A Omitted Proofs

A.1 Proof of Theorem 4.1

Proof. Let $\mathcal{I}_{\mathrm{PFC}}$ a given instance of $\mathrm{PFC}(k,p)$, $SOL_{\mathrm{PFC}} = (S^*_{\mathrm{PFC}}, \phi^*_{\mathrm{PFC}})$ the optimal solution of $\mathcal{I}_{\mathrm{PFC}}$ and $\mathrm{OPT}_{\mathrm{PFC}}$ its corresponding optimal value. Also, for $\mathrm{Cluster}(k,p)$ and for any instance of it, the optimal value is denoted by $\mathrm{OPT}_{\mathrm{Cluster}}$ and the corresponding solution by $\mathrm{SOL}_{\mathrm{Cluster}} = (S^*_{\mathrm{Cluster}}, \phi^*_{\mathrm{Cluster}})$.

The proof closely follows that from Bera et al. [2019]. First running the color-blind α approximation algorithm results in a set of centers S, an assignment ϕ, and a solution value that is at most $\alpha\mathrm{OPT}_{\mathrm{Cluster}} \leq \alpha\,\mathrm{OPT}_{\mathrm{PFC}}$. Note that $\mathrm{OPT}_{\mathrm{Cluster}} \leq \mathrm{OPT}_{\mathrm{PFC}}$ since $\mathrm{PFC}(k,p)$ is a more constrained problem than $\mathrm{Cluster}(k,p)$. Now we establish the following lemma:

Lemma A.1. $\mathrm{OPT}_{\text{FA-PFC}} \leq (\alpha+2)\,\mathrm{OPT}_{\mathrm{PFC}}$

Proof. The lemma is established by finding the instance satisfying the inequality. Let $\phi'(v) = \arg\min_{i\in S} d(i, \phi^*_{\mathrm{PFC}}(v))$, i.e. an assignment that routes the vertices from the optimal center to the nearest center in color-blind solution S. For any point v the following holds:

$$\begin{aligned} d(v,\phi'(v)) &\leq d(v,\phi^*_{\mathrm{PFC}}(v)) + d(\phi^*_{\mathrm{PFC}}(v),\phi'(v)) \\ &\leq d(v,\phi^*_{\mathrm{PFC}}(v)) + d(\phi^*_{\mathrm{PFC}}(v),\phi(v)) \\ &\leq d(v,\phi^*_{\mathrm{PFC}}(v)) + d(v,\phi^*_{\mathrm{PFC}}(v)) + d(v,\phi(v)) \\ &= 2d(v,\phi^*_{\mathrm{PFC}}(v)) + d(v,\phi(v)) \end{aligned}$$

stacking the distance values in the vectors $\vec{d}(v,\phi'(v)), \vec{d}(v,\phi^*_{\mathrm{PFC}}(v))$, and $\vec{d}(v,\phi(v))$. By the virtue of the fact that $\big(\sum_{v\in\mathcal{C}} x^p(v)\big)^{1/p}$ is the ℓ_p-norm of the associated vector $\vec{x}$ and since each entry in $\vec{d}(v,\phi'(v))$ is non-negative, the triangular inequality for norms implies:

$$\begin{aligned} \big(\sum_{v\in\mathcal{C}} d^p(v,\phi'(v))\big)^{1/p} &\leq 2\big(\sum_{v\in\mathcal{C}} d^p(v,\phi^*_{\mathrm{PFC}}(v))\big)^{1/p} \\ &+ \big(\sum_{v\in\mathcal{C}} d^p(v,\phi(v))\big)^{1/p} \end{aligned}$$

It remains to show that ϕ' satisfies the fairness constraints 3b, for any color h_ℓ and any center i in S, denote $N(i) = \{j \in S^*_{\mathrm{PFC}} | \arg\min_{i'\in S} d(i',j) = i\}$, then we have:

$$\frac{\sum_{v\in\phi'^{-1}(i)} p_v^{h_\ell}}{|\phi'^{-1}(i)|} = \frac{\sum_{j\in N(i)} \Big(\sum_{v\in\phi^{*-1}_{\mathrm{PFC}}(j)} p_v^{h_\ell}\Big)}{\sum_{j\in N(i)} |\phi^{*-1}_{\mathrm{PFC}}(j)|}$$

It follows by algebra and the lower and upper fairness constrain bounds satisfied by ϕ^*_{PFC}:

$$\begin{aligned} l_{h_\ell} &\leq \min_{j\in N(i)} \frac{\Big(\sum_{v\in\phi^{*-1}_{\mathrm{PFC}}(j)} p_v^{h_\ell}\Big)}{|\phi^{*-1}_{\mathrm{PFC}}(j)|} \\ &\leq \frac{\sum_{j\in N(i)} \Big(\sum_{v\in\phi^{*-1}_{\mathrm{PFC}}(j)} p_v^{h_\ell}\Big)}{\sum_{j\in N(i)} |\phi^{*-1}_{\mathrm{PFC}}(j)|} \\ &\leq \max_{j\in N(i)} \frac{\Big(\sum_{v\in\phi^{*-1}_{\mathrm{PFC}}(j)} p_v^{h_\ell}\Big)}{|\phi^{*-1}_{\mathrm{PFC}}(j)|} \\ &\leq u_{h_\ell} \end{aligned}$$

This shows that there exists an instance for FA-PFC that both satisfies the fairness constraints and has cost $\leq 2\,\mathrm{OPT}_{\mathrm{PFC}} + \alpha\mathrm{OPT}_{\mathrm{Cluster}} \leq (\alpha+2)\,\mathrm{OPT}_{\mathrm{PFC}}$. □

Now combining the fact that we have an α approximation ratio for the color-blind problem, along with an algorithm that achieves a γ violation to FA-PFC with a value equal to the optimal value for FA-PFC, the proof for theorem 4.1 is complete. □

A.2 General Theorem for Lower Bounded Deterministic Fair Clustering

Before stating the theorem and proof, we introduce some definitions. Let FA-PFC-LB denote the fair assignment problem with lower bounded cluster sizes. Specifically, in FA-PFC-LB(S, p, L) we are given a set of clusters S and we seek to find an assignment $\phi : \mathcal{C} \to S$ so that the fairness constraints 8b are satisfied, in addition to constraint 8c for lower bounding the cluster size by at least L.

Note that although we care about the deterministic case, the statement and proof hold for the probabilistic case. Since the deterministic case is a special case of the probabilistic, the proof follows for the deterministic case as well.

