The Panel Complexity of Sortition: Is 12 Angry Men Enough?

Johannes Brustle
Sapienza University of Rome
brustle@diag.uniroma1.it

Simone Fioravanti
Sapienza University of Rome
fioravanti@diag.uniroma1.it

Tomasz Ponitka
Tel Aviv University
tomaszp@mail.tau.ac.il

Jeremy Vollen
Northwestern University
vollen@northwestern.edu

Abstract

Sortition is the practice of delegating public decision-making to randomly selected panels. Recently, it has gained momentum worldwide through its use in citizens' assemblies, sparking growing interest within the computer science community. One key appeal of sortition is that random panels tend to be more *representative* of the population than elected committees or parliaments. Our main conceptual contribution is a novel definition of representative panels, based on the Wasserstein distance from statistical learning theory. Using this definition, we develop a framework for analyzing the *panel complexity* problem—determining the required panel size to ensure desirable properties. We focus on three key desiderata: (1) that efficiency at the panel level extends to the whole population, measured by social welfare; (2) that fairness guarantees for the panel translate to fairness for the population, captured by the core; and (3) that the probability of an outlier panel, for which the decision significantly deviates from the optimal one, remains low. We establish near-tight panel complexity guarantees for these desiderata across two fundamental social choice settings: facility location and participatory budgeting.

1 Introduction

A citizens' assembly is a panel of randomly selected citizens convened to deliberate on a policy question and collectively express a position. While the practice dates back millennia, it is not lacking in recent examples, including the Irish Citizens' Assembly (2016–2018), for which 100 randomly selected citizens addressed key legal and policy issues—such as abortion, aging, and climate change—leading to a national referendum on abortion and a government-declared climate emergency. Many other citizens' assemblies have been convened worldwide on a range of policy topics [OECD, 2020]. A defining feature of citizens' assemblies is their use of sortition—random selection of participants—an idea that has gained increasing attention in the computer science community [e.g., Benadè et al., 2019, Flanigan et al., 2020, 2021a,b, 2024, Halpern et al., 2025, Baharav and Flanigan, 2024, Walsh and Xia, 2012, Caragiannis et al., 2024, Ebadian et al., 2022, Meir et al., 2021, Cheng et al., 2018, 2017, Ebadian and Micha, 2025].

An important question in sortition is how to determine the optimal panel size. Naturally, the logistical and financial burden of a citizens' assembly increases with its size. At the same time, smaller panels are less likely to represent the entire population accurately. Moreover, other applications of sortition, such as jury selection, offer little guidance on navigating this trade-off. In the U.S., juries almost always consist of 12 members—a tradition dating back to 725, when King Morgan of Gla-Morgan, Wales, is said to have chosen this number to mirror the 12 Apostles [Gorski, 2012]. The central topic

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

of this paper is *panel complexity*: how many participants must a uniformly sampled panel have for its decision to satisfy a specified population-level guarantee? We treat panel members as the scarce resource and seek bounds in terms of accuracy and the decision setting.[1]

To precisely formulate the panel complexity problem, we must first establish a criterion for evaluating panels. Much of the literature on sortition focuses on the design and analysis of sortition algorithms with respect to two key desiderata: *representativeness* and *selection fairness*. Broadly, representativeness refers to ensuring that a citizens' assembly reflects the diverse perspectives of the underlying population [Young, 2002, Williams, 2000], while selection fairness is the ideal that posits that all citizens should have an equal chance of participation. In this work, we focus on the case where the panel members are selected uniformly at random from the population, ensuring selection fairness. Thus, our analysis is primarily concerned with achieving representativeness.

Much of the prior work defines representative panels based on compliance with exogenous quotas for demographic features [Flanigan et al., 2020, 2021a,b, 2024]. However, specifying quotas—and even determining which features should have quotas—may be difficult before resolving the question of panel size. Representativeness has also been defined in terms of the expected loss of a random panel of a given size relative to the optimal panel of the same size [Ebadian et al., 2022], or through the notion of the (approximate) core [Ebadian and Micha, 2025]. However, these definitions are not suitable for our context, as they do not provide a meaningful way to compare the representativeness of panels of different sizes. Thus, we ask:

Question 1: *What is the right measure of panel representativeness for analyzing panel complexity?*

Sortition-selected panels are often tasked with concrete decisions, such as legal or policy matters. We draw inspiration from Fishkin [2018], who argues that a citizens' assembly's decision should approximate what would emerge if the entire population deliberated. The panel size required to reach this end will necessarily depend on the particulars of the policy decision at hand, such as the outcome space and the structure of preferences. Hence, it is essential to study panel complexity in the context of specific *social choice settings*.

A recent work of Caragiannis, Micha, and Peters [2024] bounded the panel size required for uniform sampling to approximately maximize social welfare in facility location with finitely many alternatives.[2] In this model, individuals have ideal points in a metric space. A sampled panel might choose a public facility to minimize its members' total travel distance; one can then evaluate that location using the entire population's travel distances. Panel complexity asks how many sampled citizens suffice for these two objectives to approximately agree. While this is a promising result, it has two key limitations: (1) its bounds depend on the number of alternatives and thus do not account for scenarios with infinitely many alternatives, such as facility location in the Euclidean space; and (2) it does not capture preferences that go beyond distances in a metric space, such as those in the *participatory budgeting* problem [Fain et al., 2016, Goel et al., 2019], where a fixed budget must be distributed across a set of potential projects, and individuals differ in their preferences over how much funding each project should receive. Hence, we ask:

Question 2: *What is the panel complexity in social choice settings beyond facility location in finite spaces?*

