# A Useful Facts and Lemmas

**Fact A.1.** *Given non-negative numbers $a_1, a_2, \ldots, a_n$ and positive numbers $b_1, b_2, \ldots, b_n$, then:*

$$\min_{i\in[n]} \frac{a_i}{b_i} \leq \frac{\sum_{i\in[n]} a_i}{\sum_{i\in[n]} b_i} \leq \max_{i\in[n]} \frac{a_i}{b_i}$$

# B MaxFlowGF

We start with the following Lemma. First, note that $x_{qj}$ is a decision variable if $x_{qj} = 1$ then point $j$ is assigned to center $q$ and if $x_{qj} = 0$ then it is not. In an integral solutions $x_{qj} \in \{0, 1\}$, but a fractional LP solution could instead have values in $[0, 1]$. We use the bold symbol $\mathbf{x}$ for the collection of value $\{x_{qj}\}_{q\in Q, j\in C}$:

**Lemma 3.** *Given a fractional solution $\mathbf{x}^{\textit{frac}}$ that satisfies the* ***GF*** *constraints at an additive violation of at most $\rho$, then if there exists an integral solution $\mathbf{x}^{\textit{integ}}$ that satisfies:*

$$\forall q \in Q : \left\lfloor \sum_{j\in C} x^{\textit{frac}}_{qj} \right\rfloor \leq \sum_{j\in C} x^{\textit{integ}}_{qj} \leq \left\lceil \sum_{j\in C} x^{\textit{frac}}_{qj} \right\rceil \tag{3}$$

$$\forall q \in Q, h \in \mathcal{H} : \left\lfloor \sum_{j\in C^h} x^{\textit{frac}}_{qj} \right\rfloor \leq \sum_{j\in C^h} x^{\textit{integ}}_{qj} \leq \left\lceil \sum_{j\in C^h} x^{\textit{frac}}_{qj} \right\rceil \tag{4}$$

*Then this integral solution $\mathbf{x}^{\textit{integ}}$ satisifies the* ***GF*** *constraints at an additive violation of at most $\rho + 2$.*

*Proof.* Since the fractional solution satisfies the **GF** constraints at an additive violation of $\rho$, then we have the following:

$$-\rho + \Big(\beta_h \sum_{j\in C} x^{\text{frac}}_{qj}\Big) \leq \sum_{j\in C^h} x^{\text{frac}}_{qj} \leq \Big(\alpha_h \sum_{j\in C} x^{\text{frac}}_{qj}\Big) + \rho$$

We start with the upper bound:

$$\begin{aligned}
\sum_{j\in C^h} x^{\text{integ}}_{qj} &\leq \left\lceil \sum_{j\in C^h} x^{\text{frac}}_{qj} \right\rceil \\
&\leq \sum_{j\in C^h} x^{\text{frac}}_{qj} + 1 \\
&\leq \alpha_h \sum_{j\in C} x^{\text{frac}}_{qj} + \rho + 1 \\
&\leq \alpha_h (\sum_{j\in C} x^{\text{integ}}_{qj} + 1) + \rho + 1 \\
&\leq \alpha_h \sum_{j\in C} x^{\text{integ}}_{qj} + (\alpha_h + \rho + 1) \\
&\leq \alpha_h \sum_{j\in C} x^{\text{integ}}_{qj} + (\rho + 2)
\end{aligned}$$

Now we do the lower bound:

$$
\begin{aligned}
\sum_{j\in C^h} x^{\text{integ}}_{qj} &\geq \left\lfloor \sum_{j\in C^h} x^{\text{frac}}_{qj} \right\rfloor \\
&\geq \sum_{j\in C^h} x^{\text{frac}}_{qj} - 1 \\
&\geq \beta_h \sum_{j\in C} x^{\text{frac}}_{qj} - \rho - 1 \\
&\geq \beta_h (\sum_{j\in C} x^{\text{integ}}_{qj} - 1) - (\rho + 1) \\
&\geq \beta_h \sum_{j\in C} x^{\text{integ}}_{qj} - (\beta_h + \rho + 1) \\
&\geq \beta_h \sum_{j\in C} x^{\text{integ}}_{qj} - (\rho + 2)
\end{aligned}
$$

□

The LP solution given to MAXFLOWGF satisfies the **GF** constraints at an additive violation of $\rho$, we want to show that the output integral solution satisfies the above conditions of Eqs (3 and 4). The MAXFLOWGF($\mathbf{x}^{\text{LP}}, C, Q$) subroutine is similar to that shown in [20, 5, 10]. Specifically, given an LP solution $\mathbf{x}^{\text{LP}} = \{x^{\text{LP}}\}_{q\in Q, j\in Q}$, a set of points $C$, and a set of centers $Q$ and a color assignment function $\chi : C \rightarrow \mathcal{H}$ which assigns to each point in $C$ exactly one color in the set of colors $\mathcal{H}$, we construct the flow network $(V, A)$ according to the following:

1. $V = \{s, t\} \cup C \cup \{(q, q^h)|q \in Q, h \in \mathcal{H}\}$.
2. $A = A_1 \cup A_2 \cup A_3 \cup A_4$ where $A_1 = \{(s, j)|j \in C\}$ with upper bound of 1. $A_2 = \{(j, (q, q^h))|j \in C, x_{qj} > 0\}$ with upper bound of 1. The arc set $A_3 = \{((q, q^h), q)|q \in Q, h \in \mathcal{H}\}$ with lower bound $\left\lfloor \sum_{j\in C^h} x^{\text{LP}}_{qj} \right\rfloor$ and upper bound of $\left\lceil \sum_{j\in C^h} x^{\text{LP}}_{qj} \right\rceil$. As for $A_4 = \{(q, t)|q \in Q\}$ the lower and upper bounds are $\left\lfloor \sum_{j\in C} x^{\text{LP}}_{qj} \right\rfloor$ and $\left\lceil \sum_{j\in C} x^{\text{LP}}_{qj} \right\rceil$.

In the above all lower and upper bounds of the network are integral, therefore if we can show a feasible solution to the above then there must exist an integral flow assignment which also satisfies the constraints. By the construction of the network we have the following fact about any max flow integral solution $\mathbf{x}^{\text{integ}}$.

**Fact B.1.**

$$\forall q \in Q : \left\lfloor \sum_{j\in C} x^{LP}_{qj} \right\rfloor \leq \sum_{j\in C} x^{integ}_{qj} \leq \left\lceil \sum_{j\in C} x^{LP}_{qj} \right\rceil \tag{5}$$

$$\forall q \in Q, h \in \mathcal{H} : \left\lfloor \sum_{j\in C^h} x^{LP}_{qj} \right\rfloor \leq \sum_{j\in C^h} x^{integ}_{qj} \leq \left\lceil \sum_{j\in C^h} x^{LP}_{qj} \right\rceil \tag{6}$$

Accordingly, the following theorem immediately holds:

**Theorem B.1.** *Given an LP solution to* MAXFLOWGF *that satisfies the* ***GF*** *constraints at an additive violation of $\rho$ and a clustering cost of $R$, then the output integral solution satisfies the* ***GF*** *constraints at an additive violation of $\rho + 2$ and a clustering cost of at most $R$.*

*Proof.* The guarantee for the additive violation of **GF** follows immediately from Lemma 3 and Fact B.1. The guarantee for the clustering cost holds, since a point (vertex) $j$ is not connected to a center vertex $(q, q^h)$ unless $x_{qj} > 0$ which can only be the case if $d(j, q) \leq R$. □

# C OMITTED PROOFS

We restate the following lemma and give its proof:

**Lemma 1.** *Given a non-empty cluster $C$ with center $i$ and radius $R$ that satisfies the **GF** constraints at an additive violation of $\rho$ and a subset of points $Q$ ($Q \subset C$). Then the clustering $(Q, \phi)$ where $\phi = \text{DIVIDE}(C, Q)$ has the following properties: (1) The **GF** constraints are satisfied at an additive violation of at most $\frac{\rho}{|Q|} + 2$. (2) Every center in $Q$ is* active. *(3) The clustering cost is at most $2R$. If $|Q| = 1$ then guarantee (1) is for the additive violation is at most $\rho$.*

*Proof.* We first consider the case where $|Q| > 1$. We prove the following claim[2]:

**Claim 1.** *For the fractional assignment $\{x^{\textit{frac}}_{qj}\}_{q\in Q, j\in C}$ such that:*

$$\forall q \in Q, \forall h \in \mathcal{H} : \sum_{j\in C^h} x^{\textit{frac}}_{qj} = \frac{|C^h|}{|Q|} = T_h$$

*It holds that: (1) $\forall q \in Q : \sum_{j\in C} x^{\textit{frac}}_{qj} \geq 1$, (2) **GF** constraints are satisfied at an additive violation of $\frac{\rho}{|Q|}$.*

*Proof.* Now we prove the first property

$$\forall q \in Q : \sum_{j\in C} x^{\text{frac}}_{qj} = \sum_{h\in\mathcal{H}} \sum_{j\in C^h} x^{\text{frac}}_{qj} = \frac{1}{|Q|} \sum_{h\in\mathcal{H}} |C^h| = \frac{|C|}{|Q|} \geq 1 \quad (\text{since } Q \subset C) \tag{7}$$

.