Theorem A.1. *Given an α approximation algorithm for the color blind clustering problem* $\mathrm{Cluster}(k, p)$ *and a γ violating algorithm for the fair assignment problem with lower bounded cluster sizes* FA-PFC-LB(S, p, L)*, a solution with approximation ratio $\alpha + 2$ and violation at most γ can be achieved for the deterministic fair clustering problem with lower bounded cluster size* $\mathrm{DFC}_{\mathrm{LB}}(k, p)$.

Proof. First running the color-blind α approximation algorithm results in a set of centers S, an assignment ϕ, and a solution value that is at most $\alpha \mathrm{OPT}_{\mathrm{Cluster}} \leq \alpha\, \mathrm{OPT}_{\mathrm{DFC}_{\mathrm{LB}}}$.

Now we establish the equivalent to lemma A.1 for this problem:

Lemma A.2. *For the fair assignment problem with lower bounded cluster sizes* FA-PFC-LB*, we have that* $\mathrm{OPT}_{\text{FA-PFC-LB}} \leq (\alpha + 2)\, \mathrm{OPT}_{\mathrm{DFC}_{\mathrm{LB}}}$

Proof. The proof is very similar to the proof for lemma A.1. Letting $SOL^*_{\mathrm{DFC}_{\mathrm{LB}}} = (S^*_{\mathrm{DFC}_{\mathrm{LB}}}, \phi^*_{\mathrm{DFC}_{\mathrm{LB}}})$ denote the optimal solution to $\mathrm{DFC}_{\mathrm{LB}}$ with optimal value $\mathrm{OPT}_{\mathrm{DFC}_{\mathrm{LB}}}$. Similarly, define the assignment $\phi'(v) = \arg\min_{i \in S} d(i, \phi^*_{\mathrm{DFC}_{\mathrm{LB}}}(v))$, i.e. an assignment which routs vertices from the optimal center to the closest center in the color-blind solution. By identical arguments to those in the proof of lemma A.1, it follows that:

$$\big(\sum_{v \in \mathcal{C}} d^p(v, \phi'(v))\big)^{1/p} \leq 2\big(\sum_{v \in \mathcal{C}} d^p(v, \phi^*_{\mathrm{DFC}_{\mathrm{LB}}}(v))\big)^{1/p} + \big(\sum_{v \in \mathcal{C}} d^p(v, \phi(v))\big)^{1/p}$$

and that:

$$l_{h_\ell} \leq \frac{\sum_{v \in \phi'^{-1}(i)} p_v^{h_\ell}}{|\phi'^{-1}(i)|} \leq u_{h_\ell}$$

What remains is to show that each cluster is lower bounded by L. Here we note that a center in S will either be allocated the vertices of one or more centers in $S^*_{\mathrm{DFC}_{\mathrm{LB}}}$ or it would not be allocated any vertices at all. If it is not allocated any vertices, then it is omitted as a center (since no vertices are assigned to it). If vertices for a center or more are routed to it, then it will have a cluster of size $\sum_{j \in N(i)} |\phi^{*-1}_{\mathrm{DFC}_{\mathrm{LB}}}(j)| \geq L$. This follows since any center in the optimal solution to $\mathrm{DFC}_{\mathrm{LB}}$ must satisfy the lower bound L. □

Now combining the fact that we have an α approximation ratio for the color-blind problem, along with an algorithm that achieves a γ violation to FA-PFC-LB with value equal to the optimal value for FA-PFC-LB, the proof for theorem A.2 is complete. □

A.3 Proof of the theorem 4.2 (Two-Color and Metric Membership Violation)

Proof. For a given center i, every vertex $q \in C_i$ is assigned some vertices and adds value $\sum_{j \in \phi^{-1}(i,q)} R_j x^q_{ij}$ to the entire average (expected) value of cluster i where $\phi^{-1}(i, q)$ refers to the subset in $\phi^{-1}(i)$ assigned to q. After the rounding, $\sum_{j \in \phi^{-1}(i,q)} R_j x^q_{ij}$ will become $\sum_{j \in \phi^{-1}(i,q)} R_j \bar{x}^q_{ij}$. Denoting $\max_{j \in \phi^{-1}(i,q)} R_j$ and $\min_{j \in \phi^{-1}(i,q)} R_j$ by $R^{max}_{q,i}$ and $R^{min}_{q,i}$, respectively. The following

bounds the maximum violation:

$$\sum_{q=1}^{|C_i|}\Big(\sum_{j\in\phi^{-1}(i,q)} R_j\bar{x}_{ij}^q\Big) - \sum_{q=1}^{|C_i|}\Big(\sum_{j\in\phi^{-1}(i,q)} R_j x_{ij}^q\Big)$$
$$= \sum_{q=1}^{|C_i|}\sum_{j\in\phi^{-1}(i,q)}\Big(R_j\bar{x}_{ij}^q - R_j x_{ij}^q\Big)$$
$$\leq \sum_{q=1}^{|C_i|} R_{q,i}^{max} - R_{q,i}^{min}$$
$$\Big(R_{1,i}^{max} - R_{1,i}^{min}\Big) + \Big(R_{2,i}^{max} - R_{2,i}^{min}\Big)$$
$$+ \Big(R_{3,i}^{max} - R_{3,i}^{min}\Big) + \cdots + \Big(R_{|C_i|,i}^{max} - R_{|C_i|,i}^{min}\Big)$$
$$\leq \Big(R_{1,i}^{max} - R_{1,i}^{min}\Big) + \Big(R_{1,i}^{min} - R_{2,i}^{min}\Big)$$
$$+ \Big(R_{2,i}^{min} - R_{3,i}^{min}\Big) + \cdots + \Big(R_{|C_i|-1,i}^{min} - R_{|C_i|,i}^{min}\Big)$$
$$\leq R_{1,i}^{max} - R_{|C_i|,i}^{min}$$
$$\leq R - 0 = R$$

where we invoked the fact that $R_{k,i}^{max} \leq R_{k-1,i}^{min}$. By following the reverse logic we see that the maximum drop is $-R$. For the probabilistic case, simply $R = 1$. □

A.4 Proof of theorem 4.3 (Lower Bound on the Additive Violation for the Two Color and Metric Membership Case)

