Proportional Representation in Metric Spaces and Low-Distortion Committee Selection

Yusuf Kalayci[1], **David Kempe**[1], **Vikram Kher**[2]

[1]University of Southern California
[2]Yale University *
kalayci@usc.edu, David.M.Kempe@Gmail.com, vikram.kher@yale.edu

Abstract

We introduce a novel definition for a small set R of k points being *representative* of a larger set in a metric space. Given a set V (e.g., documents or voters) to represent, and a set C of possible representatives, our criterion requires that for any subset S comprising a θ fraction of V, the average distance of S to their best $\theta \cdot k$ points in R should not be more than a factor γ compared to their average distance to the best $\theta \cdot k$ points among all of C. This definition is a strengthening of proportional fairness and core fairness, but - different from those notions - requires that large cohesive clusters be represented proportionally to their size.

Since there are instances for which - unless γ is polynomially large - no solutions exist, we study this notion in a resource augmentation framework, implicitly stating the constraints for a set R of size k as though its size were only k/α, for $\alpha > 1$. Furthermore, motivated by the application to elections, we mostly focus on the *ordinal* model, where the algorithm does not learn the actual distances; instead, it learns only for each point v in V and each candidate pairs c, c' which of c, c' is closer to v. Our main result is that the EXPANDING APPROVALS RULE of Aziz and Lee is (α, γ) representative with $\gamma \leq 1 + 6.71 \cdot \frac{\alpha}{\alpha-1}$.

Our results lead to three notable byproducts. First, we show that the EXPANDING APPROVALS RULE achieves constant proportional fairness in the ordinal model, giving the first positive result on metric proportional fairness with ordinal information. Second, we show that for the core fairness objective, the EXPANDING APPROVALS RULE achieves the same asymptotic tradeoff between resource augmentation and approximation as the recent results of Li et al., which used full knowledge of the metric. Finally, our results imply a very simple single-winner voting rule with metric distortion at most 44.

1 Introduction

Selecting representatives for a large set is a common and central problem across a wide range of application areas. As three paradigmatic applications, consider selecting a small set of documents (such as pictures or text) representing a much larger collection, selecting a committee of representatives for a large population, or selecting locations for several public facilities to serve the population of a city. Naturally, there are many ways of defining what it means for a candidate set to be "representative"; we discuss some key definitions from past work in Section 5.

A commonly accepted notion of representation is based on *proportionality* (Humphreys 1911; Moulin 2003): subgroups of the population should be represented in the selected set proportionally to their size. That is, if a cohesive subset S comprises a θ fraction of the documents/population, then (at least) roughly a θ fraction of the representative set should be similar to the members of S. In terms of documents, this implies that by examining the representative documents, a user can accurately assess the contents of the document collection. For committee elections, it states that large like-minded groups of the population should be suitably represented in the committee. And for the location of public facilities, it implies that dense population centers should be sufficiently served with nearby facilities. Indeed, notions of fairness or representation based on this intuition have been studied extensively, as discussed in Section 5.

We are particularly interested in the common setting in which the documents or candidates/population are embedded in a metric space in which distances capture dissimilarity. For documents, this is often the result of feature-based embeddings applied to the documents, and for the selection of public facilities, the metric is naturally derived from geographic proximity or transportation times. For voters and candidates in elections, the idea of considering all agents as embedded in a metric space, and their preferences being reflective of the distances, was first articulated as *single-peaked preferences* (where the metric space is the line, e.g., representing a one-dimensional left-to-right spectrum of opinions) (Black 1948; Moulin 1980), but also subsequently generalized to other metrics (Barberà, Gul, and Stacchetti 1993; Merrill and Grofman 1999).

Our first main contribution (in Section 2) is a natural and novel definition of what it means for a set R of k points to "represent" a larger set V in a metric space; our notion is a strengthening of the notion of *core fairness* proposed recently by Li et al. (2021).

Our second main contribution (in Section 4) is to show that a natural algorithm (a special case of the EXPANDING APPROVALS RULE of Aziz and Lee (2020)) achieves strong representativeness guarantees for the new definition. It does

*Work done while at the University of Southern California.

so even though it works in the *ordinal* model (see Section 3), in which the algorithm only learns, for each voter/document/citizen, the ranking of potential representatives by increasing distances (but not the distances themselves). As immediate corollaries of our analysis, we obtain the first algorithm with constant proportional fairness for metric costs in the ordinal model, an algorithm with ordinal information achieving — up to constants — the same parameter tradeoff for approximate core fairness as the one of Li et al. (2021) (which had access to the full metric), as well as an extremely simple single-winner voting rule with constant metric distortion; see Section 4.

Finally, we show that when the algorithm has access to all the distances, a slight modification of the GREEDY CAPTURE algorithm of Chen et al. (2019) provides improved constants in the representativeness guarantees. Due to space constraints, this result is deferred to the full version (Kalayci, Kempe, and Kher 2023), as are all proofs missing from the main body of this paper.

2 The Key Fairness Concepts

Our first main contribution is a new definition of a representative set in a metric space. Our definition is a strengthening of the notion of core fairness proposed by Li et al. (2021), and — as that definition does — naturally recovers an approximate median as a special case for $k = 1$.

We consider settings in which a set V should be "well represented" by a subset of a set[1] C; for examples, see Section 1. We write $n = |V|$ and $m = |C|$ for the sizes of the sets, and $k < m$ for the size of the subset of C that is to be selected. For concreteness in our nomenclature, we will refer to V as *voters* and C as *candidates* throughout, although we do not exploit any specific properties of this domain.

We assume that $V \cup C$ is embedded in a (pseudo-)metric space $(V \cup C, d)$.[2] The goal is to pick a set $R \subseteq C$ of k *representatives* to ensure that each sufficiently large subset $S \subseteq V$ is "well represented", in a sense we define next. In keeping with the voting-related nomenclature, we will refer to R as a *committee* and to S as a *coalition*.