Since the **GF** constraints given center $i$ are satisfied at an additive violation of $\rho$, then we have:

$$\forall h \in \mathcal{H} : -\rho + \beta_h |C| \leq |C^h| \leq \alpha_h |C| + \rho \tag{8}$$

Therefore, since the amount of color for each center in $Q$ with the fractional assignment can be obtained by dividing by $|Q|$, then we have:

$$\forall h \in \mathcal{H}, \forall q \in Q : -\frac{\rho}{|Q|} + \beta_h \sum_{j\in C} x^{\text{frac}}_{qj} \leq \sum_{j\in C^h} x^{\text{frac}}_{qj} \leq \alpha_h \sum_{j\in C} x^{\text{frac}}_{qj} + \frac{\rho}{|Q|} \tag{9}$$

Therefore the **GF** constraints are satisfied at an additive violation of $\frac{\rho}{|Q|}$. ☐

Denoting the assignment $\phi$ resulting from DIVIDE by $\{x^{\text{integ}}_{qj}\}_{q\in Q, j\in C}$, then the following claim holds:

**Claim 2.**

$$\forall q \in Q : \left\lfloor \sum_{j\in C} x^{\textit{frac}}_{qj} \right\rfloor \leq \sum_{j\in C} x^{\textit{integ}}_{qj} \leq \left\lceil \sum_{j\in C} x^{\textit{frac}}_{qj} \right\rceil$$

$$\forall q \in Q, h \in \mathcal{H} : \left\lfloor \sum_{j\in C^h} x^{\textit{frac}}_{qj} \right\rfloor \leq \sum_{j\in C^h} x^{\textit{integ}}_{qj} \leq \left\lceil \sum_{j\in C^h} x^{\textit{frac}}_{qj} \right\rceil$$

*Proof.* For any color $h$ we have $|C_h| = a_h|Q| + b_h$ where $a_h$ and $b_h$ are non-negative integers and $b_h$ is the remainder of dividing $|C_h|$ by $Q$ ($b_h \in \{0, 1, \ldots, |Q| - 1\}$). It follows that $\sum_{j\in C^h} x^{\text{frac}}_{qj} = T_h = a_h + \frac{b_h}{|Q|}$. DIVIDE gives each center either $\sum_{j\in C^h} x^{\text{integ}}_{qj} = a_h = \lfloor T_h \rfloor = \left\lfloor \sum_{j\in C^h} x^{\text{frac}}_{qj} \right\rfloor$ or $\sum_{j\in C^h} x^{\text{integ}}_{qj} = a_h + 1 = \lceil T_h \rceil = \left\lceil \sum_{j\in C^h} x^{\text{frac}}_{qj} \right\rceil$. This proves the second condition.

For the first condition, note that $|C| = \sum_{h\in\mathcal{H}} (a_h|Q| + b_h) = (\sum_{h\in\mathcal{H}} a_h)|Q| + a|Q| + b$ where we set $\sum_{h\in\mathcal{H}} b_h = a|Q| + b$ with $a$ and $b$ being non-negative integers. $b$ is the remainder and has values

[2] In our notation $x_{qj} \in [0, 1]$ denotes the assignment of point $j$ to center $q$.

in $\{0, 1, \ldots, |Q|-1\}$. Accordingly, the sum of the remainders across the colors is $a|Q|+b$. Since the remainders are added "successivly" across the centers (see Figure 2) and $a$ is divisible by $|Q|$, then for any center $q \in Q$ either $\sum_{j\in C} x^{\text{integ}}_{qj} = (\sum_{h\in\mathcal{H}} a_h) + a$ or $\sum_{j\in C} x^{\text{integ}}_{qj} = (\sum_{h\in\mathcal{H}} a_h) + a + 1$. Note that $\sum_{j\in C} x^{\text{frac}}_{qj} = \sum_{h\in\mathcal{H}} T_h = (\sum_{h\in\mathcal{H}} a_h) + a + \frac{b}{|Q|}$. Therefore, $\left\lfloor \sum_{j\in C} x^{\text{frac}}_{qj} \right\rfloor = (\sum_{h\in\mathcal{H}} a_h) + a$ and $\left\lceil \sum_{j\in C} x^{\text{frac}}_{qj} \right\rceil = (\sum_{h\in\mathcal{H}} a_h) + a + 1$. This proves, the first condition. $\square$

By Claim 2 and Lemma 3 it follows that for each center $q \in Q$ the assignment $\{x^{\text{integ}}_{qj}\}_{q\in Q, j\in C}$ satisfies the **GF** constraints at an additive violation of $\frac{\rho}{|Q|} + 2$, this proves the first guarantee.

By Claim 2 and guarentee (1) of Claim 1, then $\forall q \in Q : \sum_{j\in C} x^{\text{integ}}_{qj} \geq \left\lfloor \sum_{j\in C} x^{\text{frac}}_{qj} \right\rfloor \geq 1$. Therfeore, every center $q \in Q$ is *active* proving the second guarantee.

Guarantee (3) follows since $\forall j \in C : d(j, \phi(j)) \leq d(j, i) + d(i, \phi(j)) \leq 2R$.

Now if $|Q| = 1$, then guarantee (2) follows since the cluster $C$ is non-empty. Guarantee (3) follows similarly to the above. The additive violation in the **GF** constraint on the other hand is $\rho$ since the single center $Q$ has the exact set of points that were assigned to the original center $i$. $\square$

We restate the next lemma and give its proof:

**Lemma 2.** *Solution $(S', \phi')$ of line (3) in algorithm 2 has the following properties: (1) It satisfies the* ***GF*** *constraint at an additive violation of 2, (2) It has a clustering cost of at most $(1 + \alpha_{\textbf{DS}})R^*_{\textbf{GF+DS}}$ where $R^*_{\textbf{GF+DS}}$ is the optimal clustering cost (radius) of the optimal solution for* ***GF+DS****, (3) The set of centers $S'$ is a subset (possibly proper subset) of the set of centers $\bar{S}$, i.e. $S' \subset S$.*

*Proof.* We begin with the following claim which shows that there exists a solution that only uses centers from $\bar{S}$ to satisfy the **GF** constraints exactly and at a radius of at most $(1 + \alpha_{\textbf{DS}})R^*_{\textbf{GF+DS}}$. Note that this claim has non-constructive proof, i.e. it only proves the existence of such a solution:

**Claim 3.** *Given the set of centers $\bar{S}$ resulting from the $\alpha_{\textbf{DS}}$-approximation algorithm, then there exists an assignment $\phi_0$ from points in $\mathcal{C}$ to centers in $\bar{S}$ such that the following holds: (1) The* ***GF*** *constraint is exactly satisfied (additive violation of 0). (2) The clustering cost is at most $(1 + \alpha_{\textbf{DS}})R^*_{\textbf{GF+DS}}$.*

*Proof.* Let $(S^*_{\textbf{GF+DS}}, {\phi^*}_{\textbf{GF+DS}})$ be an optimal solution to the **GF+DS** problem. $\forall i \in S^*_{\textbf{GF+DS}}$ let $N(i) = \arg\min_{\bar{i}\in\bar{S}} d(i, \bar{i})$, i.e. $N(i)$ is the nearest center in $\bar{S}$ to center $i$ (ties are broken using the smallest index). $\phi_0$ is formed by assigning all points which belong to center $i \in S^*_{\textbf{GF+DS}}$ to $N(i)$. More formally, $\forall j \in \mathcal{C} : {\phi^*}_{\textbf{GF+DS}}(j) = i$ we set $\phi_0(j) = N(i)$. Note that it is possible for more than one center $i$ in $S^*_{\textbf{GF+DS}}$ to have the same nearest center in $\bar{S}$. We will now show that $\phi_0$ satisfies the **GF** constraint exactly. Note first that if a center $\bar{i} \in \bar{S}$ has not been assigned any points by $\phi_0$, then it is empty and trivially satisfies the **GF** constraint exactly. Therefore, we assume that $\bar{i}$ has a non-empty cluster. Denote by $N^{-1}(\bar{i})$ the set of centers $i \in S^*_{\textbf{GF+DS}}$ for which $\bar{i}$ is the nearest center, then using Fact A.1 and the fact that every cluster in $(S^*_{\textbf{GF+DS}}, {\phi^*}_{\textbf{GF+DS}})$ satisfies the **GF** constraint exactly we have:

$$\beta_h \leq \min_{i\in N^{-1}(\bar{i})} \frac{|C^h_i|}{|C_i|} \leq \frac{\sum_{i\in N^{-1}(\bar{i})} |C^h_i|}{\sum_{i\in N^{-1}(\bar{i})} |C_i|} = \frac{|C^h_{\bar{i}}|}{|C_{\bar{i}}|} \leq \max_{i\in N^{-1}(\bar{i})} \frac{|C^h_i|}{|C_i|} \leq \alpha_h \tag{10}$$