Proof. Proof. Consider the following instance (in Figure 5) with 5 points. Points 2 and 4 are chosen as the centers and both clusters have the same radius. The entire set has average color: $\frac{2(0)+2(\frac{3R}{4})+R}{2+2+1} = \frac{\frac{5R}{2}}{5} = \frac{R}{2}$. If the upper and lower values are set to $u = l = \frac{R}{2}$, then the fractional assignments for cluster 1 can be: $x_{21} = 1, x_{22} = 1, x_{23} = \frac{1}{2}$, leading to average color $\frac{\frac{3R}{4}+0+\frac{R}{2}}{1+1+\frac{1}{2}} = \frac{R}{2}$. For cluster 2 we would have: $x_{43} = \frac{1}{2}, x_{44} = 1, x_{45} = 1$ and the average color is $\frac{R(\frac{3}{4}+\frac{1}{2})}{\frac{5}{2}} = \frac{\frac{5R}{4}}{\frac{5}{2}} = \frac{R}{2}$. Only assignments x_{23} and x_{43} are fractional and hence will be rounded. WLOG assume that $x_{23} = 1$ and $x_{43} = 0$. It follows that the change (violation) in the assignment $\sum_j r_j x_{ij}$ for a cluster i will be $\frac{R}{2}$. Consider cluster 1, the resulting color is $\frac{3R}{4} + R = \frac{7R}{4}$, the change is $|\frac{7R}{4} - \frac{5R}{4}| = \frac{R}{2}$. Similarly, for cluster 2 the change is $|\frac{5R}{4} - \frac{3R}{4}| = \frac{R}{2}$.

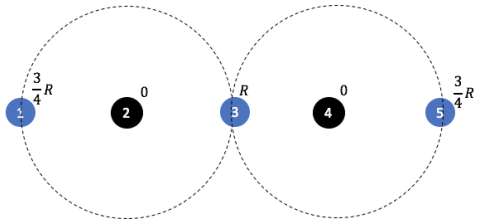

Figure 5: Points 2 and 4 have been selected as centers by the integer solution. Each points has its value written next to.

□

A.5 Proof of Theorem 4.4

Proof. First, each cluster C_i has an amount of color h_ℓ equal to $S_{C_i}^{h_\ell}$ with $\mathbb{E}[S_{C_i}^{h_\ell}] = \sum_{v\in C_i} p_v^{h_\ell}$ according to theorem B.2. Furthermore, since the cluster is valid it follows that: $l_{h_\ell} \leq \mathbb{E}[S_{C_i}^{h_\ell}] \leq u_{h_\ell}$.

Define $l_{\min} = \min_{h_\ell \in \mathcal{H}}\{l_{h_\ell}\} > 0$, then for any $\delta \in [0,1]$ by Theorem B.1 we have:

$$\Pr(|S_{C_i}^{h_\ell} - \mathbb{E}[S_{C_i}^{h_\ell}]| > \delta\,\mathbb{E}[S_{C_i}^{h_\ell}]) \leq 2e^{-\mathbb{E}[S_{C_i}^{h_\ell}]\delta^2/3}$$
$$\leq 2\exp(-\frac{\delta^2}{3}\sum_{v\in C_i} p_v^{h_\ell}) \leq 2\exp(-\frac{\delta^2}{3}Ll_{\min})$$

This upper bounds the failure probability for a given cluster. For the entire set we use the union bound and get:

$$\Pr\left(\left\{\exists i \in \{1,\ldots,k\}, h_\ell \in \mathcal{H} \text{ s.t. } |S_{C_i}^{h_\ell} - \mathbb{E}[S_{C_i}^{h_\ell}]| > \delta\,\mathbb{E}[S_{C_i}^{h_\ell}]\right\}\right)$$
$$\leq 2k|\mathcal{H}|\exp(-\frac{\delta^2}{3}Ll_{\min}) \leq 2\frac{n}{L}|\mathcal{H}|\exp(-\frac{\delta^2}{3}Ll_{\min})$$
$$\leq 2|\mathcal{H}|n^{1-r}\exp(-\frac{\delta^2}{3}l_{\min}n^r)$$

It is clear that given r, δ, and $l_{\min}$ there exists a constant c such that the above is bounded by $\frac{1}{n^c}$. Therefore, the result holds with high probability. □

A.6 Proof of Theorem 4.5

Proof. First, given an instance $\mathcal{I}_{\mathrm{PFC}}$ with optimal value $\mathrm{OPT}_{\mathrm{PFC}}$ the clusters in the optimal solution would with high probability be a valid solution for the deterministic setting, as showed in Theorem 4.4. Moreover the objective value of the solution is unchanged. Therefore, the resulting deterministic instance would have $\mathrm{OPT}_{\mathrm{DFC_{LB}}} \leq \mathrm{OPT}_{\mathrm{PFC}}$. Hence, the algorithm will return a solution with cost at most $(\alpha+2)\,\mathrm{OPT}_{\mathrm{DFC_{LB}}} \leq (\alpha+2)\,\mathrm{OPT}_{\mathrm{PFC}}$.

For the solution $\mathrm{SOL}_{\mathrm{DFC_{LB}}}$ returned by the algorithm, each cluster is of size at least L, and the Chernoff bound guarantees that the violation in expectation is at most ϵ with high probability. □

B Further details on Independent Sampling and Large Cluster Solution

Here we introduce more details about independent sampling. In section B.1 we discuss the concentration bounds associated with the algorithm. In section B.2 we show that relaxing the upper and lower bounds might be necessary for the algorithm to have a high probability of success. Finally, in section B.3 we show that not enforcing a lower bound when solving the deterministic fair instance may lead to invalid solutions.

B.1 Independent Sampling and the Resulting Concentration Bounds

We recall the Chernoff bound theorem for the sum of a collection of independent random variables.

Theorem B.1. *Given a collection of n many binary random variables where $\Pr[X_j = 1] = p_j$ and $S = \sum_{j=1}^n X_j$. Then $\mu = \mathbb{E}[S] = \sum_{j=1}^n p_j$ and the following concentration bound holds for $\delta \in (0,1)$:*

$$\Pr(|S - \mu| > \delta\mu) \leq 2e^{-\mu\delta^2/3} \tag{9}$$

In the following theorem we show that although we do not know the true joint probability distribution $\mathcal{D}_{\mathrm{True}}$, sampling according to the marginal probability $p_v^{h_\ell}$ for each point v results in the amount of color having the same expectation for any collection of points. But furthermore, the amount of color would have a Chernoff bound for the independently sampled case.