We write $d_{\text{sum}}(X, Y) = \sum_{x \in X, y \in Y} d(x, y)$ for the sum of distances between all pairs in $X \times Y$. When $X = \{x\}$ is a singleton, we write $d_{\text{sum}}(x, Y) = d_{\text{sum}}(\{x\}, Y)$.

2.1 Proportional Representation

Recall that our goal is to ensure that all sufficiently large coalitions of voters are well represented. Specifically, if a coalition S comprises a θ fraction of all *voters*, at least a θ fraction of the *committee* should be approximately closest to S as a whole. To phrase this requirement cleanly, we recall the definition of the *Hare Quota* $p = \lceil n/k \rceil$, the number of voters/documents represented by any one selected candidate/document/facility. As articulated by Li et al. (2021) in their definition of (approximate) *core fairness*, any coalition S whose size is at least the Hare Quota should have at least one roughly satisfactory representative in R. Our generalization requires that, for any positive integer t, each member of a coalition S of size $|S| \geq t \cdot p$ should have at least t representatives in R, such that no other t candidates are much better for the coalition compared to their individual best t members in R.

Definition 2.1 (γ-proportionally representative committee). *A committee R is called* γ-proportionally representative *if for every coalition $S \subseteq V$ of size at least $t \cdot p$, the committee satisfies:*

$$\sum_{v \in S} \min_{\substack{R'_v \subseteq R \\ |R'_v| = t}} d_{sum}(v, R'_v) \leq \gamma \cdot \min_{\substack{C' \subseteq C \\ |C'| = t}} d_{sum}(S, C'). \qquad (1)$$

Definition 2.1 captures a notion of approximate *stability*: no sufficiently large coalition S of voters could find an alternative committee for themselves of corresponding size t which they strongly prefer over their individually best size-t subcommittees of R. Note that proportional representation (Definition 2.1) is a more demanding requirement than approximate core fairness in the sense of Li et al. (2021), as defined in Definition 2.3; we will elaborate on this in more detail below. A positive feature of Definition 2.1 is that it does not require a notion of "cohesive" coalitions which must be represented. Rather, a requirement for *all* sufficiently large coalitions is given; however, the requirements for very spread-out coalitions are typically trivially satisfied, because such coalitions do not have attractive alternatives to deviate to.

While natural and intuitive, our definition of proportional representation is unfortunately too demanding; Li et al. (2021) already showed that there are instances for which the $(1, \Omega(\sqrt{n}))$-core[3] is empty. Since Definition 2.1 has additional constraints, we can in general not hope for $\gamma = o(\sqrt{n})$.

Therefore, as in Li et al. (2021), we relax Definition 2.1 by allowing for *resource augmentation* — see Jiang, Munagala, and Wang (2020) for another example and discussion of the use of resource augmentation to deal with impossibility of proportional representation. Specifically, for a *resource augmentation parameter* $\alpha \geq 1$, we consider the problem of selecting a committee of size k, but require the weaker stability guarantee for a committee of (smaller) size k/α, as captured by the following definition.

Definition 2.2 ((α, γ)-proportionally representative committee). *For a resource augmentation parameter $\alpha \geq 1$, a committee R is called* (α, γ)-proportionally representative

[1] $V = C$ — the most natural case when selecting documents, and a case corresponding to peer selection in the context of elections — is of course allowed.

[2] Recall that a *metric* is a non-negative function d on pairs satisfying that $d(x, x) = 0$ for all x, *symmetry* ($d(x, y) = d(y, x)$ for all x, y), triangle inequality ($d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, z), and *positivity* ($d(x, y) > 0$ whenever $x \neq y$). A pseudo-metric is allowed to violate positivity, i.e., multiple points of the metric space can be at distance 0 from each other.

[3] The instances described in Li et al. (2021) require $n = \Theta(k^2)$. In the full version (Kalayci, Kempe, and Kher 2023), we give a simple class of instances showing a slightly more fine-grained lower bound of $\Omega(\min(k, n/k))$, thus giving a lower bound for the entire range of k.

if for every coalition $S \subseteq V$ of size at least $t \cdot \alpha \cdot p$, the committee satisfies:

$$\sum_{v \in S} \min_{\substack{R'_v \subseteq R \\ |R'_v| = t}} d_{sum}(v, R'_v) \le \gamma \cdot \min_{\substack{C' \subseteq C \\ |C'| = t}} d_{sum}(S, C'). \tag{2}$$

Definition 2.2 is a strengthening of the definition of approximate core in Li et al. (2021), defined as follows:

Definition 2.3 (Approximate Core)**.** *For a resource augmentation parameter $\alpha \ge 1$, a committee R is in the (α, β)-*core *if for every coalition $S \subseteq V$ of size at least $\alpha \cdot p$, the committee R satisfies:*

$$\sum_{v \in S} \min_{r_v \in R} d(v, r_v) \le \beta \cdot \min_{c \in C} d_{sum}(S, c). \tag{3}$$

Thus, the definition of the (α, β)-core[4] is obtained by requiring only the constraints for $t = 1$ in Definition 2.2.

After augmenting resources, the definition becomes less stringent, and one may ask if approximation in the objective can be avoided, i.e., whether $\gamma = 1$ (or $\beta = 1$) can be achieved. We show that this is in general not possible; in the full version (Kalayci, Kempe, and Kher 2023), we prove the following lower bound, which already holds for the weaker notion of approximate core in the sense of Li et al. (2021):

Proposition 2.4. *For every $\alpha \ge 1$, there are instances for which the $(\alpha, 1 + 1/(2\alpha))$-core is empty, and so, no $(\alpha, 1 + 1/(2\alpha))$-proportional representation exists.*

Furthermore, in the full version (Kalayci, Kempe, and Kher 2023), we show that when $\alpha \to 1$, a large blowup in γ is unavoidable in general. Again, we show this lower bound result already for the approximate core:

Proposition 2.5. *There are α arbitrarily close to 1 and corresponding instances for which the $(\alpha, 1/(4(\alpha - 1)))$-core is empty, and so, no $(\alpha, 1/(4(\alpha - 1)))$-proportional representation exists.*

Note that Proposition 2.5 provides an asymptotically matching lower bound for Theorem 19 of Li et al. (2021).

