ± 0.014	0.017 ± 0.004
Regular FP	$\mathbf{0.781 \pm 0.005}$	0.764 ± 0.01	0.775 ± 0.006	$\mathbf{0.051 \pm 0.029}$	$\mathbf{0.075 \pm 0.011}$	$\mathbf{0.005 \pm 0.006}$
0.0-Fair GR	0.773 ± 0.009	$\mathbf{0.796 \pm 0.006}$	0.771 ± 0.011	0.072 ± 0.044	0.098 ± 0.032	0.011 ± 0.012
0.025-Fair GR	0.779 ± 0.005	0.792 ± 0.007	0.772 ± 0.003	0.052 ± 0.032	0.091 ± 0.02	0.017 ± 0.012
0.05-Fair GR	0.78 ± 0.003	0.792 ± 0.007	$\mathbf{0.773 \pm 0.004}$	0.078 ± 0.0314	0.094 ± 0.028	0.023 ± 0.01
Regular GR	$\mathbf{0.781 \pm 0.005}$	0.785 ± 0.004	0.773 ± 0.008	$\mathbf{0.049 \pm 0.043}$	$\mathbf{0.073 \pm 0.038}$	$\mathbf{0.008 \pm 0.01}$

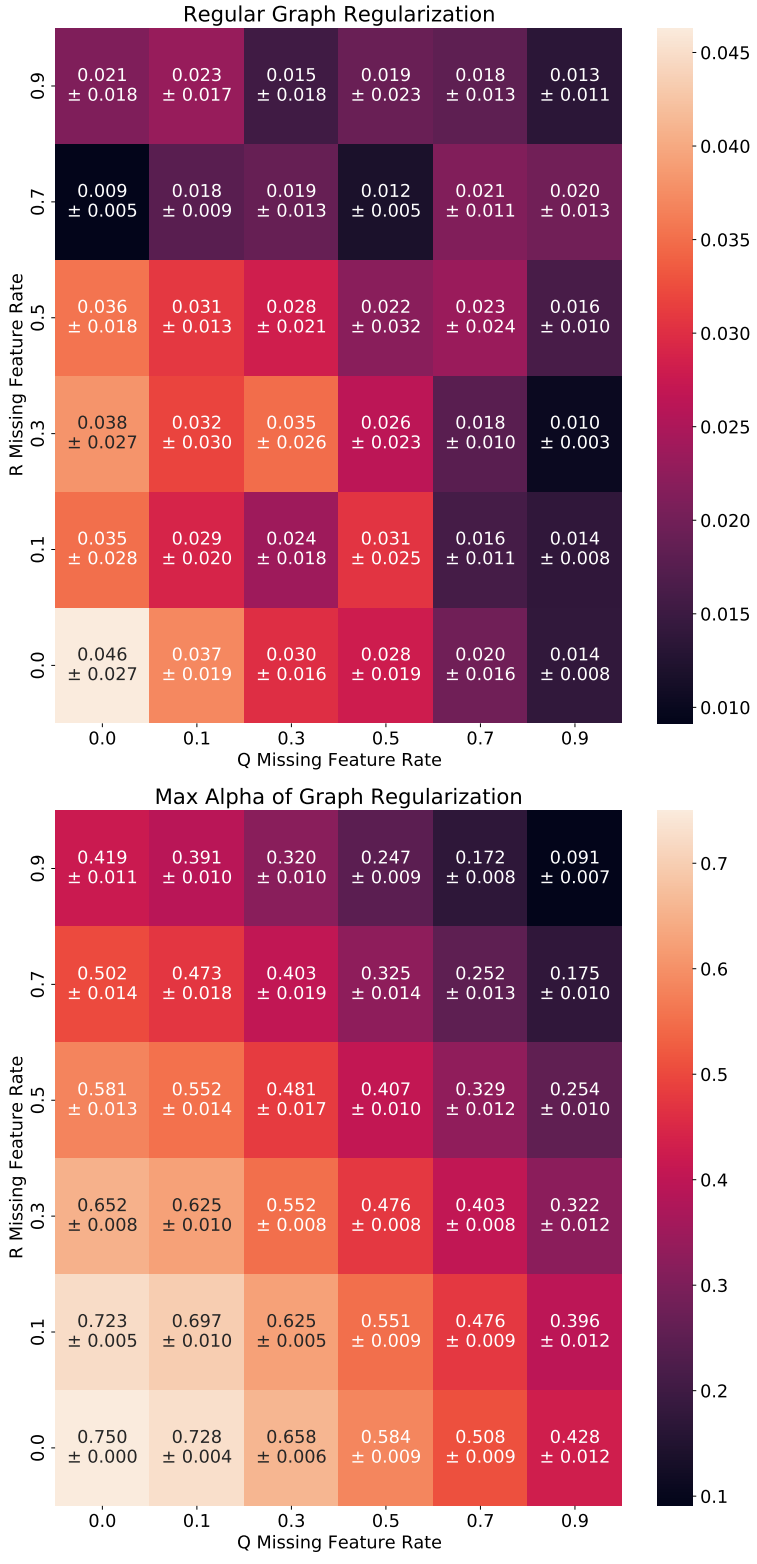