Theorem B.2. *Let $\Pr_{\mathcal{D}_{\mathrm{True}}}[X_1 = x_1, \ldots, X_n = x_n]$ equal the probability that $(X_1 = x_1, \ldots, X_n = x_n)$ where X_i is the random variable for the color of vertex i and $x_i \in \mathcal{H}$ ($\mathcal{H}$ being the set of colors) is a specific value for the realization and the probability is according to the true unknown joint probability distribution $\mathcal{D}_{\mathrm{True}}$. Using $X_i^{h_\ell}$ for the indicator random variable of color h_ℓ for vertex i, then for any collection of points C, the amount of color h_ℓ in the collection is $S_{\mathcal{D}_{\mathrm{True}}}^{h_\ell} = \sum_{i\in C} X_{i,\mathcal{D}_{\mathrm{True}}}^{h_\ell}$ when sampling according to $\mathcal{D}_{\mathrm{True}}$ and it is $S_{\mathcal{D}_{\mathrm{Indep}}}^{h_\ell} = \sum_{i\in C} X_{i,\mathcal{D}_{\mathrm{Indep}}}^{h_\ell}$ when independent sampling is done. We have that:*

- *In general:* $\Pr_{\mathcal{D}_{\text{True}}}[X_1 = x_1, \ldots, X_n = x_n] \neq \Pr_{\mathcal{D}_{\text{Indep}}}[X_1 = x_1, \ldots, X_n = x_n]$.
- *Expectations agree on the of amount of color:* $\mathbb{E}[S^{h_\ell}_{\mathcal{D}_{\text{True}}}] = \mathbb{E}[S^{h_\ell}_{\mathcal{D}_{\text{Indep}}}]$.
- *The amount of color has a Chernoff bound for the independently sampled case* $S^{h_\ell}_{\mathcal{D}_{\text{Indep}}}$.

Proof. The first point follows since we simply don't have the same probability distribution. The second is immediate from the linearity of expectations and the fact that both distributions agree in the marginal probabilities ($\Pr_{\mathcal{D}_{\text{True}}}[X_i = h_\ell] = \Pr_{\mathcal{D}_{\text{Indep}}}[X_i = h_\ell] = p_i^{h_\ell}$):

$$\begin{aligned}\mathbb{E}[S^{h_\ell}_{\mathcal{D}_{\text{Indep}}}] &= \mathbb{E}\left[\sum_{i \in C} X^{h_\ell}_{i,\mathcal{D}_{\text{Indep}}}\right] = \sum_{i \in C} \mathbb{E}\left[X^{h_\ell}_{i,\mathcal{D}_{\text{Indep}}}\right] \\ &= \sum_{i \in C} p_i^{h_\ell} = \sum_{i \in C} \mathbb{E}\left[X^{h_\ell}_{i,\mathcal{D}_{\text{True}}}\right] = \mathbb{E}[S^{h_\ell}_{\mathcal{D}_{\text{True}}}]\end{aligned}$$

The last point follows from the fact that $S^{h_\ell}_{\mathcal{D}_{\text{Indep}}}$ is a sum of independent random variables and therefore the Chernoff bound has to hold (B.1). □

B.2 Relaxing the Upper and Lower Bounds

Suppose for an instance $\mathcal{I}_{\text{PFC}}$ of probabilistic fair clustering that there exists a color h_ℓ for which the the upper and lower proportion bounds are equal, i.e. $l_{h_\ell} = u_{h_\ell}$. Suppose the optimal solution $SOL_{\text{PFC}} = (S^*_{\text{PFC}}, \phi^*_{\text{PFC}})$, has a cluster C_i which we assume can be made arbitrarily away than the other points. The Chernoff bound guaranteed by independent sampling would not be useful since the realization has to precisely equal the expectation, not be within a δ of the expectation. In this case sampling will not result in cluster C_i having a balanced color and therefore the points in C_i would have to merged with other points (if possible, since the entire instance maybe infeasible) to have a cluster with balance equal to l_{h_ℓ} and u_{h_ℓ} for color h_ℓ. Since we assumed cluster C_i can be made arbitrarily far away the cost of deterministic instance generated can be arbitrarily worse.

Note, that we do not really need $l_{h_\ell} = u_{h_\ell}$. Similar arguments can be applied if $l_{h_\ell} \neq u_{h_\ell}$, by assuming the that optimal solution has a cluster C_i (which is arbitrarily far away) whose balance either precisely equals l_{h_ℓ} or u_{h_ℓ}. Simply note that with independent sampling would result in violation to the bounds for cluster C_i.

Therefore, in the worst case relaxing the bounds is necessary to make sure that a valid solution would remain valid w.h.p. in the deterministic instance generated by independent sampling.

B.3 Independent Sampling without Lower Bounded Cluster Sizes Could Generate Invalid Solutions

To show that enforcing a lower bound on the cluster size is necessary, consider the case shown in figure 6:(a) where the outlier points in the top-right have probability 0.45 of being white, whereas the other points have probability 1 of being white. Let the lower and upper bounds for the white color be $l_{\text{white}} = 0.6$ and $u_{\text{white}} = 1$, respectively. Since the outlier points don't have the right color balance, they are merged with the other points, although that leads to a higher cost.

However, independent sampling would result in the outlier points being white with probability $(0.45)(0.45) \simeq 0.2$. This makes the points have the right color balance and therefore the optimal solution for deterministic fair clustering would have these points merged as shown in figure 6:(b). However, the cluster for the two outlier points is not a valid cluster for the probabilistic fair clustering instance

Therefore, forcing a lower bound is necessary to make sure that a solution found in deterministic fair clustering instance generated by independent sampling is w.h.p. valid for the probabilistic fair clustering instance.

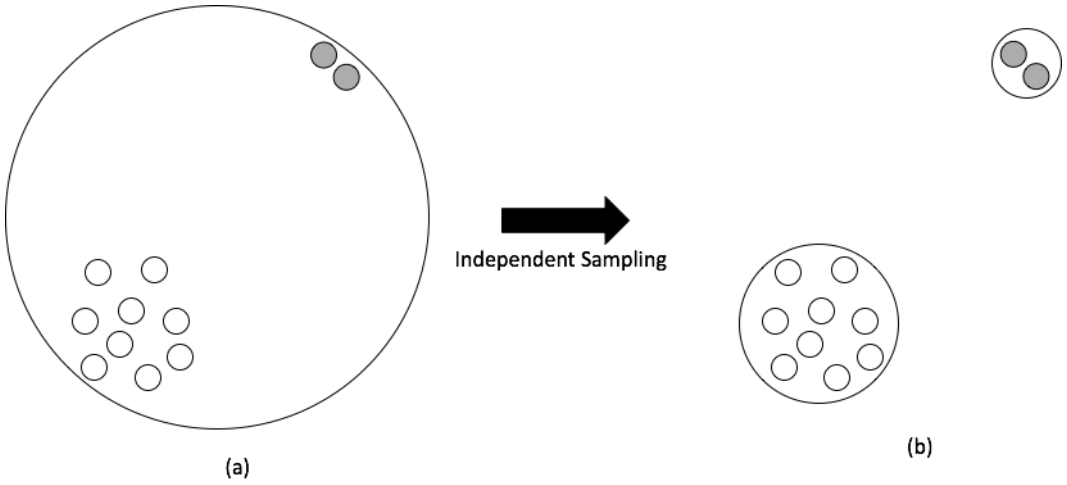

Figure 6: (a): The two outlier points at the top-right have probabilities 0.45 of being white, whereas the rest have probabilities 1. All points are merged together to form a balanced cluster. (b): An instance of same points with the colors resulting from independent sampling. The two outlier points have been merged to form their own cluster.