The proves guarantee (1) of the lemma. Now we prove guarantee (2), we denote by $R^*_{\textbf{DS}}$ the optimal clustering cost for the **DS** constrained problem. We can show that $\forall j \in \mathcal{C}$:

$$\begin{aligned}
d(j, \phi_0(j)) &\leq d(j, {\phi^*}_{\textbf{GF+DS}}(j)) + d({\phi^*}_{\textbf{GF+DS}}(j), \phi_0(j)) \\
&\leq d(j, {\phi^*}_{\textbf{GF+DS}}(j)) + d({\phi^*}_{\textbf{GF+DS}}(j), N({\phi^*}_{\textbf{GF+DS}}(j))) && (\text{ since } \phi_0(j) = N({\phi^*}_{\textbf{GF+DS}}(j)) \text{ }) \\
&\leq R^*_{\textbf{GF+DS}} + \alpha_{\textbf{DS}} R^*_{\textbf{DS}} && (\text{ since } \bar{S} \text{ is an } \alpha_{\textbf{DS}}\text{-approximation for } \textbf{DS} \text{ }) \\
&\leq (1 + \alpha_{\textbf{DS}}) R^*_{\textbf{GF+DS}}
\end{aligned}$$

Where the last holds since $R^*_{\textbf{DS}} \leq R^*_{\textbf{GF+DS}}$ because the set of solutions constrained by **DS** is a subset of the set of solutions constrained by **GF+DS**. $\square$

Now we can prove the lemma. By the above claim, it follows that when ASSIGNMENTGF is called, the LP solution from line (3) of algorithm block 3 satisfies: (1) The **GF** constraints exactly and (2) Has a clustering cost of at most $(1+\alpha_{\mathbf{DS}})R^*_{\mathbf{GF+DS}}$. This is because LP (2) includes all integral assignments from $\mathcal{C}$ to $\bar{S}$ including $\phi_0$. Since this LP assignment is fed to MAXFLOWGF it follows by Theorem B.1 that the final solution satisfies: (1) The **GF** constraint at an additive violation of 2, (2) Has a clustering cost of at most $(1+\alpha_{\mathbf{DS}})R^*_{\mathbf{GF+DS}}$. Guarantee (3) holds since some centers may become closed (assigned no points) and therefore $S' \subset \bar{S}$ (possibly being a proper subset). □

We restate the following theorem and give its proof:

**Theorem 4.1.** *Given an $\alpha_{\mathbf{DS}}$-approximation algorithm for the* ***DS*** *problem, then we can obtain an $2(1+\alpha_{\mathbf{DS}})$-approximation algorithm that satisfies* ***GF*** *at an additive violation of 3 and satisfies* ***DS*** *simultaneously.*

*Proof.* By Lemma 2 above, the set of centers $S'$ is a subset (possibly proper) subset of $S$ and therefore the **DS** constraints may no longer be satisfied. Algorithm 2 select points from each color $h$ so that when they are added to $S'$, then for each color $h$ the set of centers is at least $\beta_h k$. Since these new centers are opened using the DIVIDE subroutine then it follows that they are all active (guarantee (2) of Lemma 1).

Further, by guarantee (3) of Lemma 1 for DIVIDE we have for any point $j$ assigned to a new center $q$ that $d(j,q) \leq 2d(j,\phi'(j)) \leq 2(1+\alpha_{\mathbf{DS}})R^*_{\mathbf{GF+DS}}$.

Finally, by guarantee (1) of Lemma 1 DIVIDE is called over a cluster that satisfies **GF** at an additive violation of 2 and therefore the resulting additive violation is at most $\max\{2, \frac{2}{|Q_i|}+2\}$. Since $2 \leq \frac{2}{|Q_i|}+2 \leq \frac{2}{2}+2 = 3$. The additive violation is at most 3. □

We restate the next theorem and give its proof:

**Theorem 4.2.** *If we have a solution $(\bar{S},\bar{\phi})$ of cost $\bar{R}$ that satisfies the* ***GF*** *constraints where the number of non-empty clusters is $|\bar{S}| = \bar{k} \leq k$, then we can obtain a solution $(S,\phi)$ that satisfies* ***GF*** *at an additive violation of 2 and* ***DS*** *simultaneously with cost $R \leq 2\bar{R}$.*

*Proof.* We point out the following fact:

**Fact C.1.** *Every cluster in $(\bar{S},\bar{\phi})$ has at least one point from each color.*

*Proof.* This holds, since given a center $i \in \bar{S}$ we have $|\bar{C}_i| > 0$ and therefore $\forall h \in \mathcal{H} : |\bar{C}^h_i| \geq \beta_h |\bar{C}_i| > 0$ and therefore $|\bar{C}^h_i| \geq 1$ since it must be an integer. □

We note that the values $\{\beta_h, \alpha_h\}_{h\in\mathcal{H}}$ and $k$ must lead to a feasible **DS** problem, i.e. there exist positive integers $g_h$ such that $\sum_{h\in\mathcal{H}} g_h = k$ and $\forall h \in \mathcal{H} : \beta_h k \leq g_h \leq \alpha_h k$. Accordingly, since lines (4-13) in algorithm 4 can always pick a point of some color $h$ such that the upper bound $\alpha_h k$ is not exceeded for every cluster $i$. Therefore the following fact must hold

**Fact C.2.** *By the end of line (13) we have $\forall i \in \bar{S} : |Q_i| \geq 1$.*

Further, the final $s_h$ values are valid for **DS**:

**Claim 4.** *By the end of line (13) the values of $s_h$ satisfy: (1) $\sum_{h\in\mathcal{H}} s_h \leq k$, (2) $\forall h \in \mathcal{H} : \beta_h k \leq s_h \leq \alpha_h k$.*

*Proof.* Lines (4-13) add values to $s_h$ if the lower bound $\beta_h k$ for color $h$ is not satisfied. If the lower bound is satisfied for all colors, then points of some color $h$ are added provided that adding them would not exceed the upper bound of $\alpha_h k$ (see line 5). Therefore, by the end of line (13) for any color $h \in \mathcal{H} : s_h \leq \alpha_h k$ and either $s_h \geq \beta_h k$ or $s_h < \beta_h k$[3].

If by the end of line (13) we have $\forall h \in \mathcal{H} : s_h \geq \beta_h k$, then the algorithm moves to line (22). Otherwise, it will keep picking points and incrementing $s_h$ until $\forall h \in \mathcal{H} : s_h \geq \beta_h k$.

[3]To see why we could have $s_h < \beta_h k$, consider the case where $\bar{k} < k$ and therefore there would not be enough clusters to so that we can add points for each color.

Further, since such valid **DS** values exist it must be that the above satisfies $\sum_{h\in\mathcal{H}} s_h \leq k$ and $\forall h \in \mathcal{H} : s_h \leq \alpha_h k$. This concludes the proof for the claim. $\square$

By Lemma 1 for DIVIDE the new centers $S = \cup_{i\in\bar{S}} Q_i$ are all active (guarantee 2 of DIVIDE) and since the values of $s_h$ are valid (Claim 4 above), therefore $S$ satisfies the **DS** constraints.

Since the assignment in each cluster in the new solution $(S, \phi)$ is formed using DIVIDE over the clusters of $(\bar{S}, \bar{\phi})$ then by guarantee 1 of DIVIDE, each cluster $(S, \phi)$ satisifes **GF** at an additive violation of 2. Finally, the clustering cost is at most $R \leq 2\bar{R}$ (guarantee 3 of DIVIDE). $\square$

We restate the following corollary and give its proof:

**Corollary 1.** *Given an $\alpha_{\textbf{GF}}$-approximation algorithm for **GF**, then we can have a $2\alpha_{\textbf{GF}}$-approximation algorithm that satisfies **GF** at an additive violation of 2 and **DS** simultaneously.*

*Proof.* Using the previous theorem (Theorem 4.2) the solution $(\bar{S}, \bar{\phi})$ has a cost of $\bar{R} \leq \alpha_{\textbf{GF}}\, \mathrm{OPT}_{\textbf{GF}}$. The post-processed solution that satisfies **GF** at an additive violation of 2 and **DS** simultaneously has a cost of $R \leq 2\bar{R} \leq 2\alpha_{\textbf{GF}}\, \mathrm{OPT}_{\textbf{GF}} \leq 2\alpha_{\textbf{GF}}\, \mathrm{OPT}_{\textbf{GF+DS}}$. The last inequality follows because $\mathrm{OPT}_{\textbf{GF}} \leq \mathrm{OPT}_{\textbf{GF+DS}}$ which is the case since both problems minimize the same objective, however by definition the constraint set of **GF** + **DS** is a subset of the constraint set of **GF**. $\square$

Before we proceed, we define the following clustering instance which will be used in the proof:

**Definition 1.** *$\ell$-Community Instance: The $\ell$-community instance is a clustering instance where the set of points $\mathcal{C}$ can be partitioned into $\ell$ communities (subsets) $\{C_1^{CI}, \dots, C_\ell^{CI}\}$ of coinciding points (points within the same community are separated by a distance of 0). Further, the communities are of equal size, i.e. $\forall i \in \ell : |C_i^{CI}| = \frac{n}{\ell}$ . Moreover, the distance between any two points belonging to different communities in the partition is at least $R > 0$.*

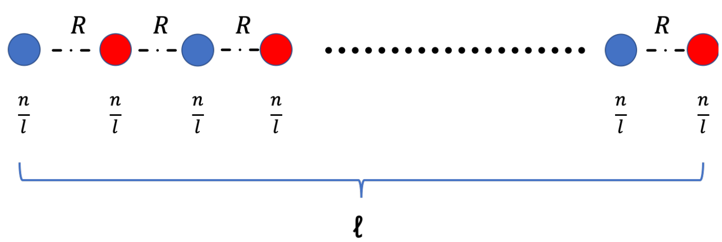


Figure 5: An $\ell$-community instance to show Price of Fairness (**GF**) and incompatibility between **GF** and other fairness constraints when $k$ is even.