While we cannot achieve $\alpha \approx 1$ or $\gamma \approx 1$ without a large blowup in the other parameter, we still seek to design (polynomial-time) algorithms for computing committees achieving a good tradeoff between α and γ; indeed, this is the main goal of our work, studied in Section 4.

2.2 Proportional Representation, Core Fairness, and Proportionally Fair Clustering

As discussed above, proportional representation is a natural strengthening of the notion of core fairness. Another closely related concept is *proportional fairness*, introduced in the field of clustering by Chen et al. (2019), and studied further by Micha and Shah (2020). Here, the setup is the same, and the committee R is construed as *cluster centers*:

Definition 2.6 (γ-proportional fairness)**.** *A committee R of size k is* γ-proportionally fair *if for every voter coalition S of size at least p and every alternate candidate c, at least one voter $v \in S$ satisfies* $\min_{r \in R} d(v, r) \le \gamma \cdot d(v, c)$.

Li et al. (2021) already pointed out that every committee in the $(1, \beta)$-core according to their definition is also β-proportionally fair. Since $(1, \beta)$-proportional representation implies membership in the $(1, \beta)$-core, it is strictly stronger than β-proportional fairness.

We discuss key differences between Definitions 2.1, 2.3 and 2.6, as well as their implications. We believe that they justify Definition 2.1 as a more suitable notion of representativeness of a committee.

1. Regardless of the size of S and k, the definition of approximate core (and thus also proportionally fair clustering) only requires the existence of a single candidate c that is preferred in order to "satisfy" S. As a result, a large committee R could significantly distort the composition of V.

   For example, consider $k = 100$, with two clusters V_1, V_2, containing 99% and 1% of the voters, respectively. The two clusters are far from each other, and each has a large number of possible candidates, all at distance 1 from each voter in the cluster. Then, R could contain 99 candidates close to V_2 and one candidate close to V_1, while satisfying Definition 2.3 and Definition 2.6 with $\beta = 1$. The exact same types of distances can be used to obtain similarly non-proportional outcomes for Example 1 in Li et al. (2021). In a sense, neither Definition 2.3 nor Definition 2.6 include a notion of *proportionality*, and thus, they fail to achieve it.
2. Definition 2.6 suffers from a second weakness, which is fixed by Definition 2.3: for small k (in particular, $k = 1$), it is extremely lenient. In fact, for $k = 1$, it allows the chosen candidate to be any candidate not Pareto-dominated by another; among others, this includes any candidate ranked first by at least one voter. This is a very weak requirement for a chosen candidate being "representative". In contrast, any candidate in the $(1, \beta)$-core (and thus any candidate who is $(1, \beta)$-representative) must be a β-approximate median of the voters, a much more meaningful sense of being representative.
3. The distinction between the definitions can also be viewed through the lens of whether the utility from deviations is transferable within the deviating coalition, as pointed out by Li et al. (2021). While proportionally fair clustering implicitly assumes non-transferable utilities, Definition 2.3 as well as Definition 2.1 correspond to transferable utilities, leading to a larger set of "deviation threats". Notice that the question of whether utility is transferable also arises in more "classical" definitions of the core (e.g., Peters (2015)).

3 The Ordinal Information Model

While the assumption of a known metric is reasonable for finding representative documents or the location of public facilities, for the election of a committee, the metric space is a useful modeling tool, but typically not known explicitly. Rather, the voters are assumed to *rank* the candidates by non-decreasing distance from themselves, and the algorithm has access to the rankings, but not the distances. Despite this limited information (after all, many different met-

[4] We use γ in place of β in Li et al. (2021) to emphasize the difference in the definitions.

rics may be consistent with the rankings), an algorithm or voting rule should select an approximately optimal solution, in our case, a proportionally representative set R exhibiting a good tradeoff between α and γ.

This framework is called the *ordinal information model* of the *metric distortion*[5] framework (Anshelevich et al. 2018, 2021a,b; Caragiannis, Shah, and Voudouris 2022), contrasting it with the *cardinal* model, in which the metric is known explicitly. The worst-case loss in the objective function due to the lack of information is called *(metric) distortion*, and has been studied for numerous optimization problems. Most notable is the by now extensive line of work on the distortion of *single-winner elections* (Anshelevich, Bhardwaj, and Postl 2015; Anshelevich et al. 2018; Munagala and Wang 2019; Gkatzelis, Halpern, and Shah 2020; Kizilkaya and Kempe 2022; Gkatzelis, Latifian, and Shah 2023); however, more complex objectives have also been studied (Anshelevich et al. 2021b,a; Anshelevich and Zhu 2021; Anari, Charikar, and Ramakrishnan 2023).

Viewed in this context, under the ordinal information model, our goal of selecting a committee R can be viewed as a natural extension of the metric distortion objective to multi-winner elections; a detailed comparison to other recently proposed multi-winner distortion objectives is given in Section 5.

We now define the concepts formally. The algorithm learns, for each voter $v \in V$, a *ranking* of candidates $\succ_v$. Voters rank candidates by non-decreasing distances, so $c \succ_v c'$ implies that $d(v,c) \leq d(v,c')$. We use $\pi_v(c)$ to denote the position of candidate c in v's ranking, with $\pi_v^{-1}(1)$ being v's most preferred candidate. We write $\succ_V = (\succ_v)_{v \in V}$ for the vector of all voters' rankings, and refer to it as the *ranked-choice profile*. An *election* consists of the triple $(V, C, \succ_V)$. We say that a pseudo-metric d is *consistent* with the ranked-choice profile $\succ_V$ if it satisfies that $d(v,c) \leq d(v,c')$ whenever $c \succ_v c'$, for all v, c, c'.

An ordinal committee selection rule f receives as input the committee size k and the election $(V, C, \succ_V)$, and outputs a committee $R \subseteq C$ of size k.