C Example on Forming the Network Flow Graph for the Two-Color (Metric Membership) Case

Suppose we have two centers and 5 vertices and that the LP solution yields the following assignments for center 1: $x_{11} = 0.3, x_{12} = 0.6, x_{13} = 0.7, x_{14} = 0, x_{15} = 1.0$ and the following assignments for center 2: $x_{21} = 0.7, x_{22} = 0.4, x_{23} = 0.3, x_{24} = 1.0, x_{25} = 0$. Further let the probability values be: $p_1 = 0.7, p_2 = 0.8, p_3 = 0.4, p_4 = 0.9, p_5 = 0.1$. The following explains how the network flow graph is constructed.

Cluster 1: First we calculate $|C_1| = \left\lceil \sum_{j \in \mathcal{C}} x_{1j} \right\rceil = \lceil 2.6 \rceil = 3$, this means the we will have 3 vertices in C_1. The collection of vertices having non-zero assignments to center 1 are $\{1, 2, 3, 5\}$, sorting the vertices by a non-increasing order according to their probability we get $\vec{A}_1 = [2, 1, 3, 5]$. Now we follow algorithm 1, this leads to the graph shown in figure 7.

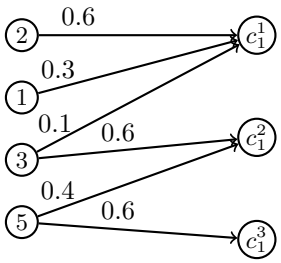

Figure 7: Graph constructed in cluster 1. For clarity, we write above each edge the assignment is "sends" to the vertex in C_1. Notice how each vertex in C_1 receives a total assignment of 1, except for the last vertex c_1^3.

Cluster 2: We follow the same procedure for cluster 2. First we calculate $|C_2| = \left\lceil \sum_{j \in \mathcal{C}} x_{1j} \right\rceil = \lceil 2.4 \rceil = 3$, this means the we will have 3 vertices in C_2. The collection of vertices having non-zero assignments to center 2 are $\{1, 2, 3, 4\}$, sorting the vertices by a non-increasing order according to

their probability we get $\vec{A}_2 = [4, 2, 1, 3]$. Now we follow algorithm 1, this leads to the graph shown in figure 8 Now we construct the entire graph by connecting the edges from each vertex in C_1 to the

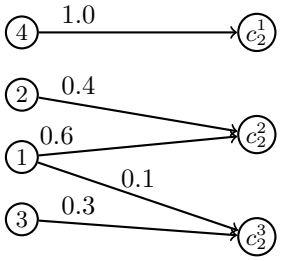

Figure 8: Graph constructed in cluster 2. For clarity, we write above each edge the assignment is "sends" to the vertex in C_2. Notice how each vertex in C_2 receives a total assignment of 1, except for the last vertex c_2^3.

vertex for center 1 and each vertex in C_2 to the vertex for center 2. Finally, we connect the vertices for 1 and 2 to the vertex t. This leads to the graph in figure 9. Note that the edge weights showing the sent assignment are not put as they have no significance once the graph is constructed.

The entire graph is constructed by the union of both subgraphs for clusters 1 and 2, but without repeating the vertices of $\mathcal{C}$. Further, we drop the wedge weights which designated the amount of LP assignment sent, as it has no affect on the following steps. Finally, the vertices of both C_1 and C_2 are connected to their centers 1 and 2 in S, respectively, and the centers themsevles are connected to vertex t. Figure 9 shows the final constructed graph.

For the case of metric membership the procedure is unaltered, but instead of sorting according to the probability value p_v for a vertex, we sort according to the value r_v.

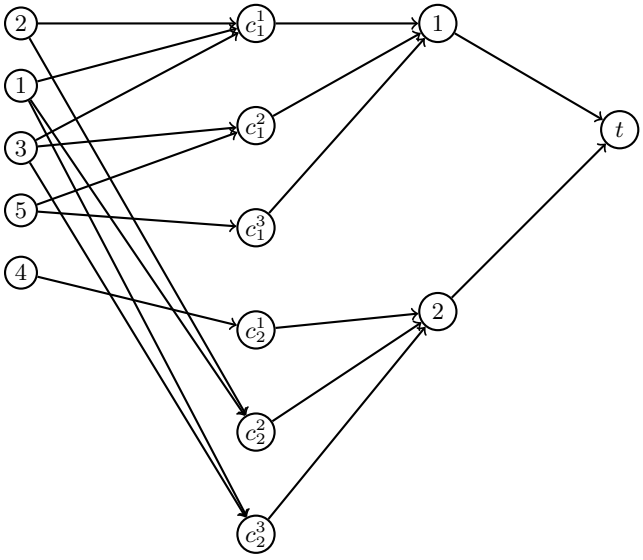

Figure 9: Diagram for the final network flow graph.

D Dependent Rounding for Multiple Colors under a Large Cluster Assumption

Here we discuss a dependent rounding based solution for the k-center problem under the large cluster assumption 4.1. First we start with a brief review/introduction of dependent rounding.

D.1 Brief Summary of Dependent Rounding

Here we summarize the properties of dependent rounding, see Gandhi et al. [2006] for full details. Given a bipartite graph $(G = (A, B), E)$ each edge $(i, j) \in E$ has a value $0 \leq x_{ij} \leq 1$ which will be rounded to $X_{ij} \in \{0, 1\}$. Further for every vertex $v \in A \cup B$ define the fractional degree as $d_v = \sum_{u:(v,u)\in E} x_{vu}$ and the integral degree as $D_v = \sum_{u:(v,u)\in E} X_{vu}$. Dependent rounding satisfies the following properties:

1. $\Pr[X_{ij} = 1] = x_{ij}$.
2. $\forall v \in A \cup B : D_v \in \{\lfloor d_v \rfloor, \lceil d_v \rceil\}$
3. $\forall v \in A \cup B$, let E_v denote any subset of edges incident on v, then $\Pr[\bigwedge_{e_v \in E_v} X_{e_v} = b] \leq \Pi_{e_v \in E_v} \Pr[X_{e_v} = b]$ where $b \in \{0, 1\}$.