The welfare guarantee above assumes that the panel selects a decision maximizing utilitarian welfare among its members. Time constraints or strategic behavior may prevent exact maximization; for example, strategyproof facility-location mechanisms generally provide approximate welfare guarantees [Chan et al., 2021]. In such circumstances, we ask what population-level guarantee follows when the panel only approximately optimizes its own objective. Moreover, a welfare-maximizing allocation need not be fair to all panel members, as emphasized in participatory budgeting [Fain et al., 2016]. This motivates studying whether an approximately fair panel decision, measured through the core, remains fair for the population. These are alternative *deliberation guarantees* for the decision reached by the panel. Therefore, we ask:

[1] While some prior works have implications for panel complexity, to the best of our knowledge, our work is the first to systematically study the panel complexity problem. We contrast our results with prior work in Section A.

[2] Their scope also includes Fair Greedy Capture and welfare guarantees for its decisions. These results differ from core fairness of the eventual decision; we discuss the broader scope in Section A.

Question 3: *What is the panel complexity for deliberation guarantees beyond exact welfare maximization?*

In this work, we address all three questions. We note that the social choice settings we study involve individual preferences—such as cost functions over budget allocations in participatory budgeting—which may not be easily observable or verifiable, unlike demographic features such as age or gender. For this reason, we focus our analysis on uniform panel selection rather than on algorithms that use individual preference information to select panels. The use of uniform panel selection naturally rules out the possibility of guarantees that hold for every panel in the support of the lottery—that is, ex-post guarantees; see Section A for more details—but it improves the robustness of our results by ensuring they do not rely on access to any individual's information. An open direction for future work is to extend our model to allow non-uniform panel selection via demographic quotas, and to analyze how this affects panel complexity; see Section 4.

Contributions We first give a novel definition of representativeness, which serves as our key tool in deriving panel complexity bounds. We define ϵ-representative panels with respect to a given feature $f : [n] \to \mathcal{X}$, where $(\mathcal{X}, d)$ is a metric space. A panel is ϵ-representative if the distribution it induces over the values of f is at most ϵ away, in Wasserstein distance (also known as the earth mover's distance; see Definition 2.1), from the distribution induced by the entire population. This is a natural notion of representativeness and is particularly well-suited for panel complexity, unlike prior definitions in the literature [Flanigan et al., 2020, 2021a,b, 2024, Ebadian et al., 2022, Ebadian and Micha, 2025]. We are able to establish that a panel of size $O((1/\epsilon)^2 \cdot (\log \ell + \log(1/\delta)))$ is ϵ-representative with probability $1 - \delta$ for any sequence of ℓ real-valued features (Theorem 2.4). Notably, this guarantee is independent of both the population size and the specific features. This theorem is instrumental in deriving all of our upper bounds on panel complexity and we believe it will be useful for analyzing panel complexity in social choice problems beyond those we study.

We study the facility location problem, where the goal is to place a single facility in a metric space $\mathcal{X}$. Each agent in the population is located at some point in $\mathcal{X}$ and seeks to minimize their distance to the facility. Our first result shows that the probability of a panel selecting a facility location far from the optimal one is small. Notably, this result holds for arbitrary, possibly infinite, metric spaces. We then use this result to show that if the panel selects an efficient allocation among its members, it remains efficient for the entire population, giving an upper bound on the panel complexity of efficiency. Specifically, we show that if the panel size k is $O\left((1/\epsilon^2) \cdot t \cdot \log(1/\epsilon)\right)$, then the expected social cost is optimal up to a multiplicative factor of $1 + \epsilon$. Notably, this multiplicative error guarantee holds for any space with bounded dimension (see Definition 3.9 for a formal definition of the notion of dimension that we use) and is almost tight. Lastly, we extend our analysis to the case of locating multiple facilities. Assuming the panel minimizes total distance to its members, we show that in any metric space $\mathcal{X}$, the loss in social cost resulting from delegation to a panel is upper bounded by the Wasserstein distance between the panel and population distributions (Lemma 3.11). This result leads to a panel complexity guarantee for social cost minimization over the line $\mathcal{X} = [0, 1]$ that is independent of the number of facilities to be selected. This application uses Wasserstein representativeness directly, beyond controlling the mean of each feature separately; Section C.1 discusses the distinction.

We also study the (divisible) participatory budgeting problem, where the goal is to allocate a budget B across m projects and each agent has a Lipschitz cost function over the space of possible budget allocations $[0, 1]^m$. We give panel complexity results on efficiency and fairness. Notably, each of the results in this section give meaningful bounds when the panel is assumed only to approximately achieve its objective (or efficiency or fairness). Due to space constraints, we defer this section (Section E) as well as most of the proofs of statements contained in the main body, to appendices.

2 Framework for Panel Complexity

In this section, we develop our framework for analyzing panel complexity, which we later use to derive panel complexity guarantees for facility location (Section 3) and participatory budgeting (Section E). The following subsections introduce the sortition model and present a novel definition of representative panels for establishing upper bounds on panel complexity.

2.1 Model of Sortition

For any $\ell \in \mathbb{N}$, let $[\ell] = \{1, \ldots, \ell\}$ denote the set of the first ℓ natural numbers. We represent a *population* of n agents as $[n]$ and refer to any subset $S \subseteq [n]$ as a *panel*. We consider features $f : [n] \to \mathcal{X}$, assigning to each agent $i \in [n]$ a value in a metric space $(\mathcal{X}, d)$. Equivalently, we may view f as a vector $(f_1, \ldots, f_n) \in \mathcal{X}^n$. These features may capture demographic characteristics or encode agents' preferences. For instance, a feature $f : [n] \to \mathcal{X}$ may represent agent i's location in an instance of the facility location problem. Similarly, a feature $f : [n] \to [0, 1]$ may specify agent i's utility for a given budget allocation in a participatory budgeting problem.