Figure 5: Heatmap of discrimination risks and maximum α (over all channels) of Graph Regularization for 36 combinations of unknown feature rates for each group in SBM. We use 0.5 relative group sizes and 0.5 inter- and intra-link rates.

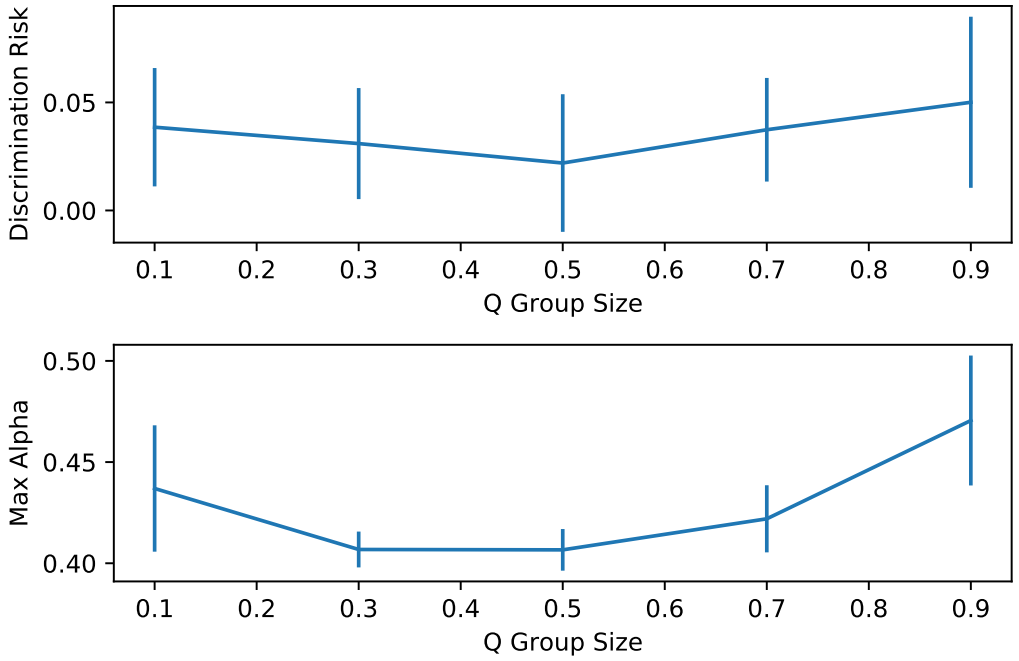

Figure 6: Plots of discrimination risk and maximum α (over all channels) of Graph Regularization vs. relative size of group Q in `SBM`. We use 0.5 unknown feature rates for both groups and 0.5 inter- and intra-link rates.

Table 7: equalized odds (**EO**) of all models averaged over all 25 combinations of unknown feature rates of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ for each group in `Credit defaulter`.