We note that property 3 implies the following theorem about the variables X_{ij} (see theorem 3.1 in Gandhi et al. [2006]):

Theorem D.1. *Let $a_1, \ldots, a_t$ be reals in $[0, 1]$, and $X_1, \ldots, X_t$ be random variables taking values in $\{0, 1\}$, and $\mathbb{E}[\sum_i a_i X_i] = \mu$. If $\Pr[\bigwedge_{i\in S} X_i = b] \leq \Pi_{i\in S} \Pr[X_i = b]$ where S is any subset of indices from $\{1, \ldots, t\}$ and $b \in \{0, 1\}$, then we have for $\delta \in (0, 1)$:*

$$\Pr\left[|\sum_i a_i X_i - \mu| \geq \delta\mu\right] \leq 2e^{-\mu\delta^2/3}$$

D.2 Multiple Color Large Cluster solution using Dependent Rounding

For the multiple color k-center problem satisfying assumption 4.1. Form the following bipartite graph $(G = (A, B), E)$, A has all vertices of of $\mathcal{C}$, and B has all of the vertices of S (the cluster centers). Further the fractional assignments x_{ij} represent the weight of the edge, $\forall (i, j) \in E$. Applying dependent rounding leads to the following theorem:

Theorem D.2. *Under assumption 4.1, the integer solution resulting from dependent rounding for the multi-color probabilistic k-center problem has: (1) An approximation ratio of $\alpha + 2$. (2) For any color h_ℓ and any cluster $i \in S$, the amount of color $S^{h_\ell}_{C_i} = \sum_{j\in\mathcal{C}} p_j^{h_\ell} X_{ij}$ is concentrated around the LP assigned color $\sum_{j\in\mathcal{C}} p_j^{h_\ell} x_{ij}$.*

Proof. For (1): Note that the approximation ratio before applying dependent rounding is $\alpha + 2$. By property 1 of dependent rounding if $x_{ij} = 0$, then $\Pr[X_{ij} = 1] = 0$ and therefore a point will not be assigned to a center it was not already assigned to by the LP.

For (2): Again by property 1 of dependent rounding $\mathbb{E}_{DR}[X_{ij}] = (1)x_{ij} + 0 = x_{ij}$ where $\mathbb{E}_{DR}$ refers to the expectation with respect to the randomness of dependent rounding, therefore for any cluster i the expected amount of color equals the amount of color assigned by the LP, i.e. $\mathbb{E}_{DR}[S^{h_\ell}_{C_i}] = \mathbb{E}_{DR}[\sum_{j\in\mathcal{C}} p_j^{h_\ell} X_{ij}] = \sum_{j\in\mathcal{C}} p_j^{h_\ell} \mathbb{E}_{DR}[X_{ij}] = \sum_{j\in\mathcal{C}} p_j^{h_\ell} x_{ij}$. It follows by property 3 of dependent rounding and theorem D.1 that $S^{h_\ell}_{C_i}$ is highly concentrated around $\mathbb{E}_{DR}[S^{h_\ell}_{C_i}]$. Specifically :

$$\Pr\left[|S^{h_\ell}_{C_i} - \mathbb{E}_{DR}[S^{h_\ell}_{C_i}]| \geq \delta\, \mathbb{E}_{DR}[S^{h_\ell}_{C_i}]\right] \leq 2e^{-\mathbb{E}_{DR}[S^{h_\ell}_{C_i}]\delta^2/3}$$

Similar to the proof of 4.4, the probability of failure can be upper bounded by:

$$\begin{aligned}
&\Pr\left(\left\{\exists i \in \{1, \ldots, k\}, h_\ell \in \mathcal{H} \,||S^{h_\ell}_{C_i} - \mathbb{E}[S^{h_\ell}_{C_i}]| > \delta\, \mathbb{E}[S^{h_\ell}_{C_i}]\right\}\right) \\
&\leq 2k|\mathcal{H}| \exp(-\frac{\delta^2}{3} L l_{\min}) \leq 2\frac{n}{L}|\mathcal{H}| \exp(-\frac{\delta^2}{3} L l_{\min}) \\
&\leq 2|\mathcal{H}| n^{1-r} \exp(-\frac{\delta^2}{3} l_{\min} n^r)
\end{aligned}$$

Therefore w.h.p the returned integral solution will be concentrated around the LP color assignments which are fair. □

E Further details on solving the lower bounded fair clustering problem

The solution for the lower bounded deterministic fair clustering problem, follows a similar two step solution framework. Step (1) is unchanged and simply amounts to running a color-blind approximation algorithm with ratio α. Step (2) sets up an LP similar to that in section 4.1.2. The constraints in 7c still remain but with deterministic (not probabilistic) color assignments, further a new constraint lower bounding the cluster size is added. Specifically, we have the following LP:

$$
\begin{aligned}
\min \sum_{v \in \mathcal{C}, i \in S} d^p(v,i) \quad \text{s.t.} & \\
l_{h_\ell} \sum_{v \in \mathcal{C}} x_{ij} \leq \sum_{v \in \mathcal{C}: \chi(v) = h_\ell} x_{ij} \quad & ,\forall i \in S, \forall\, h_\ell \in \mathcal{H} && (10)\\
\sum_{v \in \mathcal{C}: \chi(v) = h_\ell} x_{ij} \leq u_{h_\ell} \sum_{v \in \mathcal{C}} x_{ij} \quad & ,\forall i \in S, \forall\, h_\ell \in \mathcal{H} && (11)\\
\sum_{j \in \mathcal{C}} x_{ij} \geq L & ,\forall i \in S && (12)\\
\sum_{j \in \mathcal{C}} x_{ij} = 1 & ,\forall j \in \mathcal{C} &&\\
0 \leq x_{ij} \leq 1 & ,\forall i \in S, \forall j \in \mathcal{C} &&
\end{aligned}
$$

Constraints 10 and 11 are the deterministic counterparts to constraints 7c, respectively. Constraint 12 is introduced to lower bound the cluster size. The resulting solution will have an approximation ratio of $\alpha + 2$ (see A.2).

What remains is to round the solution. We apply the network flow rounding from Bercea et al. [2018] (specifically section 2.2 in Bercea et al. [2018]). This results in a violation of at most 1 in the cluster size and a violation of at most 1 per color in any give cluster (lemma 8 in Bercea et al. [2018]).

F Further Experimental Details and Results

F.1 Further Experiments for the two color case

How does labeling accuracy level p_{acc} impact this problem? Fig. 10 shows p_{acc} vs POF for $\delta = 0.2$ and $\delta = 0.1$. At $p_{\text{acc}} = \frac{1}{2}$, color assignments are completely random and the cost is, as expected, identical to color-blind cost. As p_{acc} increases, the colors of the vertices become more differentiated, causing POF to increase, eventually reaching the maximum at $p_{\text{acc}} = 1$ which is the deterministic case.

