

Fair and Representative Clustering for Data Summarization

Ashwin Singh

TU Wien
Vienna, Austria
ashwin.singh@tuwien.ac.at

Abstract

Clustering is widely used in data summarization systems such as web search and personalization to produce representations consumed by society at scale. These systems typically partition data into k clusters following an objective function, and then select one representative from each cluster. Under current fairness notions, a data summary is deemed fair if it contains a sufficient number of representatives, or cluster centers, from each demographic group. I argue that such criteria are insufficient to ensure that a clustering is representative and fail to account for the context underlying data summarization applications. To address this, I propose a more expressive fairness notion for clustering, for which I develop approximation algorithms with provable representation guarantees.

Introduction

Data summarization aims to provide a small representative subset of a dataset, most commonly by clustering and returning the cluster centers. Consider the example of searching for images of scientists, and the search results (*cluster centers*) include pictures of nine men and one woman. Such a data summary is problematic because it reinforces the association between gender and profession. To address this, Kleindessner, Awasthi, and Morgenstern (2019) first proposed the FAIR k -CENTER problem that selects a minimum number of centers per group. Following this definition, a summary that contains an equal number of search results for men and women may be considered fair.

However, this can lead to a poor representation for both men and women in terms of reporting aggregate cluster statistics: it is possible that all centers for men may be selected from clusters whose members are predominantly women (and vice versa). For instance, clusters with men as centers may have much lower median income than clusters with women as centers, even when this disparity is the opposite. Therefore, a fair clustering that only takes into account, the number of centers per group, can fail to faithfully represent the very groups it is supposed to be “fair” for. Figure 1 illustrates this limitation with a toy example of a data summary with $k = 2$ clusters and $m = 2$ groups.

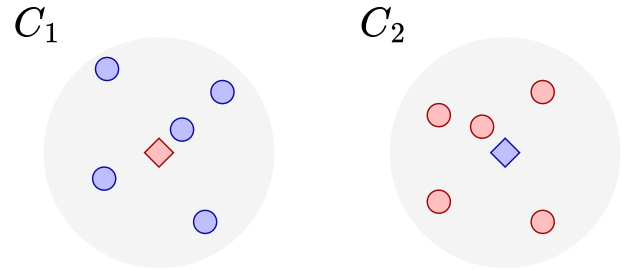


Figure 1: A data summary with $k = 2$ clusters. Circles ($\circ, \bullet$) denote data points whereas diamonds ($\diamond, \blacklozenge$) denote the selected cluster centers. Colors encode group membership. Both groups have an equal number of centers but each center represents more points from the other group than its own.

Instead, I propose that a clustering is fair if, for each group, at least an α -fraction of its members are represented by centers from the same group. For instance, if both centers in the illustration in Figure 1 are replaced by any other point in their respective clusters, 80% of members from each group would be represented by a same-group center. Formalizing this notion is only the first step, as fair clustering with representation raises two types of challenges.

Contextual Challenges. What is the minimum α for a clustering to be considered representative? What is the maximum permissible relaxation of the clustering objective to achieve an α representation? Answers to these questions are context-dependent, and require careful analysis of data summarization applications. In scenarios where cluster-membership is low-stakes (e.g. music recommendations), it may be acceptable to substantially relax the clustering objective in favor of fairness. In high-stakes scenarios, however (e.g. loan approval), the maximum permissible relaxation for the clustering objective to attain fairness may be limited.

Theoretical Challenges. Standard clustering objectives (k -center, k -median, and k -means) are already NP-hard, and imposing representation constraints makes the resulting optimization problem even less tractable.

Together, these challenges motivate the central research questions in my current work, which I describe below:

RQ1. What notions of fairness are appropriate for evaluating group representation in clustering, and when are they important? Maximizing α is the same as maximizing the minimum representation of any group, which is inspired by Rawls’ veil of ignorance (1971). However, this is only one possible criterion, and the literature on fair allocation and computational social choice offers several alternatives, such as *envy-freeness* and *equity*, which may be more appropriate in other settings (e.g., committee selection).

RQ2. Given an appropriate fair representation criterion, how can we design efficient algorithms with provable guarantees for the following two problems?

1. What is the maximum achievable representation (α) for a δ -additive relaxation of the clustering objective?
2. What is the minimum additive relaxation (δ) of the clustering objective required to satisfy an α representation.

