Proportionally Fair Clustering

Xingyu Chen [1] **Brandon Fain** [1] **Liang Lyu** [1] **Kamesh Munagala** [1]

Abstract

We extend the fair machine learning literature by considering the problem of proportional centroid clustering in a metric context. For clustering n points with k centers, we define fairness as proportionality to mean that any n/k points are entitled to form their own cluster if there is another center that is closer in distance for all n/k points. We seek clustering solutions to which there are no such justified complaints from any subsets of agents, without assuming any a priori notion of protected subsets. We present and analyze algorithms to efficiently compute, optimize, and audit proportional solutions. We conclude with an empirical examination of the tradeoff between proportional solutions and the k-means objective.

1. Introduction

The data points in machine learning are often real human beings. There is legitimate concern that traditional machine learning algorithms that are blind to this fact may inadvertently exacerbate problems of bias and injustice in society (Julia Angwin & Lauren Kirchner, 2016). Motivated by concerns ranging from the granting of bail in the legal system to the quality of recommender systems, researchers have devoted considerable effort to developing fair algorithms for the canonical supervised learning tasks of classification and regression (Dwork et al., 2012; Kleinberg et al., 2016; Hardt et al., 2016; Kleinberg et al., 2017; Zafar et al., 2017a; Corbett-Davies et al., 2017; Pleiss et al., 2017; Zafar et al., 2017b; Kearns et al., 2018; Goel et al., 2018; Hashimoto et al., 2018).

We extend this work to a canonical problem in unsupervised learning: centroid clustering. In centroid clustering, we want to partition data into k clusters by choosing k "centers" and then matching points to one of the centers. This is a classic context for clustering work (Gonzalez, 1985; Shmoys et al., 1997; Charikar et al., 2002; Arya et al., 2004), and is perhaps best known as the setting for the celebrated k-means heuristic (independently discovered many times, see (Jain, 2010) for a brief history). We provide a novel group based notion of fairness as proportionality, inspired by recent related work on the fair allocation of public resources (Aziz et al., 2017; Conitzer et al., 2017; Fain et al., 2018; Garg et al., 2018). We suppose that data points represent the individuals to whom we wish to be fair, and that these agents prefer to be clustered accurately (that is, they prefer their cluster center to be representative of their features). A solution is fair if it respects the entitlements of groups of agents, where we assume that a subset of agents is entitled to choose a center for themselves if they constitute a sufficiently large fraction of the population with respect to the total number of clusters (*e.g.*, $1/k$ of the population, if we are clustering into k groups). The guarantee must hold for *all* subsets of sufficient size, and therefore does not hinge on any particular a priori knowledge about which points should be protected. This is in line with other recent observations that information about which individuals should be protected may not be available in practice (Hashimoto et al., 2018).

Consider a motivating example where proportional clustering might be preferable to more standard clusterings that try to minimize the k-means or k-median objective. Suppose there are 3 spherical clusters in the data: A, B, and C, and we are computing a 3-clustering. A, B, and C each contain 1/3 of the total data. The radius of A is very large compared to the radii of B and C, and A is very far away from B and C compared to the radius of A. The radii of B and C are very small, and B and C are close relative to the radius of A. More simply, A is a large sphere very far away from two small spheres B and C, which are close together.

Simply placing centers at the middle of A, B, and C is proportional. However, the global k-means or k-median minimizer places 1 center for B and C to share, and uses the remaining 2 centers to cover A. Such a solution is arbitrarily not-proportional as the radii of B and C become arbitrarily small. Essentially, the global optimum forces B and C to share a center in order to pay for the high variance in A.

To interpret this example, suppose we are clustering home locations to decide where to build public parks. B and C are

[1] Department of Computer Science, Duke University, Durham, North Carolina, USA. Correspondence to: Brandon Fain <btfain@cs.duke.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

dense urban centers, and A is a suburb. Minimizing total distance seems reasonable, but the global optimum builds 2 parks for A, and only 1 that B and C must share. Alternatively, A, B, and C might represent clusters of patients in a medical study. Both solutions distinguish A from B and C, but the global optimum obscures the secondary difference between B and C. In both instances, B or C could represent a protected group (e.g., home location may be racially divided, and race or sex could cause differences in medical data), in which case proportionality provides a guarantee even if we do not have access to this information.

1.1. Preliminaries and Definition of Proportionality

We have a set $\mathcal{N}$ of $|\mathcal{N}| = n$ individuals or data points, and a set $\mathcal{M}$ of $|\mathcal{M}| = m$ feasible cluster centers. We will sometimes consider the important special case where $\mathcal{M} = \mathcal{N}$ (i.e., where one is only given a single set of points as input), but most of our results are for the general case where we make no assumption about $\mathcal{M} \cap \mathcal{N}$. For all $i, j \in \mathcal{N} \cup \mathcal{M}$, we have a distance $d(i, j)$ satisfying the triangle inequality. Our task is centroid clustering as treated in the classic k-median, k-means, and k-center problems. We wish to open a set $X \subseteq \mathcal{M}$ of $|X| = k$ centers (assume $|\mathcal{M}| \geq k$), and then match all points in $\mathcal{N}$ to their closest center in X. For a particular solution X and agent $i \in \mathcal{N}$, let $D_i(X) = \min_{x \in X} d(i, x)$. In general, a good clustering solution X will have small values of $D_i(X)$, although the aforementioned objectives differ slightly in how they measure this. In particular, the k-median objective is $\sum_{i \in \mathcal{N}} D_i(X)$, the k-means objective is $\sum_{i \in \mathcal{N}} (D_i(X))^2$, and the k-center objective is $\max_{i \in \mathcal{N}} D_i(X)$.