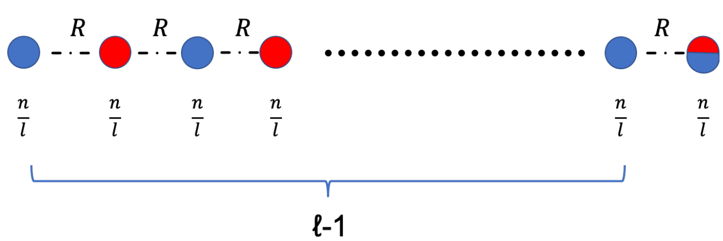


Figure 6: An $\ell$-community instance to show Price of Fairness (**GF**) and incompatibility between **GF** and other fairness constraints when $k$ is odd.

Figures 5 and 6 show two examples of the $\ell$-community instances. When clustering with a value of $k$, the given $\ell$-community instance with $k = \ell$ is arguably the most "natural" clustering instance where the clustering output is the communities $\{C_1^{CI}, \dots, C_k^{CI}\}$.

The following fact clearly holds for any $\ell$-community instance:

**Fact C.3.** *If we cluster an $\ell$-community instance with $k = \ell$ then: (1) The set of optimal solutions are $(S_{CI}, \phi_{CI})$ where $S_{CI}$ has exactly one center from each community $\{C_1^{CI}, \dots, C_k^{CI}\}$. Further, points are assigned to a center in the same community. (2) Clustering cost of $(S_{CI}, \phi_{CI})$ is 0. (3) Any solution other than $(S_{CI}, \phi_{CI})$ has a clustering cost of at least $R > 0$.*

We restate the following proposition and give its proof:

**Proposition 5.1.** *For any value of $k \geq 2$, imposing **GF** can lead to an unbounded* PoF *even if we allow an additive violation of $\Omega(\frac{n}{k})$.*

*Proof.* Consider the case where $k \geq 2$ is even and refer to Figure 5 where we have $\ell = k$ communities that alternate from red to blue color. Further, by Fact C.3 the optimal solution has a clustering cost of 0. The optimal solution would have one center in each of the $k = \ell$ communities, and assign points to its closest center.

If we set the lower and upper proportion bounds to $\frac{1}{2}$ for both colors, then to satisfy **GF** each cluster should have both red and blue points. There must exists a cluster $C_i$ of size $|C_i| \geq \frac{n}{k}$, it follows that to satisfy the **GF** constraints at an additive violation of $\rho$, then $|C_i^{\text{blue}}| \geq \frac{1}{2}|C_i| - \rho = \frac{n}{2k} - \rho$ and similarly we would have $|C_i^{\text{red}}| \geq \frac{n}{2k} - \rho$. By setting $\rho = \frac{n}{2k} - \epsilon$ for some constant $\epsilon > 0$, then we have $|C_i^{\text{blue}}|, |C_i^{\text{red}}| > 0$. This implies that a point will be assigned to a center at a distance $R > 0$ and therefore the PoF is unbounded.

For a value of $k$ that is odd, see the example of Figure 6. Here instead the last community has the same number of red and blue points. We call the cluster whose center is in the last community $C_{\text{last}}$. If $|C_{\text{last}}| \neq \frac{n}{k}$, then there are points assigned to the center of $C_{\text{last}}$ from other communities incurring cost $R > 0$ or points in the last community are assigned to other centers at distance $R > 0$. If $|C_{\text{last}}| = \frac{n}{k}$, then in the remaining $k-1$ communities with total of $n - \frac{n}{k}$ points, $k-1$ centers are chosen. There must exists a cluster $C_i$ of size $|C_i| \geq \frac{n}{k}$. We then follow the same argument as in the even $k$ case, which is to satisfy **GF** with additive violation $\rho = \frac{n}{2k} - \epsilon$ for both color, we must have $|C_i^{\text{blue}}|, |C_i^{\text{red}}| > 0$. This means at least a point will be assigned to a center at a distance $R > 0$ and therefore the PoF is unbounded for the odd $k$ case as well. $\square$

We restate the following proposition and give its proof:

**Proposition 5.2.** *For any value of $k \geq 3$, imposing **DS** can lead to an unbounded* PoF.

*Proof.* Consider a case of the $\ell$ community instance shown in Figure 7 where $k \geq 3$ and $k = \ell$. Here all communities are blue, except for the last which has $\frac{n}{2k}$ red points and $\frac{n}{2k}$ green points. Similar to the previous proposition since it is a community instance with $\ell = k$, then by Fact C.3 the optimal solution has a clustering cost of 0 and would have one center in each community and assign each point to its closest center.

Suppose for **DS** we set $k^l_{\text{blue}}, k^l_{\text{red}}, k^l_{\text{green}} > 0$, this implies that we should pick a center of each color. This implies that we can have at most $k-2$ blue center, therefore there will be a community (composed of all blue points) where no point is picked as a center. Therefore, the clustering cost is $R > 0$ and the PoF is unbounded. $\square$

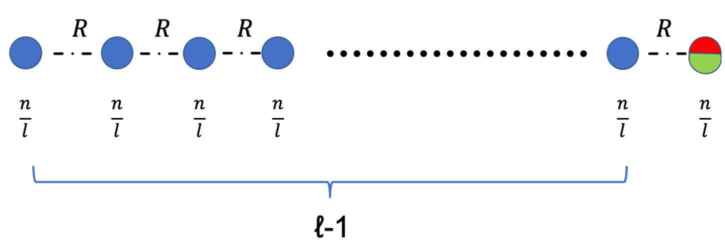


Figure 7: An $\ell$-community instance to show Price of Fairness (**DS**) and incompatibility between **DS** and other fairness constraints.

We restate the following proposition and give its proof:

**Proposition 5.3.** *For any value of $k \geq 2$, imposing **GF** on a solution that only satisfies **DS** can lead to an unbounded increase in the clustering cost even if we allow an additive violation of $\Omega(\frac{n}{k})$.*

*Proof.* The proof follows similarly to Proposition 5.1. For $k \geq 2$ and $k$ is even. Consider the same case as in Figure 5 where we have $\ell = k$. In this case, we set the upper and lower proportion bounds for both **GF** and **DS** to $\frac{1}{2}$. This implies to satisfy **DS** the number of red and blue centers should each be $\frac{k}{2}$. Thus solutions that satisfy the **DS** constraint are the optimal unconstrained solutions as specified in Fact C.3. The rest of the proof proceeds exactly as the proof for the even $k$ case in Proposition 5.1.

For $k \geq 3$ and $k$ is odd, consider the same case as in Figure 6. In this case, we set the upper and lower proportion bounds for **GF** to $\frac{1}{2}$. And we set the upper and lower bound for number of centers in **DS** constraint as $\frac{k-1}{2} + 1$ and $\frac{k-1}{2}$ respectively for both colors. Note that an optimal solution specified in Fact C.3 which chooses either a red or blue point in the right most community as a center satisfy this **DS** constraint. The rest of the proof proceeds exactly as the proof for the odd $k$ case in Proposition 5.1. □

We restate the following proposition and give its proof:

**Proposition 5.4.** *Imposing **DS** on a solution that only satisfies **GF** leads to a bounded increase in the clustering cost of at most 2 ($\text{PoF} \leq 2$) if we allow an additive violation of 2 in the **GF** constraints.*

*Proof.* This follows from Theorem 4.2 since we can always post-process a solution that only satisfies **GF** into one that satisfies both **GF** at an additive violation of 2 and **DS** simultaneously and clearly from the theorem we would have $\text{PoF} = \frac{\text{clustering cost of } \textbf{GF} \text{ post-processed solution}}{\text{clustering cost of } \textbf{GF} \text{ solution}} \leq \frac{2 \text{ clustering cost of } \textbf{GF} \text{ solution}}{\text{clustering cost of } \textbf{GF} \text{ solution}} \leq 2$. □

## D Omitted Proofs, Additional Results, and Details for Section 6

In this section, we provide more details and proofs for theorems and facts that appeared in Section 6. We present proof for theorem 6.1. We begin by giving the full definitions of the relevant fairness constraints.