Preliminaries and Technical Notation

In this section, I establish the notation needed to describe my work in more detail. Consider a metric space (X, d) with distance function $d: X \times X \rightarrow \mathbb{R}_{\geq 0}$. Let $\mathcal{G} = \{G_1, \dots, G_m\}$ be a partition of X into m groups. Given a set of k cluster centers $C = \{c_1, \dots, c_k\} \subseteq X$, let $\phi: X \rightarrow C$ assign each point $\mathbf{x} \in X$ to its nearest center in C . For each $c \in C$, define $P(c) = \{\mathbf{x} \in X : \phi(\mathbf{x}) = c\}$, the set of points assigned to c . Then, the *representation function* $f: C \times \mathcal{G} \rightarrow [0, 1]$ is given by $f(c, G) = |P(c) \cap G|/|G|$ and denotes the fraction of points of group G represented by cluster center c . The cluster centers selected from group G are given by the *allocation function* $\mathcal{A}(G) = C \cap G$.

Contributions and Ongoing Work

I first provide a formal definition for my proposed notion for fair clustering, with the k -center objective.

Definition 1 (α -REPRESENTATIVE k -CENTER). *Given a minimum representation constraint $\alpha \in [0, 1]$, a set of k centers $C \subseteq X$ is a feasible solution if it satisfies*

$$\min_{G \in \mathcal{G}} \sum_{c \in \mathcal{A}(G)} f(c, G) \geq \alpha.$$

The objective is to find a feasible solution that minimizes the k -center cost: $\max_{\mathbf{x} \in X} d(\mathbf{x}, \phi(\mathbf{x}))$.

Methods for solving the FAIR k -CENTER problem typically begin by applying the FFT algorithm of Gonzalez (1985), which computes a 2-approximation for the k -center objective. In the next step, they modify the selected cluster centers to obtain a feasible solution while incurring an additive loss to the approximation guarantee; see for e.g., Gaddekar, Gionis, and Thejaswi (2025). As my first contribution, I generalize this approach into a framework that helps me address **RQ2** for any fair representation criterion, including the α -REPRESENTATIVE k -CENTER problem. Next, I briefly describe this framework.

The key idea is to pick new centers within distance δ of the old centers, while preserving the original clustering. Then, for a cluster with old center c and new center c' , we can

say the following for any point $\mathbf{x} \in P(c)$ using the triangle inequality:

$$d(\mathbf{x}, c') \leq d(\mathbf{x}, c) + d(c, c') \leq 2\text{OPT} + \delta$$

where the final inequality uses the fact that the original centers form a 2-approximation. This construction yields a δ -additive relaxation in the clustering objective. Applying binary search to δ , we can answer **RQ2.2**.

Similarly, fixing δ restricts the candidates for new cluster centers to points $\mathbf{x} \in P(c)$ for which $d(\mathbf{x}, c) \leq \delta$. For a fixed δ , maximizing α admits a direct reduction to the Santa Claus problem (Bansal and Sviridenko 2006), where clusters are items, groups are agents, and representation function f defines the agents’ utilities for the items. This reduction allows us to address **RQ2.1**, and establishes a novel link between fair clustering and computational social choice. In particular, it enables us to borrow notions from the fair allocation of indivisible goods to reason about which representation guarantees may be more important in different data summarization applications.

In addition, we have developed an algorithm that computes an ϵk -additive approximation for **RQ2.1** in time $(1/\epsilon)^{O(m)}$, and hence runs in $\text{poly}(1/\epsilon)$ for a constant number of groups m . Our ongoing work focuses on extending these results to regimes where (i) the number of groups are not constant and can intersect, and (ii) different fair representation criteria are more appropriate (**RQ1**).

Acknowledgments

This research has been funded by the Vienna Science and Technology Fund (WWTF) [Grant ID: 10.47379/VRG23013].

References

- Bansal, N.; and Sviridenko, M. 2006. The Santa Claus problem. In *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, STOC ’06, 31–40. New York, NY, USA: Association for Computing Machinery. ISBN 1595931341.
- Gaddekar, A.; Gionis, A.; and Thejaswi, S. 2025. Fair Clustering for Data Summarization: Improved Approximation Algorithms and Complexity Insights. In *Proceedings of the ACM on Web Conference 2025*, WWW ’25, 4458–4469. New York, NY, USA: Association for Computing Machinery. ISBN 9798400712746.
- Gonzalez, T. F. 1985. Clustering to minimize the maximum intercluster distance. *Theoretical computer science*, 38: 293–306.
- Kleindessner, M.; Awasthi, P.; and Morgenstern, J. 2019. Fair k -Center Clustering for Data Summarization. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 3448–3457. PMLR.
- Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press.