To define proportional clustering, we assume that individuals prefer to be closer to their center in terms of distance (i.e., ensuring that the center is more representative of the point). Any subset of at least $r\lceil \frac{n}{k} \rceil$ individuals is entitled to choose r centers. We call a solution proportional if there does not exist any such sufficiently large set of individuals who, using the number of centers to which they are entitled, could produce a clustering among themselves that is to their mutual benefit in the sense of Pareto dominance. More formally, a *blocking coalition* is a set $S \subseteq \mathcal{N}$ of at least $r\lceil \frac{n}{k} \rceil$ points and a set $Y \subseteq \mathcal{M}$ of at most r centers such that $D_i(Y) < D_i(X)$ for all $i \in S$. It is easy to see that because $D_i(X) = \min_{x \in X} d(i, x)$, this is functionally equivalent to Definition 1; a larger blocking coalition necessarily implies a blocking coalition with a single center. We provide a brief example parsing the definition, along with its approximation, in the full version of this paper (Chen et al., 2019).

Definition 1. *Let $X \subseteq \mathcal{M}$ with $|X| = k$. $S \subseteq N$ is a* ***blocking coalition*** *against X if $|S| \geq \lceil \frac{n}{k} \rceil$ and $\exists y \in \mathcal{M}$ such that $\forall i \in S$, $d(i, y) < D_i(X)$. $X \subseteq \mathcal{N}$ is* ***proportional*** *if there is no blocking coalition against X.*

Equivalently, X is proportional if $\forall S \subseteq \mathcal{N}$ with $|S| \geq \lceil \frac{n}{k} \rceil$ and for all $y \in \mathcal{M}$, there exists $i \in S$ with $d(i, y) \geq D_i(X)$. It is important to note that this quantification is over *all* subsets of sufficient size. Hence, in attempting to satisfy the guarantee for a particular subset S, one cannot simply consider a single $i \in S$ and ignore all of the other points, as $S \backslash \{i\}$ may itself be a subset to which the guarantee applies.

Proportionality has many advantages as a notion of fairness in clustering, beyond the intuitive appeal of groups being entitled to a proportional share of centers. We name a few of these advantages explicitly.

- Proportionality implies (weak) *Pareto optimality*: namely, for any proportional solution X, there does not exist another solution X' such that $D_i(X') < D_i(X)$ for all $i \in \mathcal{N}$.
- Proportionality is *oblivious* in the sense that it does not depend on the definition of sensitive attributes or protected sub-groups.
- Proportionality is *robust* to outliers in the data, since only groups of points of sufficient size are entitled to their own center.
- Proportionality is *scale invariant* in the sense that a multiplicative scaling of all distances does not affect the set of proportional solutions.
- Approximately proportional solutions can be *efficiently computed*, and one can optimize a secondary objective like k-median subject to proportionality *as a constraint*, as we show in Section 2 and Section 3.
- Proportionality can be *efficiently audited*, in the sense that one does not need to compute the entire pairwise distance matrix in order to check for violations of proportionality, as we show in Section 4.

In the worst case, proportionality is incompatible with all three of the classic k-center, k-means, and k-median objectives; *i.e.*, there exist instances for which any proportional solution has an arbitrarily bad approximation to all objectives. We present such an instance in the the full version of this paper (Chen et al., 2019), and show in Section 5 that this behavior also arises in real-world datasets. Furthermore, as we show in Section 2 and observe empirically in Section 5, proportional solutions may not always exist. We therefore consider the natural approximate notion of proportionality that relaxes the Pareto dominance condition by a multiplicative factor.

Definition 2. *$X \subseteq \mathcal{M}$ with $|X| = k$ is ρ-**approximate proportional** (hereafter ρ-proportional) if $\forall S \subseteq \mathcal{N}$ with $|S| \geq \lceil \frac{n}{k} \rceil$ and for all $y \in \mathcal{M}$, there exists $i \in S$ with $\rho \cdot d(i, y) \geq D_i(X)$.*

1.2. Results and Outline

In Section 2 we show that proportional solutions may not always exist. In fact, one cannot get better than a 2-proportional solution in the worst case. In contrast, we give a greedy algorithm (Algorithm 1) and prove Theorem 1: The algorithm yields a $\left(1+\sqrt{2}\right)$-proportional solution in the worst case.

In Section 3, we treat proportionality as a constraint and seek to optimize the k-median objective subject to that constraint. We show how to write approximate proportionality as m linear constraints. Incorporating this into the standard linear programming relaxation of the k-median problem, we show how to use the rounding from (Charikar et al., 2002) to find an $O(1)$-proportional solution that is an $O(1)$-approximation to the k-median objective of the optimal proportional solution.

In Section 4, we show that proportionality is approximately preserved if we take a random sample of the data points of size $\tilde{O}(k^3)$, where the $\tilde{O}$ hides low order terms. This immediately implies that for constant k, we can check if a given clustering is proportional as well as compute approximately proportional solutions in near linear time, comparable to the time taken to run the classic k-means heuristic.

In Section 5, we provide a local search heuristic that efficiently searches for a proportional clustering. Our heuristic is able to consistently find nearly proportional solutions in practice. We test our heuristic and Algorithm 1 empirically against the celebrated k-means heuristic in order to understand the tradeoff between proportionality and the k-means objective. We find that the tradeoff is highly data dependent: Though these objectives are compatible on some datasets, there exist others on which these objectives are in conflict.

1.3. Related Work

Unsupervised Learning. Metric clustering is a well studied problem. There are constant approximation polynomial time algorithms for both the k-median (Jain & Vazirani, 1999; Charikar et al., 2002; Arya et al., 2004; Mettu & Plaxton, 2004; Byrka et al., 2017) and k-center objective (Gonzalez, 1985; Shmoys et al., 1997). Proportionality is a constraint on the *centers* as opposed to the data points; this makes it difficult to adapt standard algorithmic approaches for k-medians and k-means such as local search (Arya et al., 2004), primal-dual (Jain & Vazirani, 1999), and greedy dual fitting (Jain et al., 2002). For instance, our greedy algorithm in Section 2 grows balls around potential centers, which is very different from how balls are grown in the primal-dual schema (Jain & Vazirani, 1999; Mettu & Plaxton, 2004). Somewhat surprisingly, in Section 2 we show that for the problem of minimizing the k-median objective subject to proportionality as a constraint, we can extend the linear program rounding technique of (Charikar et al., 2002) to get a constant approximation algorithm. However, the additional constraints we add in the linear program formulation render the primal-dual and other methods inapplicable.