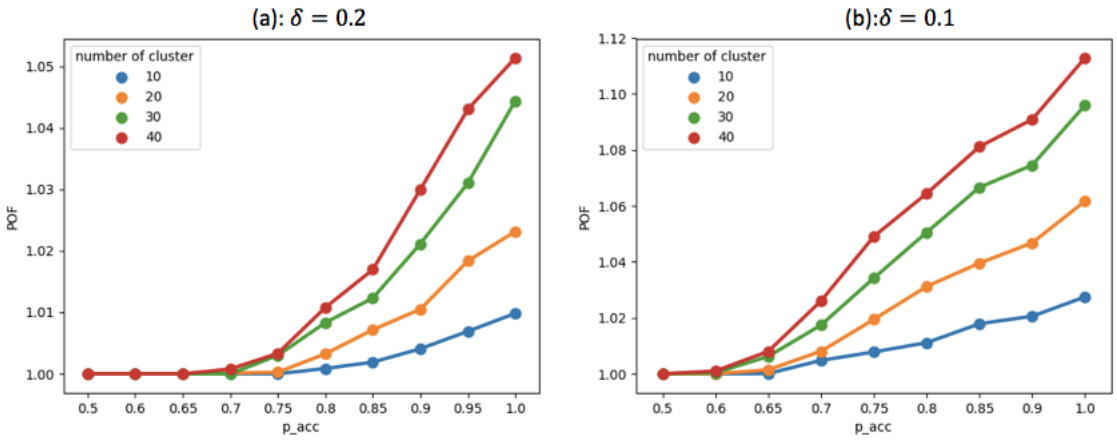

Figure 10: Plot showing p_{acc} vs POF, (a):$\delta = 0.2$ and (b):$\delta = 0.1$.

Next, we test against an "obvious" strategy when faced with probabilistic color labels: simply *threshold* the probability values, and then run a deterministic fair clustering algorithm. Fig. 11(a) shows that this may indeed work for guaranteeing fairness, as the proportions may be satisfied with small violations; however, it comes at the expense of a much higher POF. Fig. 11(b) supports this latter statement: our algorithm can achieve the same violations with smaller POF. Further, running a deterministic algorithm over the thresholded instance may result in an infeasible problem.[3]

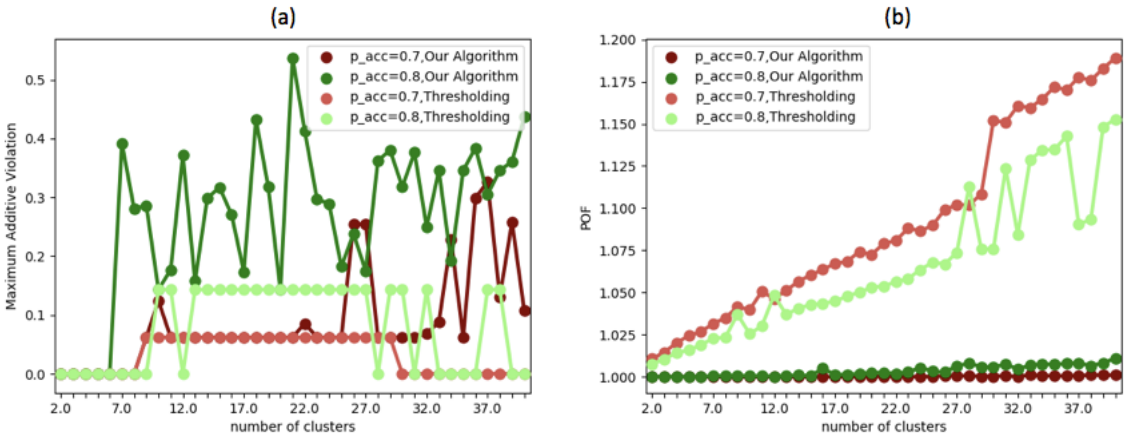

Figure 11: Comparing our algorithm to thresholding followed by deterministic fair clustering: (a)maximum violation, (b) POF.

F.2 Further Details about the Datasets and the Experimental Setup

For each dataset, the numeric features are used as coordinates and the distance between points is equal to Euclidean distance. The numeric features are normalized prior to clustering.

For metric membership in the **Adult** dataset, age is not used as a coordinate despite the fact that it is numeric since it is the fairness attribute. Similarly, for the **CreditCard** dataset, credit is not used as a coordinate.

When solving the min-cost flow problem, distances are first multiplied by a large number (1000) and then rounded to integer values. After obtaining the solution for the flow problem, the cost is calculated with the original distance values (which have not been rounded) to verify that the cost is not worse.

Although run-time is not a main concern in this paper. We find that we can solve large instances containing 100,000 points for the k-means with 5 clusters in less than 4 minutes using our commodity hardware.

F.3 Further Experiments

Here we verify the performance of our algorithm on the k-center and the k-median objectives. All datasets have been sub-sampled to 1,000 data points. For the two color probabilistic case, throughout we set $p_{\text{acc}} = 0.9$ (see section 5.2 for the definition of p_{acc}).

F.3.1 k-center

As can be seen from figure 12 our violations are indeed less than 1 matching the theoretical guarantee. Similarly, for metric membership the normalized violation is less than 1 as well, see figure 13.

[3] An intuitive example of infeasibility: consider the two color case where $p_v = \frac{1}{2} + \epsilon, \forall\, v \in \mathcal{C}$ for some small positive ϵ. Thresholding drastically changes the overall probability to 1; therefore no subset of points would have proportion around $\frac{1}{2} + \epsilon$.

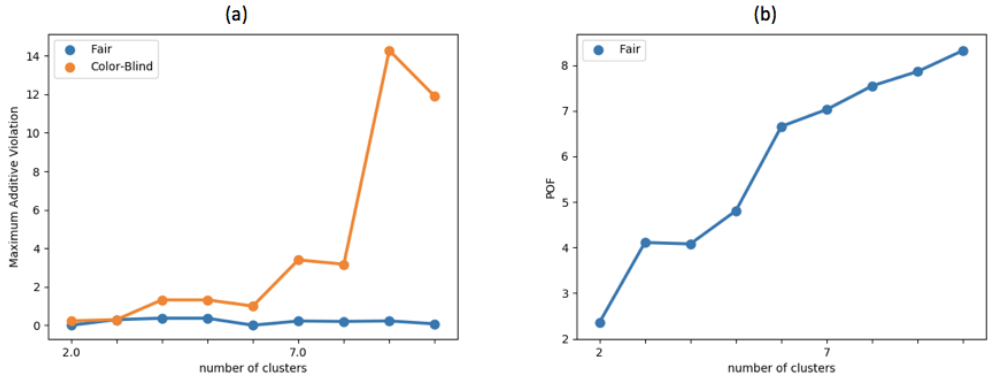

Figure 12: k-center for the two color probabilistic case using the **Bank** dataset. (a): number of clusters vs maximum violation, (b): number of clusters vs POF.