We briefly discuss our choice to consider metric spaces. To assess representativeness with respect to a given feature, we need a measure of similarity between feature values, and a metric space provides a natural formalization. Any categorical domain can be equipped with the discrete metric, under which two values are either identical or different. More informative metrics capture degrees of similarity—for example, geographical distance or absolute differences between numerical characteristics. Thus, the assumption does not require features to be numerical or Euclidean. Metric spaces are less well-equipped in scenarios with asymmetry. For example, a voter favoring a moderate policy may view an advocate of a stronger policy as an acceptable representative, while the more extreme voter may reject the moderate advocate as a substitute. A metric cannot express this asymmetry because distances are symmetric.

We study the properties of randomly selected panels of size k, where $1 \leq k \leq n$. Let $\mathcal{U}_{k,n}$ denote the uniform distribution over the space $\binom{n}{k}$ of such panels. Throughout this work, we assume that panels $S \sim \mathcal{U}_{k,n}$ are drawn uniformly at random.

2.2 Representative Panels

Consider a single feature $f : [n] \to \mathcal{X}$, where $(\mathcal{X}, d)$ is a metric space. This feature defines a discrete probability distribution over $\mathcal{X}$, determined by the frequency of each outcome. Formally, for each subset $S \subseteq [n]$, we define:

$$\phi_f^S = \frac{1}{|S|} \sum_{i \in S} \delta(f_i).$$

Here, f_i denotes the value of feature f assigned to agent i, and $\delta(f_i)$ represents the Dirac mass at f_i, i.e., the distribution ϕ_f^S assigns a probability mass of $\ell/|S|$ to each $x \in \mathcal{X}$, where ℓ is the number of occurrences of x in the vector $(f_i)_{i \in S}$. We refer to:

$\phi_f^{[n]}$ as the *population distribution* of f; and ϕ_f^S, for a panel $S \subseteq [n]$, as the *panel distribution* of f.

The Wasserstein distance, also known as the earth mover's distance or the Kantorovich-Rubinstein distance, originates from optimal transport theory Villani [2008], Chewi et al. [2024]. In general, the Wasserstein-p distances form a family of metrics that generalize the notion of optimal transport. In this work, we focus exclusively on the Wasserstein-1 distance, which we simply refer to as the Wasserstein distance. The Wasserstein distance is based on the idea that two probability measures are close if one can be transformed into the other by moving a small amount of probability mass. It is defined as follows:

Definition 2.1 (Wasserstein Distance)**.** *Let ϕ and ψ be discrete probability distributions supported on a metric space $(\mathcal{X}, d)$, with finite supports. The Wasserstein distance between ϕ and ψ is given by*

$$W(\phi, \psi) = \min_{\gamma \in \Gamma(\phi, \psi)} \mathbb{E}_{(x,y) \sim \gamma}[d(x, y)] = \min_{\gamma \in \Gamma(\phi, \psi)} \sum_{x \in \mathrm{supp}(\phi)} \sum_{y \in \mathrm{supp}(\psi)} \gamma(x, y) \cdot d(x, y), \quad (1)$$

where $\Gamma(\phi, \psi)$ is the set of all couplings *of ϕ and ψ, i.e., $\Gamma(\phi, \psi)$ includes every joint distribution γ over $\mathcal{X}^2$ such that $\sum_{y \in \mathrm{supp}(\psi)} \gamma(x, y) = \phi(x)$ for all $x \in \mathrm{supp}(\phi)$ and $\sum_{x \in \mathrm{supp}(\phi)} \gamma(x, y) = \psi(y)$ for all $y \in \mathrm{supp}(\psi)$.*

The Wasserstein distance has applications in computer vision and machine learning, where it is used to quantify the similarity between images based on pixel frequencies Arjovsky et al. [2017], Ni et al. [2009], Ruzon and Tomasi [2001].

We now define our notion of representativeness based on the Wasserstein distance.

Definition 2.2 (Representative Panels)**.** *We say that a panel $S \subseteq [n]$ is ϵ-representative of a given feature $f : [n] \to \mathcal{X}$ if the Wasserstein distance between the population distribution $\phi_f^{[n]}$ and the panel distribution ϕ_f^S is at most ϵ, i.e., it holds that $W(\phi_f^{[n]}, \phi_f^S) \leq \epsilon$.*

For instance, panel distributions close to the population distribution include those obtained by moving each point by a distance of less than ϵ or by shifting an ϵ-fraction of all points by a distance of 1.

Prior work introduced alternative notions of panel representativeness Ebadian and Micha [2025], Ebadian et al. [2022], but both approaches rely on assumptions about agents' preferences over panels. In contrast, our notion of representativeness takes a more direct approach, using the underlying population as a benchmark. Crucially, this benchmark remains unchanged as the panel size varies, enabling an analysis of panel complexity.

Remark 2.3 (Minimizing Sum of Distances is Not Representative)**.** *Ebadian and Micha [2025] consider a model in which agents are embedded in a similarity metric space. They argue that, intuitively, selecting a panel to minimize the sum of distances from each agent to their nearest panel member does not necessarily yield a representative panel [Ebadian and Micha, 2025, Example 7]. To illustrate this, they construct an example with two agents located at extreme positions A and B, far from all other agents and from each other. The remaining $n-2$ agents are evenly split between two locations, C and D, which are 10 units apart.*

For a panel size of $k = 3$, this distance-minimizing method would select the two outliers at A and B, along with only one agent from the majority group, resulting in a highly non-representative panel. This is problematic because most of the population resides in either C or D.