Method	$\mathbf{EO}_{linear} \downarrow$	$\mathbf{EO}_{mlp} \downarrow$	$\mathbf{EO}_{gcn} \downarrow$
0.0-Fair GM	0.039 ± 0.008	0.056 ± 0.019	0.013 ± 0.007
0.025-Fair GM	0.035 ± 0.014	0.058 ± 0.009	0.012 ± 0.003
0.05-Fair GM	0.048 ± 0.010	0.057 ± 0.016	0.012 ± 0.002
Regular GM	$\mathbf{0.031 \pm 0.017}$	$\mathbf{0.038 \pm 0.011}$	$\mathbf{0.004 \pm 0.005}$
0.0-Fair NM	0.039 ± 0.01	0.06 ± 0.009	0.013 ± 0.007
0.025-Fair NM	0.033 ± 0.015	0.06 ± 0.007	0.011 ± 0.004
0.05-Fair NM	0.05 ± 0.007	0.057 ± 0.014	0.012 ± 0.002
Regular NM	$\mathbf{0.031 \pm 0.018}$	$\mathbf{0.041 \pm 0.007}$	$\mathbf{0.005 \pm 0.006}$
0.0-Fair FP	0.035 ± 0.013	0.059 ± 0.013	0.013 ± 0.008
0.025-Fair FP	0.031 ± 0.016	0.062 ± 0.007	0.013 ± 0.004
0.05-Fair FP	0.043 ± 0.005	0.057 ± 0.011	0.014 ± 0.002
Regular FP	$\mathbf{0.028 \pm 0.016}$	$\mathbf{0.05 \pm 0.01}$	$\mathbf{0.004 \pm 0.005}$
0.0-Fair GR	0.051 ± 0.036	0.07 ± 0.029	0.008 ± 0.011
0.025-Fair GR	$\mathbf{0.03 \pm 0.02}$	0.06 ± 0.025	0.012 ± 0.011
0.05-Fair GR	0.048 ± 0.03	0.067 ± 0.025	0.015 ± 0.009
Regular GR	0.03 ± 0.038	$\mathbf{0.051 \pm 0.038}$	$\mathbf{0.007 \pm 0.007}$