**Definition 2.** *Neighborhood Radius [30]: For a given set of points $\mathcal{C}$ to cluster and a given number of centers $k$, the neighborhood radius of a point $j$ is the minimum radius $r$ such that at least $|C|/k$ of the points in $\mathcal{C}$ are within distance $r$ of $j$: $NR_{C,k}(j) = \min\{r : |B_r(j) \cap C| \geq |C|/k\}$,where $B_r(j)$ is the closed ball of radius $r$ around $j$.*

**Definition 3.** *Fairness in Your Neighborhood Constraint [30]: For a given set of points $\mathcal{C}$ with metric $d(.,.)$, a clustering $(S, \phi)$ is $\alpha_{NR}$-fair if for all $j \in \mathcal{C}$, $d(j, \phi(j)) \leq \alpha_{NR} \cdot NR_{\mathcal{C},k}(j)$.*

**Definition 4.** *Socially Fair [1, 24]: For a clustering problem with $k$ centers on points $\mathcal{C}$ which are from $|\mathcal{H}|$ groups and $\cup_{h \in \mathcal{H}} \mathcal{C}^h = \mathcal{C}$, the socially fair clustering optimization problem is to minimize the maximum average clustering cost across all groups: $\min\limits_{S:|S| \leq k, \phi} \max\limits_{h \in \mathcal{H}} \frac{1}{|\mathcal{C}^h|} \sum\limits_{j \in C^h} d^p(j, \phi(j))$[4].*

Note that socially fair does not optimize over assignment functions because it assumes assignment follows optimal rule: a point is assigned to the cluster center closest to it.

**Definition 5.** *An $\alpha_{SF}$-socially fair solution to a clustering problem is a solution of cost at most $\alpha_{SF}$ of the optimal socially fair solution.*

This definition allows us to bound the clustering cost of an $\alpha_{\text{SF}}$-Socially Fair solution $(S_\alpha, \phi_\alpha)$ as:

$$\max_{h \in \mathcal{H}} \frac{1}{|\mathcal{C}^h|} \sum_{j \in C^h} d^p(j, \phi_\alpha(j)) \leq \alpha_{\text{SF}} \min_{S:|S| \leq k} \max_{h \in \mathcal{H}} \frac{1}{|\mathcal{C}^h|} \sum_{j \in C^h} d^p(j, \phi(j)).$$

**Definition 6.** ***Approximately proportional**[15]: Given a set of centers $S \subseteq \mathcal{C}$ with $|S| = k$, it is $\alpha_{AP}$-approximately proportional ($\alpha_{AP}$-proportional) if $\forall U \subseteq \mathcal{C}$ and $|U| \geq \lceil \frac{n}{k} \rceil$ and for all $y \in C$, there exists $i \in U$ with $\alpha_{AP} \cdot d(i, y) \geq d(i, \phi(i))$ .*

[4] $p = 1$ for the $k$-median and $p = 2$ for the $k$-means

We restate the following theorem and give its proof:

**Theorem 6.1.** *For any value $k \geq 2$, the **fairness in your neighborhood** [30], **socially fair** constraint [1, 24] are each incompatible with **GF** even if we allow an additive violation of $\Omega(\frac{n}{k})$ in the **GF** constraint. For any value $k \geq 5$, the **proportionally fair** constraints [15] is incompatible with **GF** even if we allow an additive violation of $\Omega(\frac{n}{k})$ in the **GF** constraint.*

*Proof.* We show incompatibility of **GF** with the three Fairness notions. Recall that we consider **GF** and another fairness constraint at the same time, and incompatible means there are cases where no feasible solution exists that satisfies both constraints at the same time.

**Lemma 4.** *For any $k \geq 2$, there exist a clustering problem where no feasible solution exists that satisfies both Fairness in Your Neighborhood and **GF** even if we allow an additive violation of $\Omega(\frac{n}{k})$ in the **GF** constraint.*

*Proof.* Consider the case where $k \geq 2$, and consider the clustering problem on a $\ell$-community instance with $k = \ell$. We consider the case where lower and upper proportion bound of **GF** are set to $\frac{1}{2}$ for both colors.

**Claim 5.** *On the above mentioned clustering problem, an $\alpha_{\text{NR}}$-fair solution in the Fairness in Your Neighborhood notion for finite $\alpha_{\text{NR}}$ is a solution in the set of optimal solutions $(S_{CI}, \phi_{CI})$.*

*Proof.* By Definitions 1 and 2, for any point $j \in \mathcal{C}$, its neighborhood radius is $\text{NR}_{C,k}(j) = \min\{r : |B_r(j) \cap C| \geq |C|/k\} = 0$. This is because each point is in one of the $l$ subset, and by definition, the subset is of size $\frac{n}{l} = \frac{n}{k}$, and points in the same subset are separated by a distance 0.

For a solution on a $\ell$-community instance $(S, \phi) \in (S_{CI}, \phi_{CI})$, for any point $j \in \mathcal{C}$, because $S$ contains a center in the community where $j$ is, $d(j, S) = 0$.

By definition of $\alpha_{NR}$-fairness, this means on a $\ell$-community instance, any solution in $(S_{CI}, \phi_{CI})$ is $\alpha_{\text{NR}}$-fair with a finite $\alpha_{\text{NR}}$. This is because for any $(S, \phi) \in (S_{CI}, \phi_{CI})$, $d(j, S) = 0 \leq \alpha_{\text{NR}} \cdot \text{NR}_{C,k}(j)$ holds for $\alpha_{\text{NR}}$ equal to any finite value.

In any solution that is not in $(S_{CI}, \phi_{CI})$, there is at least a point which is assigned to a center not in the point's own community. Thus for such a solution $(S, \phi)$, there exist $j \in \mathcal{C}$, $d(i, S) = R$. Thus for $(S, \phi) \notin (S_{CI}, \phi_{CI})$, for some $j \in \mathcal{C}$, there is no finite $\alpha_{\text{NR}}$ such that $d(j, \phi(j)) = R \leq \alpha_{\text{NR}} \cdot \text{NR}_{C,k}(j)$ holds.

This shows that a solution that achieves $\alpha_{NR}$-fairness for a finite $\alpha_{NR}$ must be a solution from the set of solutions $(S_{CI}, \phi_{CI})$. □

Thus we have shown that any solution that satisfies the fairness in your neighborhood constraint approximately do not assign points to centers not in its original community.

To characterize the set of solutions that satisfy **GF** with additive $\Omega(\frac{n}{k})$ violation, we consider two cases separately: $k$ is even and $k$ is odd.

Consider the case where $k \geq 2$ is even and refer to Figure 5 where we have $\ell = k$ communities that alternate from red to blue color.

Since the lower and upper proportion bounds are set to $\frac{1}{2}$ for both colors, then to satisfy **GF** each cluster should have both red and blue points. There must exists a cluster $C_i$ of size $|C_i| \geq \frac{n}{k}$, it follows that to satisfy the **GF** constraints at an additive violation of $\rho$, then $|C_i^{\text{blue}}| \geq \frac{1}{2}|C_i| - \rho = \frac{n}{2k} - \rho$ and similarly we would have $|C_i^{\text{red}}| \geq \frac{n}{2k} - \rho$. By setting $\rho = \frac{n}{2k} - \epsilon$ for some constant $\epsilon > 0$, then we have $|C_i^{\text{blue}}|, |C_i^{\text{red}}| > 0$. This implies that a point need be assigned to a center at a distance $R > 0$ for the solution to satisfy **GF** with additive $\Omega(\frac{n}{k})$ violation. Therefore such a solution is not in the solution set that satisfies fairness in your neighborhood.

For a value of $k$ that is odd, see the example Figure 6. Here instead the last community has the same number of red and blue points. We call the cluster whose center is in the last community $C_{\text{last}}$.

If $|C_{\text{last}}| \neq \frac{n}{k}$, then there are points assigned to the center of $C_{\text{last}}$ from other communities incurring cost $R > 0$ or points in the last community are assigned to other centers at distance $R > 0$. In both cases there is at least a point assigned to a center not in its community. If $|C_{\text{last}}| = \frac{n}{k}$, then in the remaining $k-1$ communities with total of $n - \frac{n}{k}$ points, $k-1$ centers are chosen. A solution satisfying **GF** has one cluster $C_i$ with at least $\frac{1}{k-1}\left(n - \frac{n}{k}\right) = \frac{n}{k}$ points. Then we follow the same argument as in the even $k$ case. That is, to satisfy **GF** with $\rho$ additive violation on the $C_i$, $|C_i^{\text{blue}}| \geq \frac{1}{2}|C_i| - \rho = \frac{n}{2k} - \rho$, $|C_i^{\text{red}}| \geq \frac{n}{2k} - \rho$, with $\rho = \frac{n}{2k} - \epsilon$ for some constant $\epsilon > 0$, at least a point will be assigned to center at distance $R > 0$. Thus such a solution is not in the set of solutions $(S_{CI}, \phi_{CI})$. Thus the set of solutions that satisfies fairness in your neighborhood has no overlap with the set of solutions that satisfies **GF** with $\Omega(\frac{n}{k})$ additive violation. □

**Lemma 5.** *For any $k \geq 2$, there exist a clustering problem where no feasible solution exists that satisfies both Socially Fair and **GF** even if we allow an additive violation of $\Omega(\frac{n}{k})$ to the **GF** constraint.*

*Proof.* We follow a similar line of argument as in Lemma 4. Consider the case when $k \geq 2$, and consider the clustering problem on a $\ell$-community instance with $k = \ell$. We consider the case where lower and upper proportion bound of **GF** are set to $\frac{1}{2}$ for both colors.