Using our notation, this example corresponds to a feature vector $f = (-\infty, -5, \ldots, -5, 5, \ldots, 5, \infty)$, where the number of -5 values is $(n-1)/2$ and the number of 5 values is also $(n-1)/2$. To satisfy our definition of ϵ-representativeness (Definition 2.2) for any $\epsilon > 0$, the panel would need to include the entire population, i.e., it must be that $k = n$. Thus, our definition aligns with the intuition of Ebadian and Micha [2025].

In Section C.1, we present an in-depth technical analysis of our notion of representativeness. In particular, we derive general reductions from our panel selection model to independent sampling, along with a general concentration inequality. These tools allow us to establish a panel complexity bound for any collection of real-valued features, which plays a key role in deriving our main technical results. We summarize this bound in the following theorem.

Theorem 2.4 (Panel Complexity of Representativeness)**.** *Consider $\mathcal{X} = [0, 1]$ equipped with the Euclidean distance and let $f_1, \ldots, f_\ell$ be features such that $f_i : [n] \to \mathcal{X}$ for all $i \in [\ell]$. Let the panel size be:*

$$k = O\left(\left(\frac{1}{\epsilon}\right)^2 \left(\log \ell + \log\left(\frac{1}{\delta}\right)\right)\right).$$

Then it holds that:

$$\mathbb{P}_{S \sim \mathcal{U}_{k,n}}\left[S \text{ is } \epsilon\text{-representative of } f_j \text{ for all } j \in [\ell]\right] \geq 1 - \delta.$$

The proof of the theorem builds on the results developed in Section C.1 and is deferred to Section C.2.

3 Facility Location via Sortition

In this section, we analyze the facility location problem. First, we define the model. Next, we establish tail bounds for the panel distribution. We then analyze the panel complexity of social welfare maximization before extending our model to the case of multiple facilities. The proofs omitted in this section are deferred to Section D.

3.1 Model of Facility Location

An instance of facility location is a tuple $\langle \mathcal{X}, \mathcal{C}, d, (x_1, \ldots, x_n)\rangle$, where $(\mathcal{X}, d)$ denotes a metric space $\mathcal{X}$ equipped with a distance function $d : \mathcal{X} \times \mathcal{X} \to [0, 1]$. A single facility must be placed at some point $q \in \mathcal{C} \subseteq \mathcal{X}$ and each agent $i \in [n]$ has an ideal location $x_i \in \mathcal{X}$, and incurs a cost equal to the distance from q, given by $d(q, x_i)$.

Definition 3.1 (Social Cost for Facility Location)**.** *For a panel S, write $k = |S|$. We define:*

(i) The social cost of $q \in \mathcal{C}$ as $\textsc{Social-Cost}(q) = (1/n) \cdot \sum_{i=1}^{n} d(q, x_i)$*,*
(ii) The optimal social cost as $\textsc{Social-Opt} = \min_{q \in \mathcal{C}} \textsc{Social-Cost}(q)$*,*
(iii) The panel cost of $q \in \mathcal{C}$ w.r.t. a panel S as $\textsc{Panel-Cost}(q, S) = (1/k) \cdot \sum_{i \in S} d(q, x_i)$*, and*
(iv) The optimal panel cost as $\textsc{Panel-Opt}(S) = \min_{q \in \mathcal{C}} \textsc{Panel-Cost}(q, S)$*.*

Throughout this section, we consider the optimal facility location $\overline{q}(S)$ with respect to S. When we need to specify the facility location instance I, we write $\overline{q}(S, I)$. We also denote by $q^\star$ the optimal facility location for the entire population. Formally:

$$\overline{q}(S) \in \arg\min_{q \in \mathcal{C}} \textsc{Panel-Cost}(q, S) \quad \text{and} \quad q^\star \in \arg\min_{q \in \mathcal{C}} \textsc{Social-Cost}(q).$$

3.2 Bounding Panel Outliers in Facility Location

A key question in evaluating the panel-optimal facility location $\overline{q}(S)$ for a panel S drawn uniformly from the population is how far it deviates from the population-optimal facility $q^\star$. In this section, we analyze the behavior of $d(\overline{q}(S), q^\star)$ as a function of the panel size k.

This analysis is useful for obtaining guarantees on social cost in Section 3.3, and offers an alternative perspective on approximating the optimal facility location. If $q^\star$ is unique, we naturally expect $d(\overline{q}(S), q^\star) \to 0$ as $k \to n$. Moreover, if this function decays smoothly with k, it offers insight into how well a panel of size k approximates the optimal facility location.

To put this in more concrete terms, consider a government agency evaluating the feasibility of building a facility in a given region. While they may eventually be able to survey the entire population and determine the optimal location $q^\star$, logistical constraints such as land availability and permits may still pose challenges. The agency's primary concern is that, initially, they lack a clear understanding of where the facility should be placed. To gain confidence before making a large commitment, they first query a panel. This query returns a ball $B \subseteq \mathcal{X}$ centered at $\overline{q}(S)$, which, with some probability dependent on k, contains $q^\star$. If B proves to be a viable location and the agency decides to proceed, no resources are wasted, as they can refine their analysis by surveying the remaining population $[n] \setminus S$.