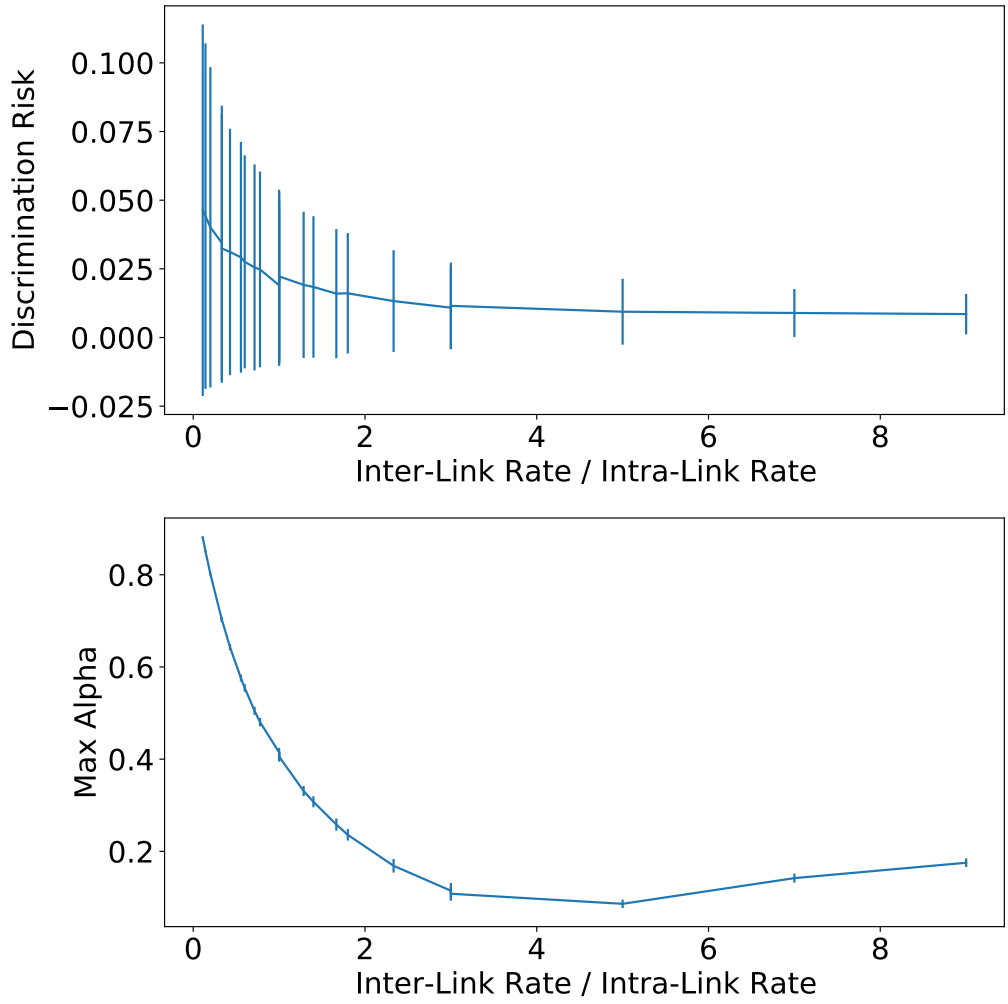

Figure 7: Plots of discrimination risk and maximum α (over all channels) of Graph Regularization vs. ratio of inter-link rate to intra-link rate in `SBM`. We use 0.5 relative group sizes and 0.5 unknown feature rates for both groups.