Our notion of proportional representation for ordinal models in committee selection rules is closely related to, and intended to be a natural generalization of, the concept of metric distortion for single-winner elections. Recall that the *metric distortion* of a single-winner voting rule f is the worst-case ratio (over all elections, and all metrics consistent with the election) of the total cost of the chosen winner relative to the total cost of the optimum candidate, i.e.,

$$\max_{(V,C,\succ_V)} \; \max_{d \text{ consistent with } \succ_V} \frac{d_{\text{sum}}(V, f(V,C,\succ_V))}{\min_{c \in C} d_{\text{sum}}(V,c)}.$$

As a result, notice that in the special case when $k = 1$ and $\alpha = 1$, a committee selection rule f is γ-proportionally representative if and only if f is a single-winner voting rule with metric distortion at most γ. This is because the only coalition of size $p = n$ is $S = V$, so the optimal γ value in Eq. (2) corresponds to the metric distortion.

4 Our Main Result

Our second main contribution — and the key technical work in this paper — is to show that a very natural algorithm, namely, a special case of the Expanding Approvals Rule of Aziz and Lee (2020), selects an (α, γ)-proportionally representative committee R of size k which achieves constant α and γ; furthermore, it does so even in the ordinal model.

The algorithm runs in iterations, parametrized by a *tolerance* parameter τ, which starts at 1 and increases in each round. In iteration τ, the algorithm considers, in some arbitrary order, each (remaining) voter as approving their top τ choices. As soon as at least $p = \lceil n/k \rceil$ of the remaining voters approve of a particular candidate c, this candidate is added to the committee, and those p voters are permanently removed from further consideration. We will say that c *covers* these voters. The fact that the algorithm processes the voters in an arbitrary order (instead of simultaneously) in each iteration achieves an (arbitrary) tie breaking implicitly, so the algorithm does not have to consider ties between multiple candidates becoming eligible for inclusion. This process continues until a committee of size k has been formed. The algorithm is described formally as Algorithm 1.

Algorithm 1: Expanding Approvals Rule

```
Input: Election $(V, C, \succ_V)$, Committee Size $k$
Output: Committee $R$
  Let $U \leftarrow V$ be the set of uncovered voters.
  Let $R \leftarrow \emptyset$ be the selected committee.
  Let $N_c \leftarrow \emptyset$ for all $c \in C$.
  for $\tau = 1, \ldots, m$ do
      for $v \in V$ in arbitrary order do
          if $v \in U$ then
              Let $c = \pi_v^{-1}(\tau)$.
              if $c \notin R$ then
                  Let $N_c \leftarrow N_c \cup \{v\}$.
                  if $|N_c| = \lceil n/k \rceil$ then
                      $R \leftarrow R \cup \{c\}$.
                      $N_{c'} \leftarrow N_{c'} \setminus N_c$ for all $c' \in C \setminus R$.
                      $U \leftarrow U \setminus N_c$.
                      We say that $N_c$ has been covered by $c$.
  if $|R| < k$ then add $k - |R|$ arbitrary candidates to $R$.
```

The running time of Algorithm 1 is essentially linear. The analysis of the running time, as well as a discussion of why precisely the algorithm is a special case of the Expanding Approvals Rule, is given in the full version (Kalayci, Kempe, and Kher 2023).

Theorem 4.1. *The* Expanding Approvals Rule *outputs a committee R of size k which is $(\alpha, \gamma(\alpha))$-proportionally representative for all $\alpha > 1$, with $\gamma(\alpha) = 1 + \frac{7+\sqrt{41}}{2} \cdot \frac{\alpha}{\alpha-1} \approx 1 + 6.71 \cdot \frac{\alpha}{\alpha-1}$.*

[5] A parallel line of work (e.g., (Boutilier et al. 2015; Boutilier and Rosenschein 2016; Procaccia and Rosenschein 2006)) considers the same tradeoff of ordinal vs. cardinal information in a setting in which the rankings are derived from positive *utilities* rather than distances/costs.

Theorem 4.1 gives guarantees only for $\alpha > 1$, and Proposition 2.5 shows that this is unavoidable for decent approximation guarantees, unless k is very small or large. When resources are not augmented, i.e., for $\alpha = 1$, we obtain the following weaker guarantee, which is proved in the full version (Kalayci, Kempe, and Kher 2023).

Theorem 4.2. *The output R of Algorithm 1 is $(1, O(n/k))$-proportionally representative.*

Our analysis directly implies several novel results. First, an immediate corollary of a key lemma in the proof of Theorem 4.1 is that the EXPANDING APPROVALS RULE achieves proportional fairness 5.71; this constitutes the first result achieving constant proportional fairness with metric costs in the ordinal information model:

Corollary 4.3. *The* EXPANDING APPROVALS RULE *is a $\frac{5+\sqrt{41}}{2} \approx 5.71$-proportionally fair clustering algorithm under the ordinal information model with metric costs.*

The constant $\frac{5+\sqrt{41}}{2}$ is larger than the best proportional fairness achievable with full knowledge of the metric space, which is $1 + \sqrt{2} \approx 2.41$ (Chen et al. 2019; Micha and Shah 2020). This is perhaps not surprising, given that in the ordinal information model, the algorithm is missing crucial information. Indeed, in the extended version (Kalayci, Kempe, and Kher 2023), we give an instance in which under the ordinal information model, no deterministic algorithm can produce a γ-proportionally fair committee for any $\gamma < 2 + \sqrt{5} \approx 4.23$; thus, there is necessarily a gap between the best proportional fairness achievable in the cardinal and ordinal models. Obtaining the best possible proportional fairness guarantees under metric costs and ordinal information is an interesting direction for future work.

A second corollary can be immediately obtained from Theorem 4.1 and Theorem 4.2: because (α, β)-representativeness implies being in the (α, β)-core, the guarantees of Theorem 4.1 and Theorem 4.2 apply verbatim to the latter. Thus, the EXPANDING APPROVALS RULE achieves approximate core fairness in the ordinal metric cost model; this, too, is the first positive result on core fairness in the ordinal metric cost model.