In (Chierichetti et al., 2017) and subsequent generalizations (Rösner & Schmidt, 2018; Bera et al., 2019), the authors consider fair clustering in terms of balance: There are red and blue points, and a balanced solution has roughly the same ratio of blue to red points in every cluster as in the overall population. The authors are motivated to extract features that cannot discriminate between status in different groups. This ensures that subsequent regression or classification on these features will be fair between these groups. In contrast, we assume that our data points prefer to be accurately clustered, and that an unfair solution provides accurate clusters for some groups, while giving other large groups low quality clusters. Finally, we note that there is a line of work in fair unsupervised learning concerned with constructing word embeddings that avoid bias (Bolukbasi et al., 2016; Caliskan et al., 2017), but these problems seem orthogonal to our concerns in clustering.

Supervised Learning. The standard model in fair supervised learning (Dwork et al., 2012; Kleinberg et al., 2016; Kleinberg et al., 2017; Zafar et al., 2017b;a) has a set of *protected agents* given as input to an algorithm which must classify agents into a positive and negative group. Most of these notions of fairness do not apply in any natural way to unsupervised learning problems. Our work further differs from the supervised learning literature in that we do not assume information about which agents are to be protected. Instead, we provide a fairness guarantee to arbitrary groups of agents, including protected groups even if we do not know their identity, similar to the ideas considered in (Kearns et al., 2018) and (Hashimoto et al., 2018).

Fair Resource Allocation. Our notion of proportionality is derived from the notion of core in economics (Scarf, 1967; Foley, 1970). The core has been adapted as a natural generalization to groups of the idea of fairness as proportionality (Fain et al., 2016; 2018), similar to other group fairness concepts for public goods that explicitly consider shared resources (Conitzer et al., 2017; Aziz et al., 2017). In clustering, the public goods are the centers themselves, and the “agents” are the data points, which share the centers. The fair clustering problem differs in that it is framed in terms of costs instead of positive utility, and agents only care about their most preferred good. That is, an agent’s cost for a clustering solution is just the distance to the closest center, as opposed to much of the previous resource allocation literature where agents have additive utility across the allocated goods. One can interpret our work as results for computing the core for a resource allocation problem where agents have a min-cost function with respect to allocations.

2. Existence and Computation of Proportional Solutions

We begin with a negative result: in the worst case, there may not be an exact proportional solution. Claim 1 is stated for arbitrary $\mathcal{N}$ and $\mathcal{M}$, but the impossibility remains even when $\mathcal{N} = \mathcal{M}$. We present both proofs in the full version of this paper (Chen et al., 2019). The idea is to create two groups of points very far away from one another with $k = 3$, ensuring that one group will be served by only one center.

Claim 1. *For all $\rho < 2$, a ρ-proportional solution is not guaranteed to exist.*

2.1. Computing a $(1 + \sqrt{2})$-Approximate Proportional Clustering

Claim 1 establishes that we should focus our attention on designing an efficient approximation algorithm. We give a simple and efficient algorithm that achieves a $\left(1 + \sqrt{2}\right)$-proportional solution, very close to the existential lower bound of 2. For notational ease, let $B(x, \delta) = \{i \in \mathcal{N} : d(i, x) \leq \delta\}$. That is, $B(x, \delta)$ is the ball (defined on $\mathcal{N}$) of distance δ about center x. For simplicity of exposition, we present Algorithm 1 as a *continuous* algorithm where a δ parameter is smoothly increasing. The algorithm can be easily discretized using priority queues.

Algorithm 1 Greedy Capture

```
$\delta \leftarrow 0$; $X \leftarrow \emptyset$; $N \leftarrow \mathcal{N}$
while $N \neq \emptyset$ do
    Smoothly increase $\delta$
    while $\exists x \in X$ s.t. $|B(x, \delta) \cap N| \geq 1$ do
        $N \leftarrow N \backslash B(x, \delta)$
    while $\exists x \in (\mathcal{M} \backslash X)$ s.t. $|B(x, \delta) \cap N| \geq \lceil \frac{n}{k} \rceil$ do
        $X \leftarrow X \cup \{x\}$
        $N \leftarrow N \backslash B(x, \delta)$
return X
```

Algorithm 1 runs in $\tilde{O}(mn)$ time.[1] In essence, the algorithm grows balls continuously around the centers, and when the ball around a center has "captured" $\lceil \frac{n}{k} \rceil$ points, we greedily open that center and disregard all of the captured points. Open centers continue to greedily capture points as their balls continue to expand. Though (Jain & Vazirani, 1999; Mettu & Plaxton, 2004) similarly expand balls about points to compute approximately optimal solutions to the k-median problem, there is a crucial difference: They grow balls around data points rather than centers.

Theorem 1. *Algorithm 1 yields a $(1 + \sqrt{2})$-proportional clustering, and there exists an instance for which this bound is tight.*

Proof. Let X be the solution computed by Algorithm 1. First note that X uses at most k centers, since it only opens a center when $\lceil \frac{n}{k} \rceil$ *unmatched* points are absorbed by the ball around that center, and this can happen at most k times. Now, suppose for a contradiction that X is not a $(1 + \sqrt{2})$-proportional clustering. Then there exists $S \subseteq \mathcal{N}$ with $|S| \geq \lceil \frac{n}{k} \rceil$ and $y \in \mathcal{M}$ such that

$$\forall i \in S, \ (1 + \sqrt{2}) \cdot d(i, y) < D_i(X). \tag{1}$$

Let r_y be the distance of the farthest agent from y in S, that is, $r_y := \max_{i \in S} d(i, y)$, and call this agent i^*. There are two cases. In the first case, $B(x, r_y) \cap S = \emptyset$ for all $x \in X$. This immediately yields a contradiction, because it implies that Algorithm 1 would have opened y. In particular, note that $S \subseteq B(y, r_y)$, so if $S \cap B(x, r_y) = \emptyset$ for all $x \in X$, then $B(y, r_y)$ would have had at least $\lceil \frac{n}{k} \rceil$ unmatched points.