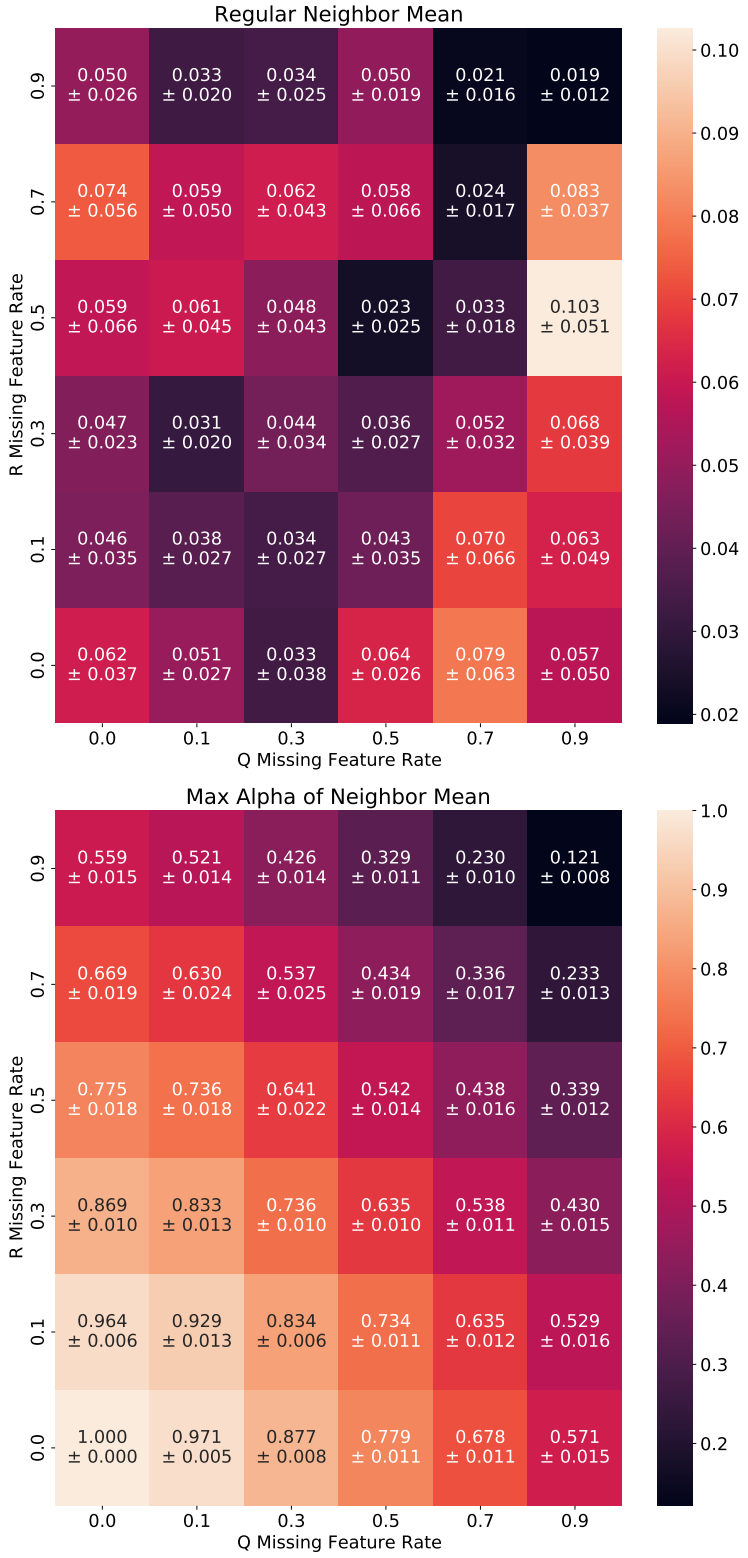

Figure 8: Heatmap of discrimination risks and maximum α (over all channels) of Neighbor Mean for 36 combinations of unknown feature rates for each group in SBM. We use 0.5 relative group sizes and 0.5 inter- and intra-link rates.

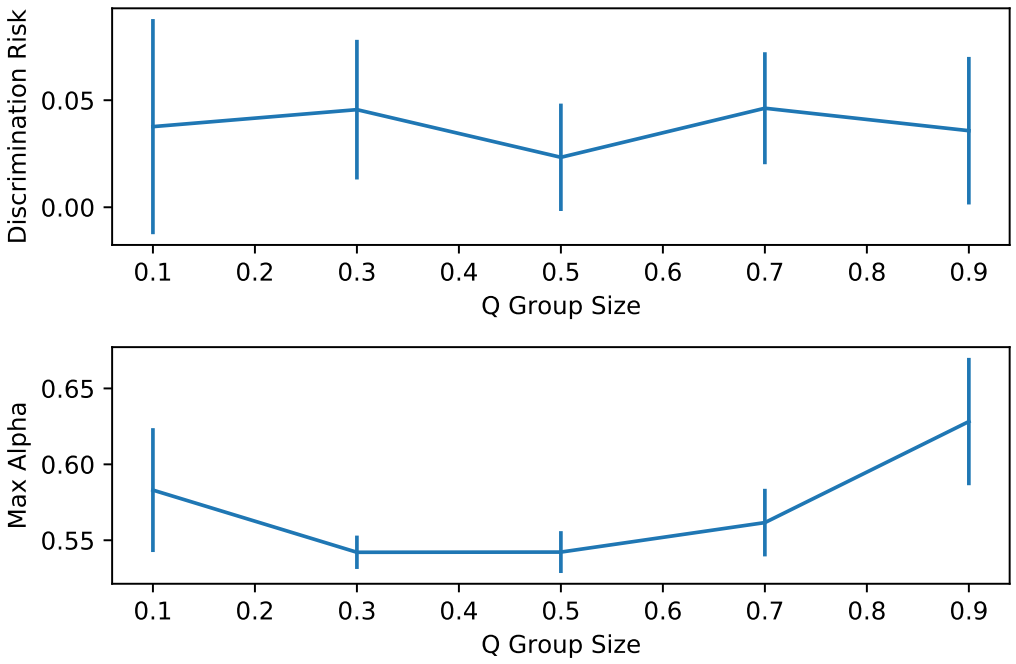

Figure 9: Plots of discrimination risk and maximum α (over all channels) of Neighbor Mean vs. relative size of group Q in `SBM`. We use 0.5 unknown feature rates for both groups and 0.5 inter- and intra-link rates.