Corollary 4.4. *The committee R output by Algorithm 1 is in the $(\alpha, \beta(\alpha))$-core for all $\alpha > 1$, with $\beta(\alpha) = 1 + \frac{7+\sqrt{41}}{2} \cdot \frac{\alpha}{\alpha-1} \approx 1 + 6.71 \cdot \frac{\alpha}{\alpha-1}$. It is also in the $(1, O(n/k))$-core.*

Note that these bounds match the information-theoretic lower bound of Proposition 2.5 and the lower bound of $\Omega(\min(k, n/k))$ for $\alpha = 1$ (detailed in the full version (Kalayci, Kempe, and Kher 2023)) up to constant factors. They mirror (with slightly worse constants) the bounds obtained by Li et al. (2021) in the model with known distances (Theorem 19).

A third corollary, gives an extremely simple single-winner voting rule with constant metric distortion.

Corollary 4.5. *For a given set of candidates C and voters V, consider the following voting rule: find a candidate c who is in the top τ positions of at least $\lceil n/2 \rceil$ voters, for the smallest possible τ. (Break ties arbitrarily.) Find a candidate c' who is in the top τ' positions of the remaining $\lfloor n/2 \rfloor$ voters, for the smallest possible τ'. (Again, break ties arbitrarily.) Return the one of c, c' preferred by a majority of voters.*

This voting rule has metric distortion at most 44 for the single-winner election $(V, C, \succ_V)$.

While the distortion guarantee of 44 given by Corollary 4.5 is worse than the (optimal) metric distortion of 3 achieved by Gkatzelis, Halpern, and Shah (2020); Kizilkaya and Kempe (2022), this voting rule is arguably even simpler than the rules previously known to achieve constant metric distortion (COPELAND (Anshelevich et al. 2018), PLURALITY-MATCHING (Gkatzelis, Halpern, and Shah 2020), PLURALITY-VETO (Kizilkaya and Kempe 2022)).

4.1 Stronger Proportional Fairness

In the analysis of Algorithm 1, a central concept is for any coalition S the set of all chosen representative candidates $r \in R$ who covered at least one voter in S. We formally define this notion; recall here that N_r is defined in Algorithm 1:

Definition 4.6 (Representatives for a Coalition)**.** *The set of representatives for the coalition $S \subseteq V$ is defined as $R[S] = \{r \in R \mid N_r \cap S \neq \emptyset\}$, i.e., $R[S] \subseteq R$ is the set of candidates $r \in R$ whose neighborhood contains at least one voter in S.*

We show that the committee R returned by Algorithm 1 satisfies a somewhat stronger notion of proportional fairness, i.e., a modification of Definition 2.6. Lemma 4.7 — proved in the full version (Kalayci, Kempe, and Kher 2023) — shows that for any coalition S, the representatives $R[S]$ are already sufficiently attractive that S will not unanimously deviate. The guarantee differs from Definition 2.6 only in the slight strengthening of replacing R with $R[S]$.

Lemma 4.7. *The committee R output by Algorithm 1 has the following stability property, with $\rho = \frac{5+\sqrt{41}}{2} \approx 5.71$: For every coalition S of size $|S| \geq p = \lceil n/k \rceil$, there exists a voter $\min_{r \in R[S]} d(v, r) \leq \rho \cdot \min_{c \in C \setminus R} d(v, c)$.*

Because $R[S] \subseteq R$, Lemma 4.7 implies that Algorithm 1 outputs a $\frac{5+\sqrt{41}}{2}$-proportionally fair clustering in the ordinal information model, which proves Corollary 4.3.

4.2 Proof of Theorem 4.1 and Corollaries

Proof of Theorem 4.1. Using Lemma 4.7, we are now ready to complete the proof of Theorem 4.1. We will show that the committee R returned by Algorithm 1 satisfies a stronger stability guarantee than the claimed (α, γ)-proportional representation. We will show that for every coalition $S \subseteq V$ of size at least $t \cdot \alpha \cdot p$, the committee satisfies:

$$\min_{\substack{R' \subseteq R \\ |R'|=t}} d_{\text{sum}}(S, R') \leq \gamma \cdot \min_{\substack{C' \subseteq C \\ |C'|=t}} d_{\text{sum}}(S, C'). \tag{4}$$

That is, for the coalition S, there is a subcommittee R' of size t which is almost as good as the best C. Note that the cost of S for R' is an upper bound on the sum of costs for voters $v \in S$ for their individually optimal size-t subcommittees.

Let S be an arbitrary coalition of size $|S| \geq t \cdot \alpha \cdot p$. Let $C^* \in \operatorname{argmin}_{C':|C'|=t} d_{\text{sum}}(S, C')$ be a set of t candidates with smallest total distance to S. We will define a perfect matching between C^* and R such that for every $c \in C^*$, its match $r \in R$ is an "approximately good" alternative. Overall, this will demonstrate that R is not much worse than C^*.

Let $c_1, c_2, \ldots, c_t$ be an enumeration of the candidates in C^*, such that the candidates in $C^* \cap R$ precede the ones in $C^* \setminus R$; apart from this requirement, the order can be arbitrary. We define the matching representatives $r_1, \ldots, r_t$ iteratively. First, for each $c_i \in C^* \cap R$, we define $r_i = c_i$. Subsequently, for each iteration i, having already defined $r_1, \ldots, r_{i-1}$, we let $S_i = S \setminus \bigcup_{j=1}^{i-1} N_{r_j}$, i.e., S_i is the set of all voters in S except those covered by the first $i-1$ selected candidates r_j. Let $T_i \in \operatorname{argmin}_{T \subseteq S_i, |T|=p} \sum_{v \in T} d(v, c_i)$ be a subset of S_i of size $|T_i| = p$ comprising the p voters in S_i closest to c_i (ties broken arbitrarily). As $|S| \geq \lceil \alpha \cdot t \cdot p \rceil$ and each N_{r_j} has size at most p, $|S_i|$ has size at least p for each i, and thus T_i is well defined. Because $c_i \notin R$ in the current case, by Lemma 4.7, there exists a voter $v_i \in T_i$ and candidate $r_i \in R[T_i] \subseteq R[S]$ such that

$$d(v_i, r_i) \leq \rho \cdot d(v_i, c_i). \tag{5}$$

Now, we confirm that the resulting assignment is indeed a matching. Observe that by definition of $R[S_i]$, we know that $S_i \cap N_{r_j} = \emptyset$ for all $j < i$; in particular, $r_j \notin R[S_i]$ for all $j < i$. This implies that $r_1, \ldots, r_t$ are distinct from each other, and so the assignment is a matching. We also remark here that Eq. (5) holds for the i with $c_i \in R$ as well, because $r_i = c_i$ implies $d(v_i, r_i) = d(v_i, c_i)$ for those candidates.