Our main result in this section is a panel complexity guarantee for general metric spaces. We provide a tail bound on the probability that $d(q^\star, \overline{q}(S))$ is large relative to $\textsc{Social-Opt}$ in general metric spaces. Note that, since the distances d are bounded by 1, it follows that $\textsc{Social-Opt} \leq 1$.

Theorem 3.2 (Panel Complexity Upper Bound)**.** *Let $\langle \mathcal{X}, \mathcal{C}, d, (x_1, \ldots, x_n) \rangle$ be a facility location instance. Fix any $T > 2$ and $\delta \in (0, 1)$. Let the panel size be:*

$$k = \frac{2 \cdot \log(1/\delta)}{\log\left(T^2/(4(T-1))\right)}.$$

Then it holds that:

$$\mathbb{P}_{S \sim \mathcal{U}_{k,n}}[d(q^\star, \overline{q}(S)) \leq T \cdot \textsc{Social-Opt}] \geq 1 - \delta.$$

We note that the denominator in the panel complexity is always positive, i.e., for any $T > 2$, we have $\log\left(T^2/(4(T-1))\right) > 0$. Furthermore, as $T \to \infty$, the denominator grows like $\log(T)$. In Section F.2, we complement Theorem 3.2 by showing that the guarantee is tight, in the sense that the claim does not hold for any $T \leq 2$.

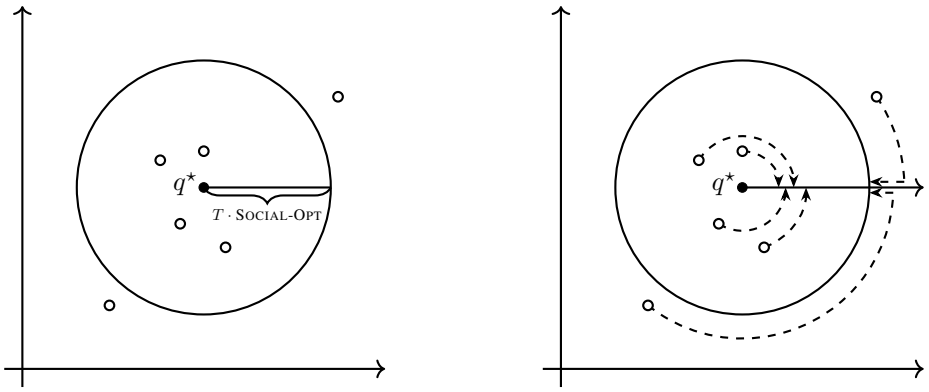
We note that the theorem above implies that, under an additional PAC-style guarantee with multiplicative error, we can obtain a multiplicative approximation of the optimal social cost.

Corollary 3.3 (Multiplicative PAC Guarantee Suffices)**.** *Let $\langle \mathcal{X}, \mathcal{C}, d, (x_1, \ldots, x_n) \rangle$ be a facility location instance. For any $0 < \epsilon < 1$, suppose the panel size satisfies $k \geq O(\log(1/\epsilon))$, and it holds that:*

$$\mathbb{P}_{S \sim \mathcal{U}_{k,n}}[\textsc{Social-Cost}(\overline{q}(S)) \leq (1 + \epsilon) \cdot \textsc{Social-Opt}] \geq 1 - \epsilon.$$

Then it follows that:

$$\mathbb{E}_{S \sim \mathcal{U}_{k,n}}[\textsc{Social-Cost}(\overline{q}(S))] \leq (1 + 5\epsilon) \cdot \textsc{Social-Opt}.$$



(i) General Instance (ii) Finite Interval Instance

Figure 1: Illustration of the reductions described in Lemmas 3.7 and 3.8, as used in the proof of Theorem 3.2. In subfigure (i), the solid black circle represents the optimal facility location $q^\star$, while the empty circles indicate the agent locations in $\mathcal{X}$. In subfigure (ii), the dashed lines show the mapping of agent locations from the original metric space $\mathcal{X}$ to their corresponding positions in the metric space $\mathcal{Y} = [0, T]$.

The proof of Corollary 3.3 follows readily by applying Theorem 3.2 and the multiplicative PAC guarantee to $\mathbb{P}_{S\sim\mathcal{U}_{k,n}}[\text{SOCIAL-COST}(\bar{q}(S)) \geq x]$. We note that Anshelevich and Postl [2017] show that even a panel consisting of a single agent, commonly referred to as random dictatorship, provides a 3-approximation to the optimal social cost in expectation. In contrast, Corollary 3.3 can be applied to obtain a $(1+\epsilon)$-approximation for any $\epsilon > 0$.

In the following, we give an idea of the proof of Theorem 3.2 by presenting the technical results that lead to it. The first step is to reduce the problem to a metric space that is easier to analyze. The reduction must preserve certain properties on distances in order to yield an upper bound on $\mathbb{P}_{S\sim\mathcal{U}_{k,n}}[d(q^\star, \bar{q}(S)) \leq T \cdot \text{SOCIAL-OPT}]$. We find that a metric map centered at a specific point p, which we will now define, is sufficient for this purpose. It is essentially a contraction that also preserves distances with respect to p, which will correspond to the population optimum.

Definition 3.4 (Metric Map)**.** *Let (X, d_X) and (Y, d_Y) be two metric spaces and $p \in X$. We call a map $\pi : X \to Y$ a* metric map *centered at p if the following two conditions are satisfied:*

(i) $d_X(x_1, x_2) \geq d_Y(\pi(x_1), \pi(x_2))$ *for all* $x_1, x_2 \in X$*, and*
(ii) $d_X(p, x) = d_Y(\pi(p), \pi(x))$ *for all* $x \in X$*.*

The following technical lemma demonstrates that, if an alternative $\bar{q}$ is far from the population optimum, it must also incur significant social cost.