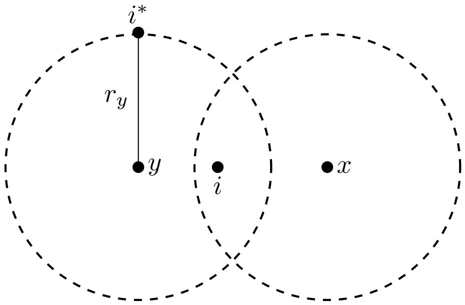


Figure 1. Diagram for Proof of Theorem 1

In the second case, $\exists x \in X$ and $\exists i \in N$ such that $i \in B(x, r_y) \cap S$. This case is drawn below in Figure 1. By the triangle inequality, $d(x, y) \leq d(i, x) + d(i, y)$. Therefore, $d(i^*, x) \leq r_y + d(i, x) + d(i, y)$. Also, $d(i, x) \leq r_y$, since $i \in B(x, r_y)$. Consider the minimum multiplicative improvement of i and i^*:

$$\begin{aligned}
&\min\left(\frac{d(i,x)}{d(i,y)}, \ \frac{d(i^*,x)}{d(i^*,y)}\right) \\
&\leq \min\left(\frac{d(i,x)}{d(i,y)}, \ \frac{r_y + d(i,x) + d(i,y)}{r_y}\right) \\
&\leq \min\left(\frac{r_y}{d(i,y)}, \ 2 + \frac{d(i,y)}{r_y}\right) \\
&\leq \max_{z \geq 0} \left(\min\left(z, \ 2 + 1/z\right)\right) = 1 + \sqrt{2}
\end{aligned}$$

which violates equation 1. It is not hard to show that there exists an instance for which Algorithm 1 yields exactly this bound, and we present this example in the full version of this paper (Chen et al., 2019). □

[1] To state running times simply, we use the convention that $f(n)$ is $\tilde{O}(g(n))$ if $f(n)$ is $O(g(n))$ up to poly-logarithmic factors.

2.2. Local Capture Heuristic

We observe that while our Greedy Capture algorithm (Algorithm 1) always produces an approximately proportional solution, it may not produce an exactly proportional solution in practice, even on instances where such solutions exist (see Figure 3a and Figure 3b). We therefore introduce a Local Capture heuristic for searching for more proportional clusterings. Algorithm 2 takes a target value of ρ as a parameter, and proceeds by iteratively finding a center that violates ρ-fairness and swapping it for the center in the current solution that is least demanded.

Algorithm 2 Local Capture Heuristic

```
input $\rho$
Initialize $X$ as a random subset of $k$ centers from $\mathcal{M}$.
repeat
    for $y \in \mathcal{M}$ do
        $S_y \leftarrow \{i \in N : \rho \cdot d_{iy} < D_i(X)\}$
        if $|S_y| \geq \lceil \frac{n}{k} \rceil$ then
            $x^* \leftarrow \text{argmin}_{x \in X} |\{i \in N : d_{ix} = D_i(X)\}|$
            $X \leftarrow (X \backslash \{x^*\}) \cup \{y\}$
until no changes occur
return X
```

Every iteration of Algorithm 2 (the entire inner for loop) runs in $\tilde{O}(mn^2)$ time. There is no guarantee of convergence (for a given input ρ, there may not even exist a ρ-proportional solution), but if Algorithm 2 terminates, then it returns a ρ-proportional solution. In our experiments (see Section 5), we search for the minimum ρ for which the algorithm terminates in a small number of iterations via binary search over possible input of ρ. In (Arya et al., 2004), the authors also evaluate a local search swapping procedure for the k-median problem, but their swap condition is based on the relative k-median objective of two solutions, whereas our swap condition is based on violations to proportionality.

3. Proportionality as a Constraint

One concern with the previous algorithms is that they may find a proportional clustering with poor global objective (e.g., k-median), even when exact proportional clusterings with good global objectives exist. For example, suppose $k = 2$ and there are two easily defined clusters, containing 40% and 60% of the data respectively. It is possible that Algorithm 1 will only open centers inside of the larger cluster. This is proportional, but undesirable from an optimization perspective (note that the "correct" clustering of such an example is still proportional). Here, we show how to address this concern by optimizing the k-median objective subject to proportionality as a constraint. Later, in Section 5, we empirically study the tradeoff between the k-means objective and proportionality on real data.

$$
\begin{aligned}
\text{Minimize} \quad & \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{M}} d(i,j) z_{ij} & & (2) \\
\text{Subject to} \quad & \sum_{j \in \mathcal{M}} z_{ij} = 1 & \forall i \in \mathcal{N} \quad & (3) \\
& z_{ij} \leq y_j & \forall j \in \mathcal{M}, \forall i \in \mathcal{N} \quad & (4) \\
& \sum_{j \in \mathcal{M}} y_j \leq k & & (5) \\
& \sum_{j' \in B(j, \gamma R_j)} y_{j'} \geq 1 & \forall j \in \mathcal{M} \quad & (6) \\
& z_{ij}, y_j \in [0,1] & \forall j \in \mathcal{M}, \forall i \in \mathcal{N} \quad & (7)
\end{aligned}
$$

Figure 2. Proportional k-median Linear Program

We consider the k-median and k-means objectives to be reasonable measures of the global quality of a solution. We see minimizing the k-center objective more as a competing notion of fairness, and so we focus on optimizing the k-median objective subject to proportionality.[2] Minimizing the k-median objective without proportionality is a well studied problem in approximation algorithms, and several constant approximations are known (Charikar et al., 2002; Arya et al., 2004; Mettu & Plaxton, 2004). Most of this work is in the model where $\mathcal{N} \subseteq \mathcal{M}$, and we follow suit in this section. We show the following.