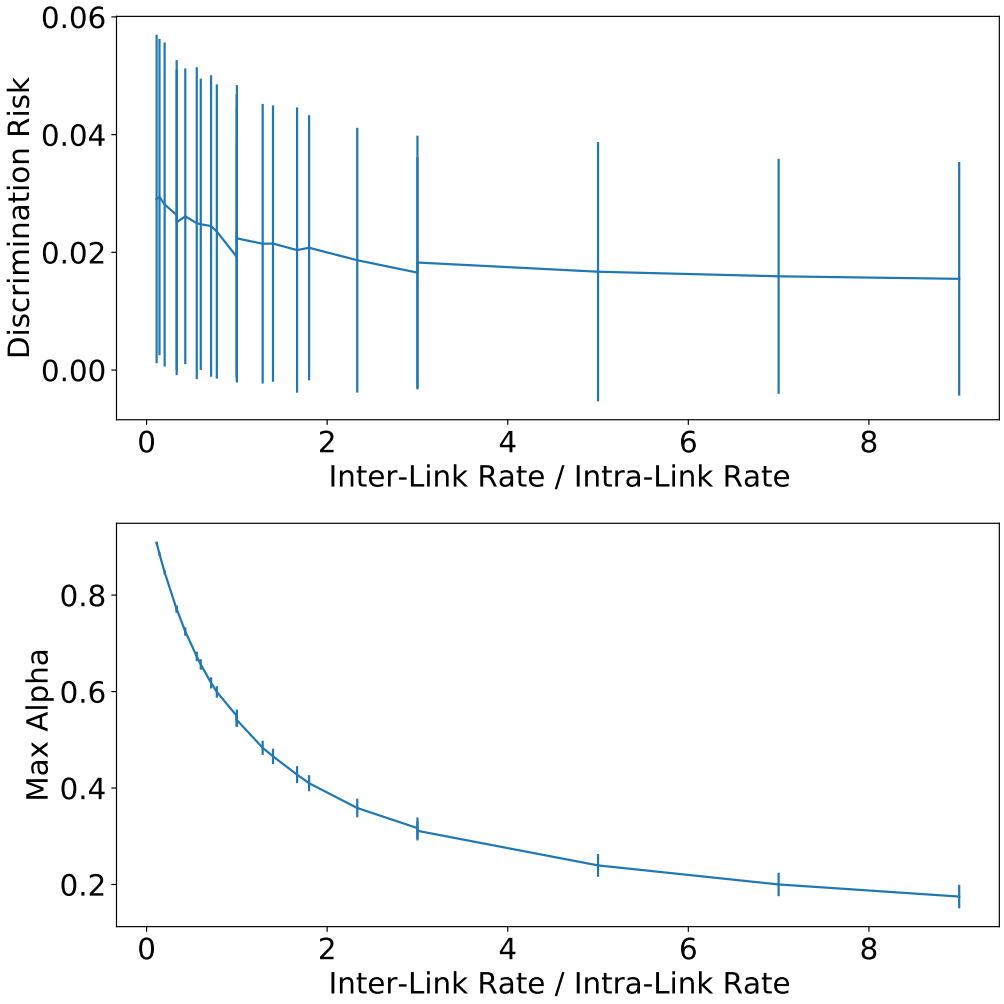

Figure 10: Plots of discrimination risk and maximum α (over all channels) of Neighbor Mean vs. ratio of inter-link rate to intra-link rate in `SBM`. We use 0.5 relative group sizes and 0.5 unknown feature rates for both groups.