In the rest of the proof, we will show that r_i is an approximately good alternative to c_i for all of S, showing that $d_{\text{sum}}(S, r_i) < \gamma(\alpha) \cdot d_{\text{sum}}(S, c_i)$. Fix an arbitrary index $i \in \{1, \ldots, t\}$. Let $\widehat{v}_i$ be a voter in T_i maximizing the distance $d(v, c_i)$ to c_i over all $v \in T_i$, i.e., a voter at (or possibly tied for) the p^{th} largest distance from c_i among voters in T_i. For any voter $v \in S$, using the triangle inequality and (5), we can bound the distance

$$\begin{aligned} d(v, r_i) &\leq d(v, c_i) + d(v_i, c_i) + d(v_i, r_i) \\ &\leq d(v, c_i) + (1 + \rho) \cdot d(v_i, c_i) \qquad (6) \\ &\leq d(v, c_i) + (1 + \rho) \cdot d(\widehat{v}_i, c_i) \\ &\leq d(v, c_i) + (1 + \rho) \cdot \max\left\{d(v, c_i), d(\widehat{v}_i, c_i)\right\}. \end{aligned}$$

Let $P_i = \{v \in S \mid d(v, c_i) < d(\widehat{v}_i, c_i)\}$ be the subset of S containing all voters v with $d(v, c_i) < d(\widehat{v}_i, c_i)$. Because T_i was chosen to be the p voters closest to c_i inside S_i, and $\widehat{v}_i$ the voter furthest from c_i in T_i, we get that $P_i \subseteq T_i \cup \bigcup_{j=1}^{i-1} N_{r_j}$ and so $|P_i| \leq p \cdot i$. Summing the distances to r_i over all voters $v \in S$, and using (in the second step) that $d(\widehat{v}_i, c_i) \leq d(v, c_i)$ for all $v \in S \setminus P_i$, we obtain the bound

$$\begin{aligned} &\sum_{v \in S} d(v, r_i) - \sum_{v \in S} d(v, c_i) \\ &\quad \leq (1 + \rho) \left(\sum_{v \in P_i} d(\widehat{v}_i, c_i) + \sum_{v \in S \setminus P_i} d(v, c_i) \right) \\ &\quad \leq (1 + \rho) \left(\frac{|P_i|}{|S| - |P_i|} + 1 \right) \sum_{v \in S \setminus P_i} d(v, c_i) \\ &\quad \leq (1 + \rho) \left(\frac{p \cdot i}{|S| - p \cdot i} + 1 \right) \sum_{v \in S \setminus P_i} d(v, c_i) \\ &\quad \leq (1 + \rho) \left(\frac{1}{\alpha - 1} + 1 \right) \sum_{v \in S \setminus P_i} d(v, c_i) \\ &\quad \leq (1 + \rho) \frac{\alpha}{\alpha - 1} \cdot \sum_{v \in S} d(v, c_i). \end{aligned}$$

In the penultimate step, we used that $\frac{p \cdot i}{|S| - p \cdot i} \leq \frac{p \cdot i}{\alpha \cdot p \cdot i - p \cdot i} = \frac{1}{\alpha - 1}$. Finally, by rearranging the inequality and summing up over all indices i, we obtain

$$\sum_{i=1}^{t} d_{\text{sum}}(S, r_i) \leq \left(1 + (1 + \rho) \cdot \frac{\alpha}{\alpha - 1} \right) \cdot \sum_{i=1}^{t} d_{\text{sum}}(S, c_i).$$

Substituting $\rho = \frac{5 + \sqrt{41}}{2}$ now completes the proof. □

As a special case of Theorem 4.1 and Theorem 4.2, we obtain Corollary 4.4, simply by focusing on just the case $t = 1$ in Theorem 4.1. In fact, by recalling that the proof of Theorem 4.1 established the somewhat stronger guarantee Eq. (4), we obtain the following Corollary 4.8, strengthening Corollary 4.4. This corollary states that no coalition of at least $\alpha \cdot p$ voters (for $\alpha > 1$) has a candidate outside R whom they strongly prefer *on average* to their *best single* candidate in $R[S]$:

Corollary 4.8. *Let R be the committee output by the* Expanding Approvals Rule. *Let $\alpha > 1$, and $\beta(\alpha) = 1 + \frac{7 + \sqrt{41}}{2} \cdot \left(\frac{\alpha}{\alpha - 1}\right) \approx 1 + 6.71 \cdot \left(\frac{\alpha}{\alpha - 1}\right)$. For any coalition S of size $|S| \geq \alpha \cdot p$,*

$$\min_{r \in R[S]} d_{sum}(S, r) \leq \beta(\alpha) \cdot \min_{c \in C \setminus R} d_{sum}(S, c).$$

Finally, we show how Theorem 4.1 implies Corollary 4.5, i.e., the existence of an extremely simple single-winner voting rule with constant distortion.