**Claim 6.** *On the above mentioned clustering problem, an $\alpha_{SF}$-fair solution in the Socially Fair notion for finite $\alpha_{SF}$ is a solution in the set of optimal solutions $(S_{CI}, \phi_{CI})$.*

*Proof.* Denote the clustering cost of an optimal solution to the a Socially Fair clustering problem as $\text{OPT}_{\text{SF}}$. By definition,

$$\text{OPT}_{\text{SF}} = \min_{S:|S|\leq k} \max_{h\in\mathcal{H}} \frac{1}{|\mathcal{C}^h|} \sum_{j\in C^h} d^p(j, \phi(j)).$$

We can formulate a problem that aims to find an $\alpha_{\text{SF}}$-socially fair solution as a constrained optimization problem. We use a dummy objective function $f$. The constraint can be set up as requiring maximum clustering costs across all colors to be upper-bounded by $\alpha_{\text{SF}}$ times that of the optimal socially fair solution $\text{OPT}_{\text{SF}}$.

The constrained program can be set up as below:

$$\begin{aligned} &\min_{S:|S|\leq k} f \\ &\text{s.t. } \max_{h\in\mathcal{H}} \frac{1}{|\mathcal{C}^h|} \sum_{j\in C^h} d^p(j, \phi(j)) \leq \alpha_{\text{SF}}\text{OPT}_{\text{SF}} \end{aligned}$$

For a clustering problem with $k$ centers on the $\ell$-community instance with $k = \ell$, in a solution $(S, \phi)$ that has one center in each subset, $d(j, \phi(j)) = 0$ for each point $j \in \mathcal{C}$. Thus this solution has clustering cost for each color $h$ as $\sum_{j\in C^h} d^p(j, \phi(j)) = 0$. Which implies that $\text{OPT}_{\text{SF}} = 0$.

Thus on the $\ell$-community instance, feasible solutions to the $\alpha_{\text{SF}}$-socially fair problem, for finite $\alpha_{\text{SF}}$, have $\max_{h\in\mathcal{H}} \frac{1}{|\mathcal{C}^h|} \sum_{j\in C^h} d^p(j, \phi(j)) = 0$. We now show $(S_{CI}, \phi_{CI})$ is the only set of solutions that have $\max_{h\in\mathcal{H}} \sum_{j\in C^h} d^p(j, \phi(j)) = 0$. Thus, they will be the only feasible solutions.

For any solution $(S, \phi)$ that is not in $(S_{CI}, \phi_{CI})$, there must be a point assigned to a center that is not in its own community. For such a point $d(j, \phi(j)) = R$. Thus $\max_{h\in\mathcal{H}} \frac{1}{|\mathcal{C}^h|} \sum_{j\in C^h} d^p(j, \phi(j)) \geq \frac{R}{\max_{h\in\mathcal{H}} |\mathcal{C}^h|}$. Therefore, an $\alpha_{\text{SF}}$-fair solution in the socially fair notion for finite $\alpha_{\text{SF}}$ must be a solution in the set of optimal solutions $(S_{CI}, \phi_{CI})$ .

□

At this point, similar to proof of Lemma 4, we had shown that any solution that satisfies socially fair approximately do not assign points to centers not in its original community on the $\ell$-community instance. The remaining of the proof is the same as that part of the proof in Lemma 4. We can use the same examples for even $k$ and odd $k$ to show that any solution that satisfies **GF** with $\Omega(\frac{n}{k})$ additive violation any $k \geq 2$ is not in the set of solutions $(S_{CI}, \phi_{CI})$. Thus the set of solutions that satisfies socially fair has no overlap with the set of solutions that satisfies **GF** with $\Omega(\frac{n}{k})$ additive violation. □

**Lemma 6.** *For any $k \geq 5$, there exist a clustering problem where no feasible solution exists that satisfies both Proportional Fairness and **GF** even if we allow an additive violation of $\Omega(\frac{n}{k})$ in the **GF** constraint.*

*Proof.* For a given value of $\alpha_{AP}$ for the proportionally fair constraint, consider Figure 8. For the **GF** constraints, the upper and lower bounds for each color to $\frac{1}{2}$ and the total number of points $n$ is always even. Consider some $k \geq 5$. It follows that the sum of cluster sizes assigned to centers on either the right side or the left side would be at least $\frac{n}{2}$, WLOG assume that it is the left side and denote the total number of points assigned to clusters on the left size by $|C_{LS}|$ and let $S_{LS}$ be the centers on the left side. The total number of points on the left side may not be assigned to a single center but rather distributed among the centers $S_{LS}$. To satisfy the **GF** constraints at an additive violation of $\rho$, it follows that the number of red points that have to be assigned to the left side is at least $\sum_{i \in S_{LS}} (\frac{1}{2}|C_i| - \rho) \geq \frac{n}{4} - k\rho$. Set $\rho = \frac{n}{4k} - \frac{n}{k^2} - 1$, then it follows that at least $\lceil \frac{n}{k} \rceil$ red points are assigned to a center on the left at a distance of at least $R$. Since the maximum distance between any two red points by the triangle inequality is $2r < \frac{R}{\alpha_{AP}}$ it follows that this set of red points forms a blocking coalition. I.e., these points would also have a lower distance from their assigned center if they were instead assigned to a red center. □

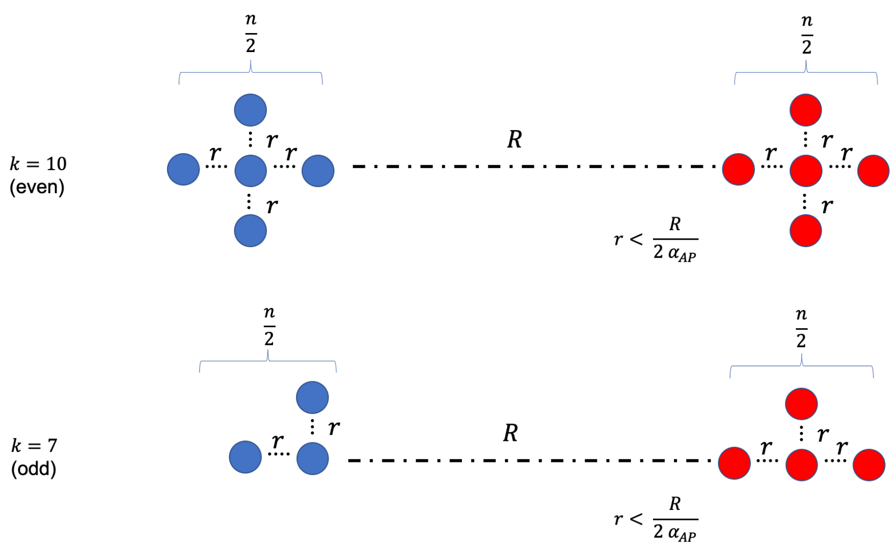


Figure 8: Instances to show incompatibility between Proportional Fairness and **GF**. We always have $n/2$ blue points on the left and $n/2$ red points on the right. For even $k$ we would have $k/2$ locations for the blue and red points each. For odd $k$ we have $\lfloor k/2 \rfloor$ blue locations and $\lceil k/2 \rceil$ red locations. For each color, there is always a location at the center at a distance $r$ from the other locations. Points of different color are at a distance of at least $R$ from each other. For any value of $\alpha_{AP}$ for the proportionally fair constraint, we set $r < \frac{R}{2\alpha_{AP}}$.

□

We restate the following theorem and give its proof:

**Theorem 6.2.** *For any value $k \geq 3$, the **fairness in your neighborhood** [30], **socially fair** [1, 24] and **proportionally fair** [15] constraints are each incompatible with **DS**.*

*Proof.* Consider a case of the $\ell$-community instance where the first $\ell - 1$ communities consist of points of only blue points. And the last community contains $\frac{n}{2\ell}$ points of red points and $\frac{n}{2\ell}$ points of green color. This $\ell$-community instance is illustrated in Figure 7. Consider the clustering problem where $k = \ell$ and **DS** constraint $k_{\text{blue}}, k_{\text{red}}, k_{\text{green}} > 0$. We establish below two claims.

**Claim 7.** *On the above mentioned clustering problem, an $\alpha_{NR}$-fair solution in the Fairness in Your Neighborhood notion for finite $\alpha_{NR}$ is a solution in the set of optimal solutions $(S_{CI}, \phi_{CI})$.*

**Claim 8.** *On the above mentioned clustering problem, an $\alpha_{SF}$-fair solution in the Socially Fair notion for finite $\alpha_{SF}$ is a solution in the set of optimal solutions $(S_{CI}, \phi_{CI})$.*

Those two claims can be proved with the same argument as in claim 5 and claim 6.