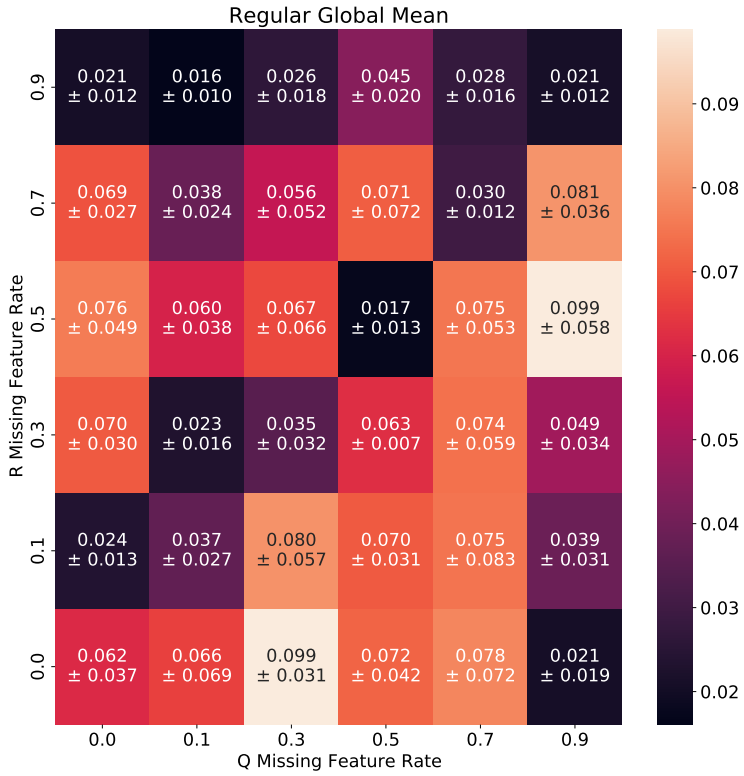

Figure 11: Heatmap of discrimination risks and maximum α (over all channels) of Global Mean for 36 combinations of unknown feature rates for each group in SBM. We use 0.5 relative group sizes. **Note:** Global Mean is not affected by graph structure.

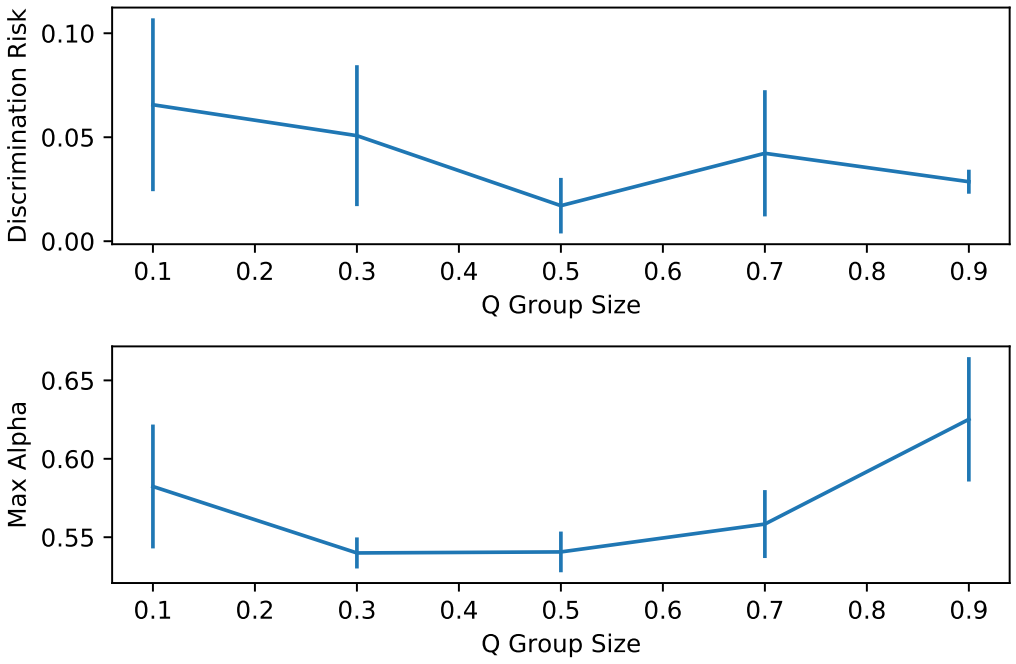

Figure 12: Plots of discrimination risk and maximum α (over all channels) of Global Mean vs. relative size of group Q in SBM. We use 0.5 unknown feature rates for both groups. **Note:** Global Mean is not affected by graph structure.