Theorem 2. *Suppose there is a ρ-proportional clustering with k-median objective c. In polynomial time in m and n, we can compute a $O(\rho)$-proportional clustering with k-median objective at most $8c$.*

In particular, we can compute a constant approximate proportional clustering with k-median objective at most eight times the minimum k-median objective proportional clustering. Note that the exact running time will depend on the algorithm used to solve the linear program. In the remainder of this section, we will sketch the proof of Theorem 2. We begin with the standard linear programming relaxation of the k-median minimization problem, and then add a constraint to encode proportionality. The final linear program is shown in Figure 2. Recall that $B(x, \delta) = \{i \in \mathcal{N} : \ d(i,x) \leq \delta\}$.

In the LP, z_{ij} is an indicator variable equal to 1 if $i \in \mathcal{N}$ is matched to $j \in \mathcal{M}$. y_j is an indicator variable equal to 1 if $j \in X$, i.e., if we want to use center $j \in \mathcal{M}$ in our clustering. Objective 2 is the k-median objective. Constraint 3 requires that every point be matched, and constraint 4 only allows a point to be matched to an open center. Constraint 5 allows

[2] A constant approximation algorithm for minimizing the k-median objective immediately implies a constant approximation algorithm for minimizing the k-means objective by running the algorithm on the squared distances (Mettu & Plaxton, 2004).

at most k centers to be opened, and constraint 7 relaxes the indicator variables to real values between 0 and 1.

Constraint 6 is the new constraint that we introduce. Our crucial lemma argues that constraint 6 approximately encodes proportionality. Let R_j be the minimum value such that $|B(j, R_j)| \geq \lceil \frac{n}{k} \rceil$. In other words, R_j is the distance of the $\lceil \frac{n}{k} \rceil$ farthest point in $\mathcal{N}$ from j.

Lemma 1. *Let X be a clustering, and let $\gamma \geq 1$. If $\forall j \in \mathcal{M}$ there exists some $x \in X$ such that $d(j, x) \leq \gamma R_j$, then X is $(1 + \gamma)$-proportional. If X is γ-proportional, then $\forall j \in \mathcal{M}$ there exists some $x \in X$ such that $d(j, x) \leq (1 + \gamma)R_j$.*

Proof. Suppose that $\forall j \in \mathcal{M}$ there exists some $x \in X$ such that $d(j, x) \leq \gamma R_j$. Suppose for a contradiction that X is not $(1 + \gamma)$-proportional. Then there exists $S \subseteq \mathcal{N}$ with $|S| \geq \lceil \frac{n}{k} \rceil$ and $j \in \mathcal{M}$ such that $\forall i \in S$, $(1+\gamma) \cdot d(i, j) < D_i(X)$. By assumption, $\exists x \in X$ such that $d(j, x) \leq \gamma R_j$, so by the triangle inequality $D_i(X) \leq d(i, j) + d(j, x) \leq d(i, j) + \gamma R_j$. Therefore, $\forall i \in S$, $\gamma \cdot d(i, j) < D_i(X) - d(i, j) \leq \gamma R_j$. However, by definition of R_j, since $|S| = \lceil \frac{n}{k} \rceil$, there must exist some $i \in S$ such that $d(i, j) \geq R_j$.

Suppose that X is γ-proportional. Let $j \in \mathcal{M}$. Consider the set S of the closest $\lceil \frac{n}{k} \rceil$ points in $\mathcal{N}$ to j. By definition of proportionality $\exists i \in S$ and $x \in X$ such that $\gamma d(i, j) \geq d(i, x)$. Therefore, by the triangle inequality, $d(j, x) \leq d(i, j) + d(i, x) \leq (1 + \gamma)d(i, j)$. By definition of S, $d(i, j) \leq R_j$, so there exists $x \in X$ such that $d(j, x) \leq (1 + \gamma)R_j$. □

Now, suppose there is a ρ-proportional clustering X with k-median objective c. Then we write the linear program shown in Figure 2 with $\gamma = \rho + 1$ in constraint 6. Lemma 1 guarantees that X is feasible for the resulting linear program, so the fractional solution has k-median objective at most c. We then round the resulting fractional solution. In (Charikar et al., 2002), the authors give a rounding algorithm for the the linear program in Figure 2 without Constraint 6. We show that a slight modification to this rounding algorithm also preserves Constraint 6 to a constant approximation.

Lemma 2. *(Proved in Full Version (Chen et al., 2019)) Let $\{y_j\}, \{z_{ij}\}$ be a fractional solution to the linear program in Figure 2. Then there is an integer solution $\{\hat{y_j}\}, \{\hat{z_{ij}}\}$ that is an 8-approximation to the objective, and that opens k centers. Furthermore, for all $j \in \mathcal{M}$, $\sum_{j' \in B(j, 27\gamma R_j)} \hat{y_{j'}} \geq 1$.*

Given Lemma 2, applying Lemma 1 again implies that the result of the rounding is $(27(1 + \rho) + 1)$-proportional, since we set $\gamma = 1 + \rho$. Since the k-median objective of the fractional solution is at most c, the fact that the k-median objective of the rounded solution is at most $8c$ follows directly from the proof from (Charikar et al., 2002). We note that the constant factor of 27 can be improved to 13 in the special case where $\mathcal{N} = \mathcal{M}$. Interestingly, the ostensibly similar primal-dual approach of (Jain & Vazirani, 1999) does not appear amenable to the added constraint of proportionality (in particular, the reduction to facility location from (Jain & Vazirani, 1999) is no longer straightforward).