However, satisfying **DS** on this with $k_{\text{blue}}, k_{\text{red}}, k_{\text{green}} > 0$ requires a center of each color be picked. Thus a solution from the set of optimal solutions $(S_{CI}, \phi_{CI})$ does not satisfy **DS** because it will only pick one point from the right most subset as a center. Thus either green points or red points will not appear in the set of centers.

On the other hand, since a solution satisfying **DS** has at least one center of each color, it will contain two centers, one green, one red chosen from the right most subset. And there are $k-2$ centers allocated to the $k-1$ communities on the left. By pigeon hole principle, one of the communities of all blue points will have no center allocated. All blue points in this subset are then assigned to a center in a nearby community, thus a **DS** satisfying solution is not in the set $(S_{CI}, \phi_{CI})$. Thus the set of **DS** satisfying solutions has no overlap with the set of solutions that satisfy either one of the two fairness constraints.

Below we use the same example to show incompatibility between **DS** and Proportional Fair.

**Claim 9.** *On the above mentioned clustering problem, there is no feasible solution exists that satisfies both Proportional Fairness and **DS**.*

*Proof.* We show a **DS** satisfying solution on above example is not proportional fair. As argued above, a solution satisfying **DS** can allocate $k-2$ centers for the $k-1$communities on the left. There will be a community of size $\frac{n}{k}$ of which all points are assigned to a nearby center not in its community. This community forms a coalition of size $\frac{n}{k}$ and would have smaller distance if they get assigned a center in their own community. Therefore a **DS** satisfying solution is not proportional fair. □

□

**Remark:** For each of the above proofs we constructed an example which is parametric in the number of centers $k$. Moreover, for these examples for the optimal unconstrained $k$-center objective to equal 0 at least $k$ centers have to be used. I.e., the points are spread over at least $k$ locations. Furthermore, it is not difficult to see in each of the above examples that an optimal solution for the unconstrained $k$-center objective satisfies fairness in your neighborhood, socially fair, and the proportionally fair constraints. In fact, it is easy to show that any $k$-center which has a radius of 0 immediately satisfies fairness in your neighborhood, socially fair, and the proportionally fair constraints. However, the same is not true for **GF** or **DS**. This indicates that the above distance-based fairness constraints can be aligned with the clustering cost whereas the same cannot be said about **GF** or **DS**.

**Compatibility between GF and DS:** One can easily show compatibility between **GF** and **DS**. Specifically, consider some values for the centers over the colors $\{k_h\}_{h\in\mathcal{H}}$ that satisfies the **DS** constraints, i.e. $\forall h \in \mathcal{H} : k_h^l \leq k_h \leq k_h^u$ and has $\sum_{h\in\mathcal{H}} k_h \leq k$. Then simply pick a set $Q_h$ of $k_h$ points of color $h$. Now if we give DIVIDE the entire dataset $\mathcal{C}$ and the set of centers $\cup_{h\in\mathcal{H}} Q_h$ as inputs, i.e. call DIVIDE$(\mathcal{C}, \cup_{h\in\mathcal{H}} Q_h)$, then by the guarantees of divide each center would be active and each cluster would satisfy the **GF** constraints at an additive violation of 2.

Our final conclusions about the incompatibility and compatibility of the constrains are summarized in Figure 9.

# E Example for Running DIVIDE:

Consider the following example running the DIVIDE subroutine. Specifically, we have a set of points $C$ with a total of $n = |C| = 38$ points. We have 3 colors (blue, red, and green) with the following points: $|C^{\text{blue}}| = 15$, $|C^{\text{red}}| = 14$, and $|C^{\text{green}}| = 9$. We have $Q \subset C$ with a total size of 4 ($|Q| = 4$). Accordingly, we have $T_{\text{blue}} = \frac{15}{4} = 3\frac{3}{4}$, $T_{\text{red}} = \frac{14}{4} = 3\frac{1}{2}$, and $T_{\text{green}} = \frac{9}{4} = 2\frac{1}{4}$. Therefore, in the beginning of the iteration for each color $h$ (line (8) in algorithm block 1) we have $b_{\text{blue}} = 3$, $b_{\text{red}} = 2$, $b_{\text{green}} = 1$. Following the execution of the algorithm, the first three centers $q = 0$ to $q = 2$ receive

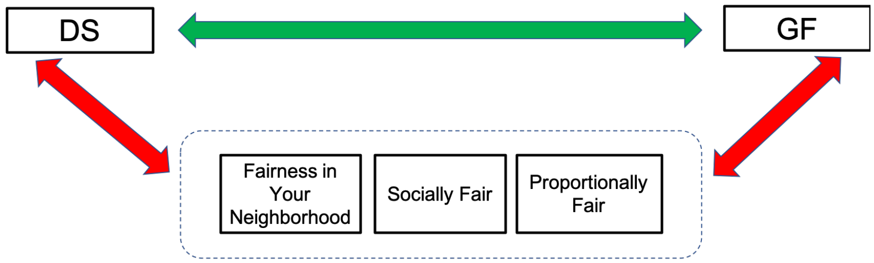


Figure 9: (In)Compatibility of clustering constraints. Red arrows indicate empty feasible set when both constraints are applied, while green arrows indicate non-empty feasibility set when both constraints are applied.

$\lceil T_{\text{blue}} \rceil$ many blue points, the last ($q = |Q| - 1$) and first center ($q = 0$) receive $\lceil T_{\text{red}} \rceil$ many red points, and center $q = 1$ receives $\lceil T_{\text{green}} \rceil$. All other assignments would be the floor of $T_{\text{h}}$. Figure 10 illustrates this.

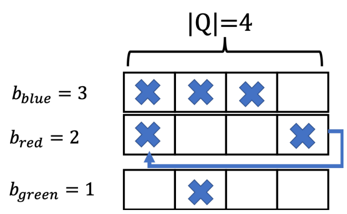


Figure 10: Diagram illustrating how DIVIDE would run over the example. The "tape" has different centers (cells) starting from $q = 0$ and ending with $q = |Q| - 1$. We go over the tape for each color $h \in \mathcal{H}$. In a given row $h$, centers marked with an **X** are assigned $\lceil T_h \rceil$ points, otherwise they are assigned $\lfloor T_h \rfloor$ points.

## F Additional Experiments Results

Here we show additional experimental results. As a reminder, the lower and upper proportion bounds for any color $h$ to $\alpha_h = (1+\delta)r_h$ and $\beta_h = (1-\delta)r_h$ for some $\delta \in [0, 1]$. Further, the **DS** constraints are set to $k^l_h = \lceil \theta r_h k \rceil$ where $\theta \in [0, 1]$ and $k^u_h = k$ for every color $h \in \mathcal{H}$.

We call our run over the **Adult** dataset in Section 7 as (**A-Adult**). In that run $\delta = 0.2$ and $\theta = 0.8$. We also, run another experiment (**B-Adult**) over the **Adult** where we set $\delta = 0.05$ and $\theta = 0.9$. Figure 11 shows the new results. We do not see a change qualitatively. It is perhaps noteworthy that the **DS-Violation** values for COLOR-BLIND and ALG-GF are even higher as well as the **GF-Violation** for ALG-DS. On the other hand, we find that our algorithms that satisfy **GF+DS** have very low (almost zero) values for **GF-Violation** and **DS-Violation** at a moderate **PoF** that is comparable to ALG-GF which satisfies only one constraint.

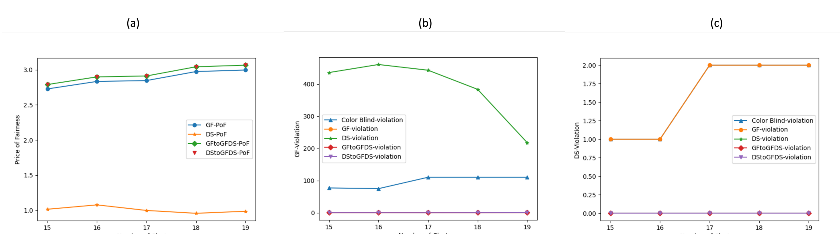


Figure 11: **B-Adult** results: (a) **PoF** comparison of 5 algorithms, with **COLOR-BLIND** as baseline; (b) **GF-Violation** comparison; (c) **DS-Violation** comparison.

Additionally, we show results over the **Census1990** dataset where we use age as the color (group) membership attribute. As done in [20] we merge the 9 age groups into 3. Specifically, groups $\{0, 1, 2\}$, $\{4, 5, 6\}$, and $\{7, 8\}$ are each merged into one group leading to total of 3 groups. Further, we subsample $6,000$ records from the dataset. We run two experiments where in the first **(A-Census1990)** we have $\delta = 0.05$ and $\theta = 0.7$ whereas in the second **(B-Census1990)** we have $\delta = 0.1$ and $\theta = 0.8$. We also use different cluster values. In terms of the 3 objective measures of **PoF**, **GF-Violation**, and **DS-Violation**, we do not see a qualitative change as can be seen from Figures 12 and 13. Specifically, the **DS** algoruthim (ALG-DS) has a low **PoF** but high **GF-Violation**. Further, COLOR-BLIND and ALG-GF have significant **DS-Violation** values. On the other hand our algorithms for **GF+DS** have low values for both **GF-Violation** and **DS-Violation** and a moderate **PoF**.