Lemma 3.5. *Let $I = \langle \mathcal{X}, \mathcal{C}, d, (x_1, \ldots, x_n)\rangle$ be a facility location instance. Let $q_1, q_2 \in \mathcal{X}$ and $T \geq 0$ be such that $d(q_1, q_2) = T \cdot \text{SOCIAL-COST}(q_1)$. Then, we have $\text{SOCIAL-COST}(q_2) > (T-1) \cdot \text{SOCIAL-COST}(q_1)$.*

Using this lemma, we can show that if the set of non-optimal alternatives is constrained to contain only outliers, in the sense that $d(q^\star, q) \geq T \cdot \text{SOCIAL-OPT}$ for any $q \neq q^*$, then finding a metric map centered at the population optimum suffices to define a valid reduction.

Lemma 3.6 (General Reduction)**.** *Let $T > 2$ and $I = \langle X, \mathcal{C}_1, d_X, (x_1, \ldots, x_n)\rangle$ be a facility location instance with population optimum $q_1^\star \in \mathcal{C}_1$. Suppose that for all $c \in \mathcal{C}_1 \setminus \{q_1^\star\}$, we have $d_X(q_1^\star, c) \geq T \cdot \text{SOCIAL-OPT}(I)$. Let (Y, d_Y) be a metric space, and let $\pi : X \to Y$ be a metric map centered at $q_1^\star$. Then, there exists an instance $J = \langle Y, \mathcal{C}_2, d_Y, (y_1, \ldots, y_n)\rangle$ such that:*

(i) for any $S \subseteq [n]$, if $d(\bar{q}(S, I), q^\star) \geq T \cdot \text{SOCIAL-OPT}(I)$, then $d_Y(\bar{q}(S, J), q^\star) \geq T \cdot \text{SOCIAL-OPT}(J)$,
(ii) $\mathcal{C}_2 = \{\pi(q_1^\star)\} \cup \{y \in Y : d_Y(\pi(q_1^\star), y) \geq T \cdot \text{SOCIAL-OPT}(J)\}$*, and*
(iii) the population optimum in J is $\pi(q_1^\star) \in \mathcal{C}_2$.

Proof. We construct J by defining $\mathcal{C}_2 = \{\pi(q_1^\star)\} \cup \{y \in Y : d_Y(\pi(q_1^\star), y) \geq T \cdot \textsc{Social-Opt}(I)\}$ and $y_i = \pi(x_i)$ for all $i \in [n]$. By Property (ii) of Definition 3.4, we have $\textsc{Social-Cost}(\pi(q_1^\star), J) = \textsc{Social-Cost}(q_1^\star, I) = \textsc{Social-Opt}(I)$. By the definition of J, it follows that $d_Y(\pi(q_1^\star), y) > 2 \cdot \textsc{Social-Opt}(I)$ for all $y \in \mathcal{C}_2 \setminus \{\pi(q_1^\star)\}$. Thus, by Lemma 3.5, we have $\textsc{Social-Cost}(y, J) > \textsc{Social-Cost}(\pi(q_1^\star), J)$ for all $y \in \mathcal{C}_2 \setminus \{\pi(q_1^\star)\}$, and therefore $\pi(q_1^\star)$ is the population optimum in J. Consequently, $\textsc{Social-Opt}(J) = \textsc{Social-Opt}(I)$.

Let $S \subseteq [n]$ be such that $\bar{q}(S, I) \neq q_1^\star$. Then, we get:

$$\begin{aligned}
\sum_{i \in S} d_Y(\pi(q_1^\star), y_i) &= \sum_{i \in S} d_Y(\pi(q_1^\star), \pi(x_i)) && \text{(by definition of } y_i\text{)} \\
&= \sum_{i \in S} d_X(q_1^\star, x_i) && \text{(by Property (ii) of Definition 3.4)} \\
&> \sum_{i \in S} d_X(\bar{q}(S, I), x_i) && \text{(since } \bar{q}(S, I) \neq q_1^\star\text{)} \\
&\geq \sum_{i \in S} d_Y(\pi(\bar{q}(S, I)), y_i) && \text{(by Property (i) of Definition 3.4)}
\end{aligned}$$

implying $\bar{q}(S, J) \neq \pi(q_1^\star)$. This concludes the proof. □

Finally, we show how to make use of Lemma 3.6 by defining a specific metric map that projects an arbitrary metric space to $[0, \infty)$. Figure 1(i) gives a visualization of this projection.

Lemma 3.7 (Reduction to the Real Line)**.** *Let $T > 2$ and $I = \langle X, \mathcal{C}_1, d_X, (x_1, \ldots, x_n)\rangle$ be a facility location instance with population optimum $q^\star$. Then, there exists an instance $J = \langle Y, \mathcal{C}_2, d_Y, (y_1, \ldots, y_n)\rangle$, where $Y = [0, \infty)$ and $d_Y(z_1, z_2) = |z_1 - z_2|$, such that:*

(i) for any $S \subseteq [n]$, if $d(\bar{q}(S, I), q^\star) \geq T \cdot \textsc{Social-Opt}(I)$, then $d_Y(\bar{q}(S, J), q^\star) \geq T \cdot \textsc{Social-Opt}(J)$,
(ii) $\mathcal{C}_2 = \{0\} \cup [T \cdot \textsc{Social-Opt}(J), \infty)$, and
(iii) the population optimum in J is 0.