4. Sampling for Linear-Time Implementations and Auditing

In this section, we study proportionality under uniform random sampling (i.e., draw $|N|$ individuals i.i.d. from the uniform distribution on $\mathcal{N}$). In particular, we show that proportionality is well preserved under random sampling. This allows us to design efficient implementations of Algorithm 1 and Algorithm 2, and to introduce an efficient algorithm for auditing proportionality. We first present the general property and then demonstrate its various applications.

4.1. Proportionality Under Random Sampling

For any $X \subseteq \mathcal{M}$ of size k and center $y \in \mathcal{M}$, define $R(\mathcal{N}, X, y) = \{i \in \mathcal{N} : D_i(X) > \rho \cdot d(i, y)\}$. Note that solution X is not ρ-proportional with respect to $\mathcal{N}$ if and only if there is some $y \in \mathcal{M}$ such that $\frac{|R(\mathcal{N},X,y)|}{|\mathcal{N}|} \geq \frac{1}{k}$. A random sample approximately preserves this fraction for all solutions X and deviating centers y. The following theorem is a consequence of Hoeffding's inequality, and we present a brief proof in the full version of this paper (Chen et al., 2019). The important idea in the proof is that we take a union bound over all possible solutions and deviations, and there are only $k\binom{m}{k}$ such combinations.

Theorem 3. *Given $\mathcal{N}$, $\mathcal{M}$ and parameter $\rho \geq 1$, fix parameters $\epsilon, \delta \in [0, 1]$. Let $N \subseteq \mathcal{N}$ of size $\Omega\left(\frac{k^3}{\epsilon^2} \log \frac{m}{\delta}\right)$ be chosen uniformly at random. Then, with probability at least $1 - \delta$, the following holds for all (X, y):*

$$\left| \frac{|R(N, X, y)|}{|N|} - \frac{|R(\mathcal{N}, X, y)|}{|\mathcal{N}|} \right| \leq \frac{\epsilon}{k}$$

In order to apply the above theorem, we say that a solution X is ρ-proportional to $(1 + \epsilon)$-deviations if for all $y \in \mathcal{M}$ and for all $S \subseteq \mathcal{N}$ where $|S| \geq (1 + \epsilon)\frac{n}{k}$, there exists some $i \in S$ such that $\rho \cdot d(i, y) \geq D_i(X)$. Note that if X is ρ-proportional to 1-deviations, it is simply ρ-proportional. We immediately have the following:

Corollary 1. *Let $N \subset \mathcal{N}$ be a uniform random sample of size $|N| = \Omega\left(\frac{k^3}{\epsilon^2} \ln \frac{m}{\delta}\right)$. Suppose $X \subseteq M$ with $|X| = k$ is ρ-proportional with respect to N. Then with probability at least $1 - \delta$, X is ρ-proportional to $(1 + \epsilon)$-deviations with respect to $\mathcal{N}$.*

4.2. Linear Time Implementation

We now consider how to take advantage of Theorem 3 to optimize Algorithm 1 and Algorithm 2. First, note that

Algorithm 1 takes $\tilde{O}(mn)$ time, which is quadratic in input size. A corollary of Theorem 3 is that we can approximately implement Algorithm 1 in nearly linear time, comparable to the running time of the standard k-means heuristic.

Corollary 2. *Algorithm 1, when run on $\mathcal{M}$ and a random sample $N \subseteq \mathcal{N}$ of size $|N| = \tilde{\Theta}\left(\frac{k^3}{\epsilon^2}\right)$, provides a solution that is $(1+\sqrt{2})$-proportional to $(1+\epsilon)$-deviations with high probability in $\tilde{O}\left(\frac{k^3}{\epsilon^2}m\right)$ time.*

We also get a substantial speedup for our Local Capture algorithm. Recall that Local Capture (Algorithm 2) is an iterative algorithm that takes a target value of ρ as a parameter, and if it converges, returns a ρ-proportional clustering. Without sampling, each iteration of Algorithm 2 takes $\tilde{O}(mn^2)$ time. Another corollary of Theorem 3 is that it is sufficient to run the Local Capture on a random sample of k^3/ϵ^2 out of the n points in $\mathcal{N}$ in order to search for a clustering that is ρ-proportional with respect to $(1+\epsilon)$-deviations.

4.3. Efficient Auditing

Alternatively, one might still want to run a non-proportional clustering algorithm, and ask whether the solution produced happens to be proportional. We call this the *Audit Problem*. Given $\mathcal{N}, \mathcal{M}$, and $X \subset \mathcal{N}$ with $|X| \leq k$, find the minimum value of ρ such that X is ρ-proportional. It is not too hard to see that one can solve the Audit Problem exactly in $O((k+m)n)$ time by computing for each $y \in \mathcal{M}$, the quantity ρ_y, the $\lceil \frac{n}{k} \rceil$ largest value of $\frac{D_i(X)}{d(i,y)}$. We subsequently find the y that maximizes ρ_y. Again, this takes quadratic time, which can be worse than the time taken to find the clustering itself.

Consider a slightly relaxed (ϵ, δ)-Audit Problem where we are asked to find the minimum value of ρ such that X is ρ-proportional to $(1+\epsilon)$-deviations with probability at least $1-\delta$. This problem can be efficiently solved by using a random sample $N \subseteq \mathcal{N}$ of points to conduct the audit.

Corollary 3. *The (ϵ, δ)-Audit Problem can be solved in $\tilde{O}\left((k+m)\frac{k^3}{\epsilon^2}\right)$ time.*

5. Implementations and Empirical Results

In this section, we study proportionality on real data taken from the UCI Machine Learning Repository (Dheeru & Karra Taniskidou, 2017). We consider three qualitatively different data sets used for clustering: Iris, Diabetes, and KDD. For each data set, we only have a single set of points given as input, so we take $\mathcal{N} = \mathcal{M}$ to be the set of all points in the data set. We use the standard Euclidean L2 distance.