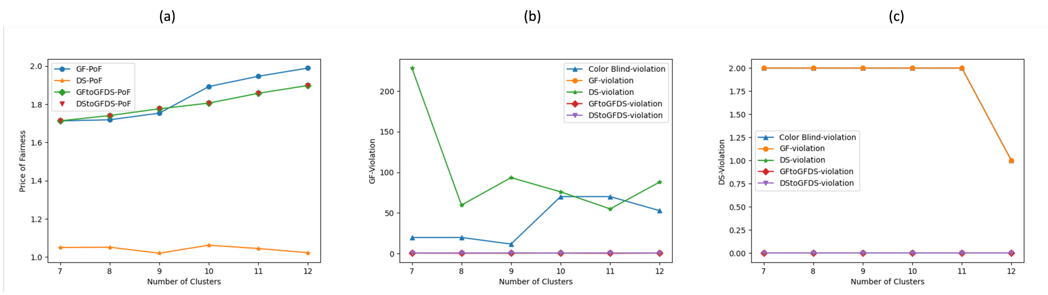


Figure 12: **A-Census1990** results: (a) **PoF** comparison of 5 algorithms, with **COLOR-BLIND** as baseline; (b) **GF-Violation** comparison; (c) **DS-Violation** comparison.

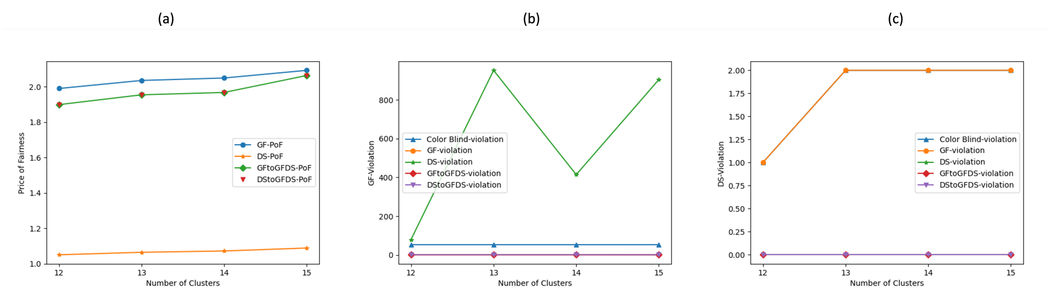


Figure 13: **B-Census1990** results: (a) **PoF** comparison of 5 algorithms, with **COLOR-BLIND** as baseline; (b) **GF-Violation** comparison; (c) **DS-Violation** comparison.

**Run-Time:** Here we show some run-time analysis results. We first calculate the "incremental" run time over **GF**. Specifically, given a solution from the ALG-GF (the algorithm for the **GF** constraints) we see that the additional run-time to post-process it to satisfy the **GF+DS** is very small in proportion. We measure $t_{\textbf{GF}\rightarrow\textbf{GF+DS}} = \frac{\text{Time to process } \textbf{GF} \text{ Solution to satisfy } \textbf{GF+DS}}{\text{Time to obtain } \textbf{GF} \text{ Solution}}$ and show the results in Figure 14 over all 4 runs. We see that the additional run time is constantly at least two orders of magnitude smaller than the time required to obtain a **GF** solution. The fact that added run-time is small further encourages a decision maker to satisfy the **DS** constraint given a solution that satisfies **GF** only.

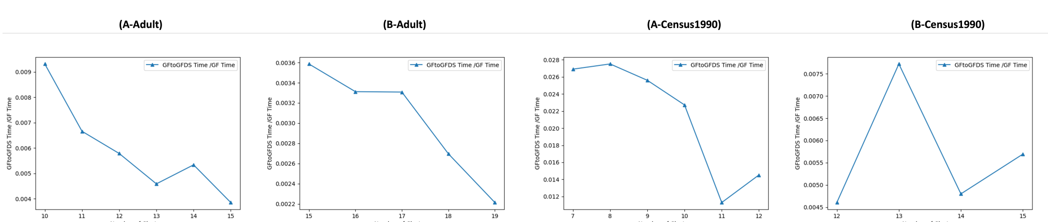


Figure 14: $t_{\textbf{GF}\rightarrow\textbf{GF+DS}}$ over the 4 runs of **(A-Adult)**, **(B-Adult)**, **(A-Census1990)**, and **(B-Census1990)**.

If we were to do the same using the ALG-DS we find the opposite. Specifically, we measure $t_{\textbf{DS}\rightarrow\textbf{GF+DS}} = \frac{\text{Time to process } \textbf{DS} \text{ Solution to satisfy } \textbf{GF+DS}}{\text{Time to obtain } \textbf{DS} \text{ Solution}}$ and find that the additional run-time required to satisfy **GF+DS** starting from a **DS** solution is orders of magnitude higher in comparison to the time required to satisfy **DS** as shown in Figure 15. We conjecture that the reason is that ALG-DS is highly optimized in terms of run-time since it runs in $O(nk)$ time [37]. On the other hand, the post-processing step (post-processing a **DS** solution to a **GF+DS**) requires solving an LP which although is done in polynomial time, can be more costly in terms of run time.

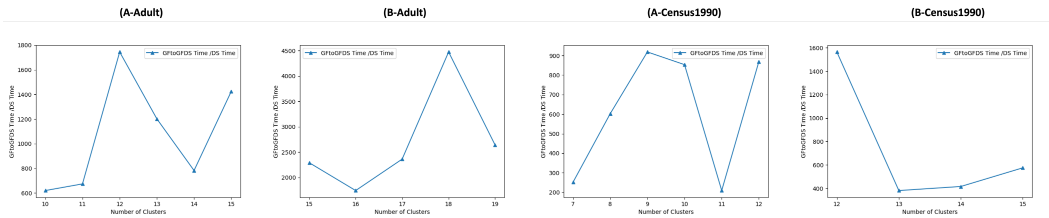


Figure 15: $t_{\textbf{DS}\rightarrow\textbf{GF+DS}}$ over the 4 runs of **(A-Adult)**, **(B-Adult)**, **(A-Census1990)**, and **(B-Census1990)**.

Finally, we show a full run-time comparison between GFTOGFDS which starts from a **GF** solution and DSTOGFDS which starts from a **DS** solution. We find that the run-times are generally comparable with one algorithm at times being faster than the other.

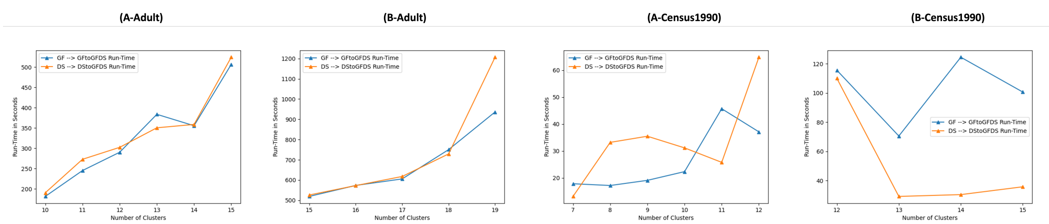


Figure 16: Full run-time comparison between GFTOGFDS and DSTOGFDS over the 4 runs of **(A-Adult)**, **(B-Adult)**, **(A-Census1990)**, and **(B-Census1990)**.

**Using a Bi-Criteria Algorithm as ALG-GF:** Our implementation of **GF** follows Bercea et al. [10] and Bera et al. [9] which would violate the **GF** constraints by at most 2. Empirically, this may cause issues for the GFTOGFDS algorithm since it requires a **GF** algorithm with zero additive violation and therefore assumes that every cluster has at least one point from each color. However, it would not cause issues as long as the resulting solution satisfies condition *of having at least one point from each color in every cluster* which would be the case if $\min_{h\in\mathcal{H},i\in\bar{S}} \beta_h|\bar{C}_i| > 2$ where $\bar{S}$ is the **GF** solution and $\bar{C}_i$ is its $i^{\text{th}}$ cluster. If the condition not met, then it is reasonable to think that the value of $k$ was set too high or that the dataset includes outlier points since the cluster sizes are very small. Furthermore, using a **GF** algorithm with an additive violation of 2 lead to a final **GF+DS** having a **GF** violation of at most 4 if the condition is satisified. However, empirically we find the **GF** violation to be generally smaller than 1. Finally, note that we treat the **GF** algorithm as a block-box and therefore it can be replaced by other algorithms such as those of [17, 40] which have no violation for **GF**. In our experiments, we run our algorithms over datasets and value of $k$ where the condition is satisifed.