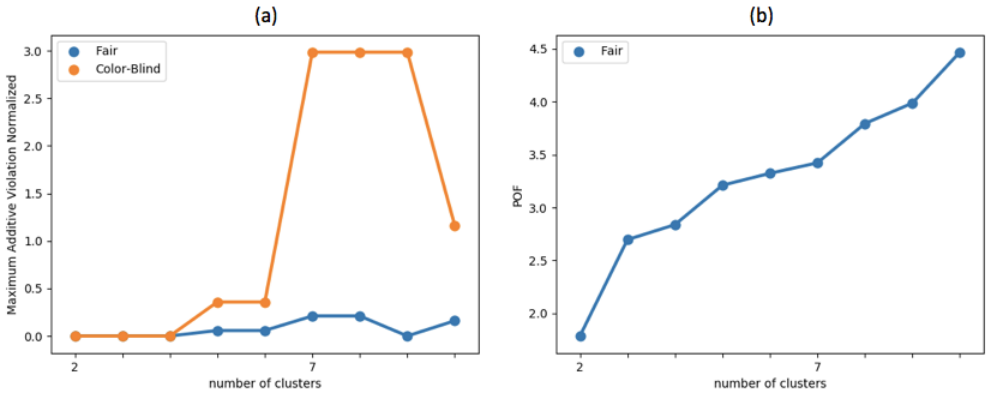

Figure 13: k-center for the metric membership problem using the **Adult** dataset (metric membership over age). (a): number of clusters vs normalized maximum violation, (b): number of clusters vs POF.

F.3.2 k-median

Similar observations apply to the k-median problems. That is, our algorithm indeed leads to small violations not exceeding 1 in keeping with the theory. See figure 14 for the two color probabilistic case and figure 15 for the metric membership case.

F.3.3 Further Experiments on the Census1990 Dataset with the Large Cluster Assumption

We ran the large cluster experiment on **Census1990** dataset the for different values of k; Fig. 16 shows a reasonable (and expected) degradation in quality.

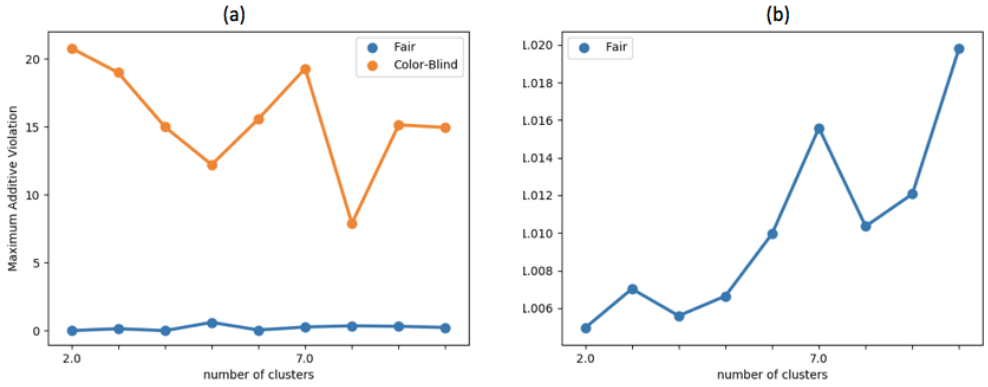

Figure 14: k-median for the two color probabilistic case using the **Bank** dataset. (a): number of clusters vs maximum violation, (b): number of clusters vs POF.

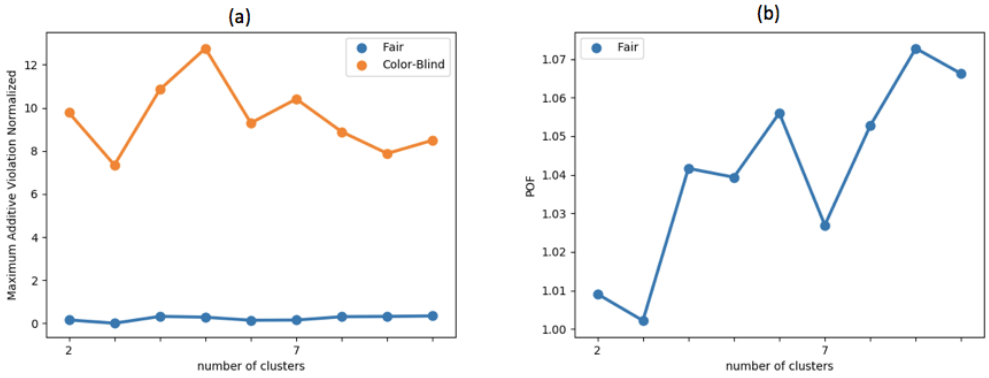

Figure 15: k-median for the metric membership problem using the **CreditCard** dataset (metric membership over credit) (a): number of clusters vs normalized maximum violation, (b): number of clusters vs POF.

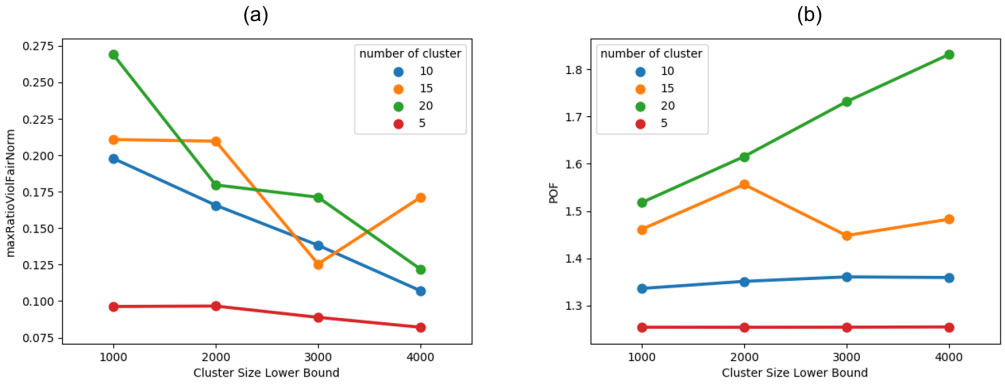

Figure 16: Results on the **Census1990** dataset for different values of k. We see a reasonable degradation in the violation (a) and POF (b) for larger values of k.