Iris. This data set contains information about the petal dimensions of three different species of iris flowers. There are 50 samples of each species.

Diabetes. The Pima Indians Diabetes data set contains information about 768 diabetes patients, recording features like glucose, blood pressure, age and skin thickness.

KDD. The KDD cup 1999 data set contains information about sequences of TCP packets. Each packet is classified as normal or one of twenty-two types of intrusions. Of these 23 classes, normal, "neptune", and "smurf" account for 98.3% of the data. The data set contains 18 million samples; we work with a subsample of 100,000 points.[3]

5.1. Proportionality and k-means Objective Tradeoff

We compare Greedy Capture (Algorithm 1) and Local Capture (Algorithm 2) with the k-means++ algorithm (Lloyd's algorithm for k-means minimization with the k-means++ initialization (Arthur & Vassilvitskii, 2007)) for a range of values of k. For the Iris data set, Local Capture and k-means++ always find an exact proportional solution (Figure 3a), and have comparable k-means objectives (Figure 4a). The Iris data set is very simple with three natural clusters, and validates the intuition that proportionality and the k-means objective are not always opposed.

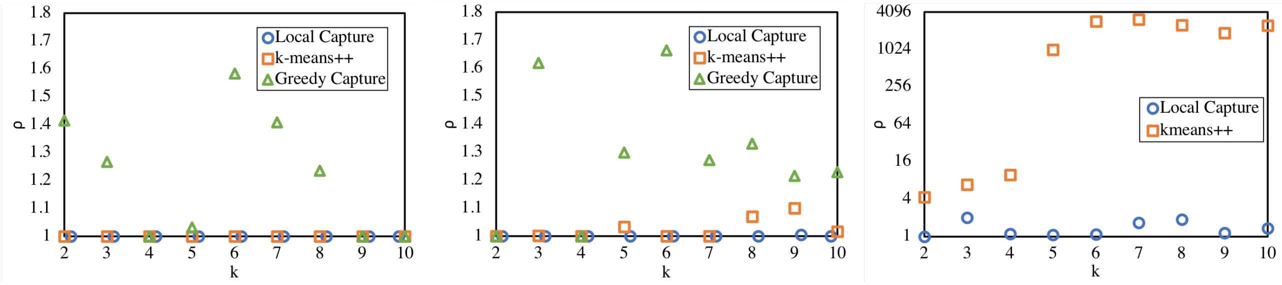
The Diabetes data set is larger and more complex. As shown in Figure 3b, k-means++ no longer always finds an exact proportional solution. Local Capture always finds a better than 1.01-proportional solution. As shown in Figure 4b, the k-means objectives of the solutions are separated, although generally on the same order of magnitude.

For the KDD data set, proportionality and the k-means object appear to be in conflict. Greedy Capture's performance is comparable to Local Capture on KDD, so we omit it for clarity. In Figures 3c and 4c, note that the gap between ρ and the k-means objective for the k-means++ and Local Capture algorithms is between three and four orders of magnitude. We suspect this is due to the presence of significant outliers in the KDD data set. This is in keeping with the theoretical impossibility of simultaneously approximating the optima on both objectives, and demonstrates that this tension arises in practice as well as theory.

5.2. Proportionality and Low k-means Objective

Note that if one is allowed to use $2k$ centers when k is given as input, one can trivially achieve the proportionality of Local Capture and the k-means objective of the k-means++

[3] We run k-means++ on this entire 100,000 point sample. For efficiency, we run our Local Capture algorithm by further sampling 5,000 points uniformly at random to treat as $\mathcal{N}$ and sampling 400 points via the k-means++ initialization to treat as $\mathcal{M}$. For the sake of a fair comparison, we generate a different sample of 400 centers using the k-means++ initialization that we use to determine the value of ρ we report for both Local Capture and the k-means++ algorithm. The k-means objective is measured on the original 100,000 points for both algorithms.



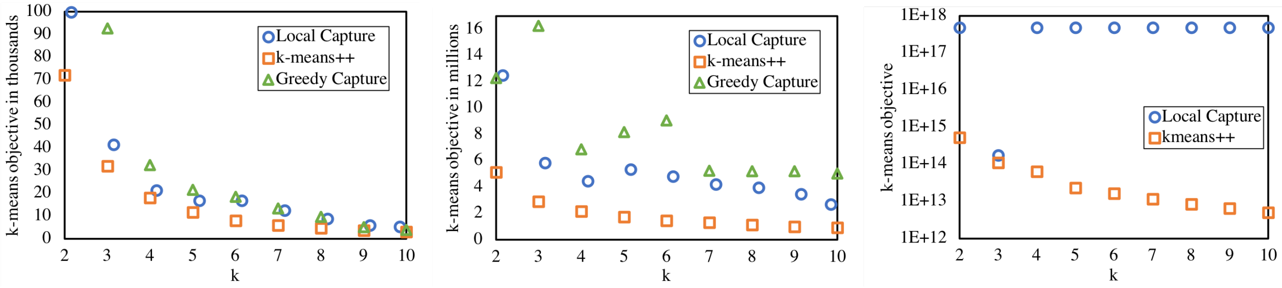
(a) Iris

(b) Diabetes

(c) KDD, geometric scale

Figure 3. Minimum ρ such that the solution is ρ-proportional

.



(a) Iris

(b) Diabetes

(c) KDD, geometric scale

Figure 4. k-means objective

algorithm by taking the union of the two solutions. Thinking in this way leads to a different way of quantifying the tradeoff between proportionality and the k-means objective: Given an approximately proportional solution, how many *extra* centers are necessary to get comparable k-means objective as the k-means++ algorithm? For a given data set, the answer is a value between 0 and k, where larger numbers indicate more incompatibility, and lower numbers indicate less incompatibility.

To answer this question, we compute the union of centers found by Local Capture and the k-means++ algorithm. We then greedily remove centers as long as doing so does not increase the minimum ρ such that the solution is ρ-proportional (defined on k, not $2k$) by more than a multiplicative factor of α, and does not increase the k-means objective by more than a multiplicative factor β.