Proof of Corollary 4.5. Consider the committee R output by Algorithm 1 when run with $k = 2$. By Theorem 4.1, applied with $\alpha = n/p \leq 2$, we get that $\min_{r \in R} d_{\text{sum}}(V, r) \leq \gamma(2) \cdot \min_{c \in C} d_{\text{sum}}(V, c)$. This implies that at least one representative in R has distortion at most $\gamma(2) \approx 14.42$ for the single-winner election. Let r_1 be the winner of the majority election between the two candidates in R, and r_2 the other candidate. Since r_1 is preferred over r_2 by at least half of the voters, Lemma 6 of Anshelevich et al. (2018) implies that $d_{\text{sum}}(V, r_1) \le 3d_{\text{sum}}(V, r_2)$. This implies that r_1 has distortion at most $3 \cdot \gamma(2) \le 44$ for the single-winner election. □

5 Related Work

Representing a large point set is also a — explicit or implicit — goal of clustering points in a metric space. A review of this very large literature is beyond the scope of our work. Of the common objective functions (most notably, k-center, k-means, and k-median), k-median (Arya et al. 2001) is closest to our objective here, since it minimizes the sum of distances of points to the respective closest selected k cluster centers. In the basic definition, like proportionally fair clustering, it suffers from the fact that a large and dense cluster can be "served" by just one selected representative.

Capacitated versions address this issue by limiting the number of points "served" by a cluster center (see, e.g., Byrka et al. (2015)). Our algorithms similarly assign to each representative c a subset of size p from V that c "covers". To appreciate the difference, suppose that roughly p points are at a large distance M from all others points which are within a unit circle. If a representative is at distance $(1-\epsilon)M$ from these remote points, a k-median solution might include it since the objective reduction by ϵM would outweigh any decisions made within the unit circle. In contrast, our proportional representation objective might exclude it since the relative improvement would be marginal. Thus, the proportional representation objective is much less sensitive to few outliers. It should also be noted that both objectives require resource augmentation for positive results (Byrka et al. 2015). Nonetheless, it would be interesting to explore the link between these two objectives and determining if one implies non-trivial guarantees for the other.

Several papers have proposed extensions of the notion of (metric) distortion to multi-winner (or committee) elections. Goel, Hulett, and Krishnaswamy (2018) consider the cost of a set R to be the sum of all distances between V and the members of R. Because this sum decomposes, it is minimized by choosing the k candidates individually closest to V (in terms of sums of distances); as a result, the set R tends to be as "homogeneous" as possible. An alternative notion was proposed by Caragiannis, Shah, and Voudouris (2022). Their definition is parametrized by $q \le k$: each individual $v \in V$ evaluates the cost of R as the cost of the q^{th} closest representative in R; the objective to minimize is then the sum of these costs. When $q = 1$, the objective coincides with (uncapacitated) k-median; in contrast, for $q = k$, a committee has low cost for v only if all of its members are close to v. Thus, in the regime of large q, the definition suffers from the same drawback as that of Goel, Hulett, and Krishnaswamy (2018): it rewards committees R (almost) all of whose members are close to the largest cluster within V. Caragiannis, Shah, and Voudouris (2022) show that while the objective can be well approximated with ordinal information in the "homogeneous" case $q > k/2$, the distortion is unbounded for $q < k/3$; though no results are shown under resource augmentation.

There has been prior work relating notions of fairness, multi-winner elections, distortion, and the core. For example, Ebadian et al. (2022) study the *utilitarian distortion* of *randomized* single-winner voting rules. They use fairness both as a tool to derive novel low-distortion randomized voting rules, and — in a different notion of proportional fairness — as an optimization goal in its own right. They show that α-approximate proportional fairness in their definition implies membership in the α-core.

In general cooperative games, the notion of core captures a stability desideratum: that the outcome be stable against deviations by any subgroup of players seeking better utilities. The exact coalitions and available "outside options" of utilities give rise to specific core solution concepts.

In social choice, core definitions often emphasize proportionality, suggesting that a θ fraction of the population should "control" an equal fraction of the outcome (Moulin 2003). In the context of electing a committee, various notions of core stability require that every sufficiently cohesive and large voter group should sufficiently "approve" of the committee. One branch of the literature explores such notions in the context of approval ballots, i.e., each voter either approves or disapproves each individual candidate. Depending on the precise definition of "cohesiveness" and "approval", many versions of *proportional representation* emerge (Aziz et al. 2017; Sánchez-Fernández et al. 2017; Skowron 2021) (see also the survey by Lackner and Skowron (2023)), which can be seen as stability or fairness measures. With ranked ballots, these definitions can be adapted by having voters approve their top τ choices, with an adjustable parameter τ. This notion lies at the heart of expanding approvals. Another research direction, illustrated by the work of Cheng et al. (2020); Jiang, Munagala, and Wang (2020), moves beyond expanding approvals, by defining the core for ranked ballots through a lexicographical ordering or costs associated with a (hidden) metric space. Chen et al. (2019); Li et al. (2021) utilize these core definitions to explore fair clustering, which are key focal points of this paper and are extensively discussed throughout.

6 Concluding Remarks and Open Questions

Aziz and Lee (2020) showed that the Expanding Approvals Rule satisfies several desirable axiomatic properties for multi-winner elections, in addition to being a natural rule in its own right. We believe that our result thus adds to the evidence for Expanding Approvals Rule being a potentially useful rule to be used in practice.

In general, the constants in our upper and lower bounds do not match. Closing the gaps for all of the notions of representation studied here would be of interest, both in the model with known metric space and with ordinal information. Another possible direction of interest is to understand under what natural conditions about the metric space a constant factor in representativeness can be achieved without resource augmentation.

Acknowledgements

Work supported by ARO MURI grant W911NF1810208. We would like to thank Evi Micha for useful conversations.

References

Anari, N.; Charikar, M.; and Ramakrishnan, P. 2023. Distortion in metric matching with ordinal preferences. In *Proc. 24th ACM Conf. on Economics and Computation*, 90–110.

Anshelevich, E.; Bhardwaj, O.; Elkind, E.; Postl, J.; and Skowron, P. 2018. Approximating optimal social choice under metric preferences. *Artificial Intelligence*, 264: 27–51.

Anshelevich, E.; Bhardwaj, O.; and Postl, J. 2015. Approximating Optimal Social Choice under Metric Preferences. In *Proc. 29th AAAI Conf. on Artificial Intelligence*, 777–783.

Anshelevich, E.; Filos-Ratsikas, A.; Shah, N.; and Voudouris, A. A. 2021a. Distortion in social choice problems: an annotated reading list. *SIGecom Exchanges*, 19(1): 12–14.