It is henceforth without loss of generality to assume that $\mathcal{X} = [0, \infty)$. But, in order to apply concentration bounds, we also cannot have points $x_i \to \infty$. Hence, as a last step, we further reduce the problem by mapping all points x_i that are larger than $T \cdot \textsc{Opt}$, to $T \cdot \textsc{Opt}$. This reduction to a finite interval instance is illustrated in Figure 1(ii).

Lemma 3.8 (Reduction to a Finite Interval)**.** *Let $T > 2$ and $I = \langle X, \mathcal{C}_1, d, (x_1, \ldots, x_n)\rangle$ be a facility location instance with $X = [0, \infty)$, $d(x, y) = |x - y|$, $\mathcal{C}_1 = \{0\} \cup [T \cdot \textsc{Social-Opt}(I), \infty)$, and population optimum at 0. Then, there exists $J = \langle Y, \mathcal{C}_2, d, (y_1, \ldots, y_n)\rangle$, where $Y = [0, T]$, such that:*

(i) for any $S \subseteq [n]$, if $d(\bar{q}(S, I), q^\star) \geq T \cdot \textsc{Social-Opt}(I)$, then $d_Y(\bar{q}(S, J), q^\star) \geq T \cdot \textsc{Social-Opt}(J)$,
(ii) $\mathcal{C}_2 = \{0, T\}$,
(iii) the population optimum in J is 0, and
(iv) $\textsc{Social-Opt}(J) \leq 1$.

3.3 Welfare Maximization in Facility Location

This section examines the social cost minimization problem in facility location. To present our main result, we first introduce the concept of Assouad dimension, which provides an upper bound on the covering number of balls at varying scales. For a formal definition of a covering in a metric space, see Definition B.2.

Definition 3.9 (Assouad dimension)**.** *We say that a metric space $(\mathcal{X}, d)$ has Assouad dimension t if for every $r < R$, we can cover any ball $B(x, R)$ for $x \in \mathcal{X}$ with $C \cdot (R/r)^t$ balls of radius r for some constant C.*

We are able to give panel complexity result for social cost minimization in facility location depending on the Assouad dimension of the metric space. Notably, any bounded subset of $\mathbb{R}^t$ has Assouad dimension t under any ℓ_p norm.

Theorem 3.10 (Panel Complexity Upper Bound)**.** *Let $\langle \mathcal{X}, \mathcal{C}, d, (x_1, \ldots, x_n)\rangle$ be a facility location instance, where $\mathcal{X}$ has Assouad dimension t. Let the panel size be:*

$$k = O\left(\left(\frac{1}{\epsilon}\right)^2 \left(t \log\left(\frac{1}{\epsilon}\right)\right)\right).$$

Then it holds that:

$$\mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[\textsc{Social-Cost}(\overline{q}(S))\right] \leq (1+\epsilon)\cdot \textsc{Social-Opt}.$$

Proof. Throughout this proof, we denote $\textsc{Opt} := \textsc{Social-Opt}$. Let T be a sufficiently large constant to ensure that Theorem 3.2 applies. Let $q^\star$ be the optimal facility location. Since $\mathcal{X}$ has Assouad dimension t, there is some sequence $q_1, \ldots, q_K \in B(q^\star, T\cdot \textsc{Opt})$ with $q_1 = q^\star$ such that $B(q^\star, T\cdot\textsc{Opt}) \subseteq \bigcup_{j=1}^K B(q_j, (\epsilon/3)\cdot\textsc{Opt})$ where $K = O((1/\epsilon)^t)$.

For $1 \leq j \leq K$, let $X_j = \textsc{Social-Cost}(q_j, S) - \textsc{Opt}$ and $Y_j = \textsc{Panel-Cost}(q_j, S) - \textsc{Panel-Cost}(q^\star, S) + (\epsilon/3)\cdot\textsc{Opt}$. Let us also define the following events:

$$A = \{S \subseteq [n] \,:\, \overline{q}(S) \in B(q^\star, T\cdot\textsc{Opt})\}, \qquad B_j = \{S\subseteq[n] \,:\, X_j \leq Y_j\}$$

for $1 \leq j \leq K$. We will argue that if $A \cap B_1 \cap \ldots \cap B_k$ holds, then $\textsc{Social-Cost}(\overline{q}(S)) \leq (1+\epsilon)\cdot\textsc{Opt}$ and that $\mathbb{P}[A\cap B_1\cap\ldots\cap B_K] \leq \epsilon$. From this the result follows by Corollary 3.3.

First, observe that $\mathbb{P}_{S\sim\mathcal{U}_{k,n}}[A] \geq 1-\epsilon/2$ by Theorem 3.2. For each $j\in[K]$ and $i\in[n]$, define a feature $f_j(i) = (1/\textsc{Opt})\cdot(d(x_i,q_j) - d(x_i,q^\star))$. Note that:

$$(1/|S|)\cdot\sum_{i\in S} f_j(i) = (1/\textsc{Opt})\cdot(\textsc{Panel-Cost}(q_j,S) - \textsc{Panel-Cost}(q^\star,S))$$
$$(1/n)\cdot\sum_{i\in[n]} f_j(i) = (1/\textsc{Opt})\cdot(\textsc{Social-Cost}(q_j,S) - \textsc{Opt}).$$

If S is $(\epsilon/3)$-representative with respect to feature f_j, then $S\subseteq B_j$ because $(1/|S|)\cdot\sum_{i\in S} f_j(i) \leq (1/n)\cdot\sum_{i\in[n]} f_j(i) + \epsilon/3$. Multiplying both sides by $\textsc{Opt}$ establishes the claim. Moreover, since $d(q^\star, q_i) \leq T\cdot\textsc{Opt}$, it follows that $|f_j(i)| \leq (1/\textsc{Opt})\cdot(T\cdot\textsc{Opt}) = T$. By Theorem 2.4, the panel S is therefore $(\epsilon/3)$-representative with respect to all K features simultaneously, implying that $\mathbb{P}_{S\sim\mathcal{U}_{k,n}}[B_1\cap\ldots\cap B_k] \geq 1-\epsilon/2$.