On the KDD dataset, we set $\alpha = 1.2$ and $\beta = 1.5$, so the proportionality of the result is within 1.2 of Local Capture in Figure 3c, and the k-means objective is within 1.5 of k-means++ in Figure 4c. We observe that this heuristic uses at most 3 extra centers for any $k \leq 10$. So while there is real tension between proportionality and the k-means objective, this tension is still not maximal. In the worst case, one might need to add k centers to a proportional solution to compete with the k-means objective of the k-means++ algorithm, but in practice we find that we need at most 3 for $k \leq 10$.

6. Conclusion and Open Directions

We have introduced proportionality as a fair solution concept for centroid clustering. Although exact proportional solutions may not exist, we gave efficient algorithms for computing approximate proportional solutions, and considered constrained optimization and sampling for further applications. Finally, we studied proportionality on real data and observed a data dependent tradeoff between proportionality and the k-means objective. While this tradeoff is in some sense a negative result, it also demonstrates that proportionality as a fairness guarantee matters in the sense that it meaningfully constrains the space of solutions.

We have shown that ρ-proportional solutions need not exist for $\rho < 2$, and always exist for $\rho \geq 1 + \sqrt{2}$. Closing this approximability gap is one outstanding question. Another is whether there is a more efficient and easily interpretable algorithm for optimizing total cost subject to proportionality, as our approach in Section 3 requires solving a linear program on the entire data set. We would ideally like a more efficient and easily interpretable primal-dual or local search type algorithm. More generally, what other fair solution concepts for clustering should be considered alongside proportionality, and can we characterize their relative advantages and disadvantages? Finally, can the idea of proportionality as a group fairness concept be adapted for supervised learning tasks like classification and regression?

Acknowledgements

Brandon Fain is supported by NSF grants CCF-1408784 and IIS-1447554. Kamesh Munagala is supported by NSF grants CCF-1408784, CCF-1637397, and IIS-1447554; and research awards from Adobe and Facebook.

References

Arthur, D. and Vassilvitskii, S. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1027–1035, 2007.

Arya, V., Garg, N., Khandekar, R., Meyerson, A., Munagala, K., and Pandit, V. Local search heuristics for k-median and facility location problems. *SIAM Journal on Computing*, 33(3):544–562, 2004.

Aziz, H., Brill, M., Conitzer, V., Elkind, E., Freeman, R., and Walsh, T. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 48(2): 461–485, 2017.

Bera, S. K., Chakrabarty, D., and Negahbani, M. Fair Algorithms for Clustering. *arXiv e-prints*, art. arXiv:1901.02393, Jan 2019.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 4349–4357. 2016.

Byrka, J., Pensyl, T., Rybicki, B., Srinivasan, A., and Trinh, K. An improved approximation for k-median and positive correlation in budgeted optimization. *ACM Transactions on Algorithms (TALG)*, 13(2):23:1–23:31, 2017.

Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Charikar, M., Guha, S., va Tardos, and Shmoys, D. B. A constant-factor approximation algorithm for the k-median problem. *Journal of Computer and System Sciences*, 65 (1):129 – 149, 2002.

Chen, X., Fain, B., Lyu, C., and Munagala, K. Proportionally Fair Clustering. *arXiv e-prints*, art. arXiv:1905.03674, 2019.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 5029–5037. 2017.

Conitzer, V., Freeman, R., and Shah, N. Fair public decision making. In *Proceedings of the 2017 ACM Conference on Economics and Computation (EC)*, pp. 629–646, 2017.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 797–806, 2017.

Dheeru, D. and Karra Taniskidou, E. UCI machine learning repository, 2017.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 214–226, 2012.

Fain, B., Goel, A., and Munagala, K. The core of the participatory budgeting problem. In *Proceedings of the 12th International Conference on Web and Internet Economics (WINE)*, pp. 384–399, 2016.

Fain, B., Munagala, K., and Shah, N. Fair allocation of indivisible public goods. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, pp. 575–592, 2018.

Foley, D. K. Lindahl's solution and the core of an economy with public goods. *Econometrica*, 38(1):66–72, 1970.

Garg, N., Goel, A., and Plaut, B. Markets for Public Decision-making. *arXiv e-prints*, art. arXiv:1807.10836, July 2018.

Goel, N., Yaghini, M., and Faltings, B. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of 2018 AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3029–3036, 2018.

Gonzalez, T. F. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293 – 306, 1985.

Hardt, M., Price, E., , and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 3315–3323. 2016.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1929–1938, 2018.

Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651 – 666, 2010.

Jain, K. and Vazirani, V. V. Primal-dual approximation algorithms for metric facility location and k-median problems. In *40th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 2–13, 1999.

Jain, K., Mahdian, M., and Saberi, A. A new greedy approach for facility location problems. In *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, pp. 731–740, 2002.

Julia Angwin, Jeff Larson, S. M. and Lauren Kirchner, P. Machine bias, 2016.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2569–2577, 2018.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. *ArXiv e-prints*, 2016.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human decisions and machine predictions. Working Paper 23180, National Bureau of Economic Research, 2017.

Mettu, R. R. and Plaxton, C. G. Optimal time bounds for approximate clustering. *Machine Learning*, 56(1-3):35–60, 2004.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 5680–5689. 2017.

Rösner, C. and Schmidt, M. Privacy Preserving Clustering with Constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, pp. 96:1–96:14, 2018.

Scarf, H. E. The core of an n person game. *Econometrica*, 35(1):pp. 50–69, 1967.

Shmoys, D. B., Tardos, E., and Aardal, K. Approximation algorithms for facility location problems. In *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing (STOC)*, pp. 265–274, 1997. ISBN 0-89791-888-6. doi: 10.1145/258533.258600.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, pp. 1171–1180, 2017a.

Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., and Weller, A. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 229–239. 2017b.