Anshelevich, E.; Filos-Ratsikas, A.; Shah, N.; and Voudouris, A. A. 2021b. Distortion in Social Choice Problems: The First 15 Years and Beyond. In *Proc. 30th Intl. Joint Conf. on Artificial Intelligence*, 4294–4301.

Anshelevich, E.; and Zhu, W. 2021. Ordinal Approximation for Social Choice, Matching, and Facility Location Problems given Candidate Positions. *ACM Transactions on Economics and Computation*, 9(2): 1–24.

Arya, V.; Garg, N.; Khandekar, R.; Meyerson, A.; Munagala, K.; and Pandit, V. 2001. Local search heuristics for k-median and facility location problems. In *Proc. 33rd ACM Symp. on Theory of Computing*.

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

Aziz, H.; and Lee, B. E. 2020. The expanding approvals rule: improving proportional representation and monotonicity. *Social Choice and Welfare*, 54(1): 1–45.

Barberà, S.; Gul, F.; and Stacchetti, E. 1993. Generalized Median Voter Schemes and Committees. *Journal of Economic Theory*, 61: 262–289.

Black, D. 1948. On the Rationale of Group Decision Making. *J. Political Economy*, 56: 23–34.

Boutilier, C.; Caragiannis, I.; Haber, S.; Lu, T.; Procaccia, A. D.; and Sheffet, O. 2015. Optimal social choice functions: A utilitarian view. *Artificial Intelligence*, 227: 190–213.

Boutilier, C.; and Rosenschein, J. S. 2016. Incomplete information and communication in voting. In Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D., eds., *Handbook of Computational Social Choice*, chapter 10, 223–257. Cambridge University Press.

Byrka, J.; Fleszar, K.; Rybicki, B.; and Spoerhase, J. 2015. Bi-Factor Approximation Algorithms for Hard Capacitated k-Median Problems. In *Proc. 26th ACM-SIAM Symp. on Discrete Algorithms*, 722–736.

Caragiannis, I.; Shah, N.; and Voudouris, A. A. 2022. The metric distortion of multiwinner voting. *Artificial Intelligence*, 313: 103802.

Chen, X.; Fain, B.; Lyu, L.; and Munagala, K. 2019. Proportionally Fair Clustering. In *Proc. 36th Intl. Conf. on Machine Learning*, 1032–1041.

Cheng, Y.; Jiang, Z.; Munagala, K.; and Wang, K. 2020. Group Fairness in Committee Selection. *ACM Transactions on Economics and Computation*, 8(4).

Ebadian, S.; Kahng, A.; Peters, D.; and Shah, N. 2022. Optimized Distortion and Proportional Fairness in Voting. In *Proc. 23rd ACM Conf. on Economics and Computation*, 563–600.

Gkatzelis, V.; Halpern, D.; and Shah, N. 2020. Resolving the Optimal Metric Distortion Conjecture. In *Proc. 61st IEEE Symp. on Foundations of Computer Science*, 1427–1438.

Gkatzelis, V.; Latifian, M.; and Shah, N. 2023. Best of Both Distortion Worlds. In *Proc. 24th ACM Conf. on Economics and Computation*, 738–758.

Goel, A.; Hulett, R.; and Krishnaswamy, A. K. 2018. Relating Metric Distortion and Fairness of Social Choice Rules. In *Proc. 13th Workshop on Economics of Networks, Systems and Computation*.

Humphreys, J. H. 1911. *Proportional Representation: A Study in Methods of Election*. Methuen & Co.

Jiang, Z.; Munagala, K.; and Wang, K. 2020. Approximately Stable Committee Selection. In *Proc. 52nd ACM Symp. on Theory of Computing*, 463–472.

Kalayci, Y. H.; Kempe, D.; and Kher, V. 2023. Proportional Representation in Metric Spaces and Low-Distortion Committee Selection. arXiv:2312.10369.

Kizilkaya, F. E.; and Kempe, D. 2022. PLURALITYVETO: A Simple Voting Rule with Optimal Metric Distortion. In *Proc. 31st Intl. Joint Conf. on Artificial Intelligence*, 349–355.

Lackner, M.; and Skowron, P. 2023. *Approval-Based Committee Voting*, 1–7. Springer International Publishing.

Li, B.; Li, L.; Sun, A.; Wang, C.; and Wang, Y. 2021. Approximate Group Fairness for Clustering. In *Proc. 38th Intl. Conf. on Machine Learning*, 6381–6391.

Merrill, S.; and Grofman, B. 1999. *A unified theory of voting: Directional and proximity spatial models*. Cambridge University Press.

Micha, E.; and Shah, N. 2020. Proportionally Fair Clustering Revisited. In *Proc. 47th Intl. Colloq. on Automata, Languages and Programming*, 85:1–85:16.

Moulin, H. 1980. On Strategy-Proofness and Single Peakedness. *Public Choice*, 35: 437–455.

Moulin, H. 2003. *Fair division and collective welfare*. MIT Press.

Munagala, K.; and Wang, K. 2019. Improved Metric Distortion for Deterministic Social Choice Rules. In *Proc. 20th ACM Conf. on Economics and Computation*, 245–262.

Peters, H. 2015. Cooperative Games with Transferable Utility. In *Game Theory*, Springer Texts in Business and Economics, chapter 9, 151–169. Springer.

Procaccia, A. D.; and Rosenschein, J. S. 2006. The distortion of cardinal preferences in voting. In *Proc. 10th Intl. Workshop on Cooperative Inform. Agents X*, 317–331.

Sánchez-Fernández, L.; Elkind, E.; Lackner, M.; Fernández, N.; Fisteus, J.; Basanta Val, P.; and Skowron, P. 2017. Proportional Justified Representation. In *Proc. 31st AAAI Conf. on Artificial Intelligence*.

Skowron, P. 2021. Proportionality Degree of Multiwinner Rules. In *Proc. 22nd ACM Conf. on Economics and Computation*, 820–840.