Finally, consider any panel $S\in A\cap B_1\cap\ldots\cap B_k$. Let j be such that $\overline{q}(S)\in B(q_j,(\epsilon/3)\cdot\textsc{Opt})$. Observe that the following derivation holds:

$$\begin{aligned}
\textsc{Social-Cost}(\overline{q}(S)) - \textsc{Opt} &\leq \textsc{Social-Cost}(q_j) - \textsc{Opt} + (1/3)\cdot\epsilon\cdot\textsc{Opt}\\
&\leq \textsc{Panel-Cost}(q_j,S) - \textsc{Panel-Cost}(q^\star,S) + (2/3)\cdot\epsilon\cdot\textsc{Opt}\\
&\leq \textsc{Panel-Cost}(\overline{q}(S),S) - \textsc{Panel-Cost}(q^\star,S) + \epsilon\cdot\textsc{Opt}\\
&= \textsc{Panel-Opt} - \textsc{Panel-Cost}(q^\star,S) + \epsilon\cdot\textsc{Opt}\\
&\leq 0 + \epsilon\cdot\textsc{Opt}
\end{aligned}$$

The first and third inequalities follow from $d(\overline{q}(S), q_j) \leq (\epsilon/3)\cdot\textsc{Opt}$, the second inequality follows from $S\in B_j$, and the fourth inequality follows from the choice of $\overline{q}(S)$. This concludes the proof. □

In Theorem F.4 of Section F.2, we also establish a nearly matching lower bound for the metric space $([0,1]^t, \ell_\infty)$, which has an Assouad dimension of t.

3.4 Extension to Multiple Facilities

In this section, we consider a setting where we can place facilities at multiple locations. An instance in this case is defined by $\langle \mathcal{X}, \mathcal{C}, d, \ell, (x_1,\ldots,x_n)\rangle$, where, as before, $(\mathcal{X}, d)$ is a metric space, $\mathcal{C}\subseteq\mathcal{X}$ is the set of possible facility locations, and $(x_1,\ldots,x_n)$ are the ideal locations of the agents in $[n]$. The additional parameter $\ell \geq 1$ specifies the desired number of facilities.

Given a set $y = (y_1, \ldots, y_\ell)$ of ℓ facility locations, it is natural to define the cost of agent i as the distance to the closest facility. That is, we define $\text{COST}_i(y) = \min_{j \in [\ell]} d(x_i, y_j)$. The social and panel costs, SOCIAL-COST(q) and PANEL-COST(q, S), along with their respective optimal values, SOCIAL-OPT and PANEL-OPT, are defined analogously to Definition 3.1. We use $\overline{y}(S) \in \mathcal{C}^\ell$ and $y^* \in \mathcal{C}^\ell$ to denote the panel-optimal and social-optimal collection of facility locations, respectively.

Our first result applies to general metric spaces and shows that if we treat the locations of individuals in the population as a feature, the expected social cost can be bounded as follows.

Lemma 3.11. *Let $\langle \mathcal{X}, \mathcal{C}, d, \ell, (x_1, \ldots, x_n) \rangle$ be a multiple facility location instance. Define $f(i) = x_i$ as the feature representing each agent's location. Let $\phi_f^{[n]}$ denote the population distribution of this feature and ϕ_f^S the panel distribution. Then, we have:*

$$\mathbb{E}_{S \sim \mathcal{U}_{k,n}} \left[\text{SOCIAL-COST}(\overline{y}(S))\right] \leq \text{SOCIAL-OPT} + \mathbb{E}_{S \sim \mathcal{U}_{k,n}} \left[W(\phi_f^{[n]}, \phi_f^S)\right]$$

This lemma is closely tied to our notion of representativeness. In particular, if a panel is ϵ-representative, it yields an additive bound on the expected social cost. The next result leverages Lemma 3.11 to upper bound the panel complexity required to obtain such an additive bound in the special case where $\mathcal{X} = [0, 1]$.

Theorem 3.12 (Panel Complexity Upper Bound)**.** *Let $\langle \mathcal{X}, \mathcal{C}, d, \ell, (x_1, \ldots, x_n) \rangle$ be a multiple facility location instance with $\mathcal{X} = [0, 1]$. Let the panel size be $k = O(1/\epsilon^2)$. Then it holds that:*

$$\mathbb{E}_{S \sim \mathcal{U}_{k,n}} \left[\text{SOCIAL-COST}(\overline{y}(S))\right] \leq \text{SOCIAL-OPT} + \epsilon.$$

The proof of the result above relies on Lemma C.1 and on known results regarding the Wasserstein metric and is deferred to Appendix D.2. Finally, in Theorem F.5 of Section F.2 we show that in the multiple facility location problem, achieving a multiplicative error guarantee is impossible.

4 Conclusion and Open Problems

In this work, we developed a framework for analyzing panel complexity and provided efficiency and fairness guarantees for two social choice problems: participatory budgeting and facility location. Our results open several interesting directions for future research.

First, throughout this work, we focused on the case where the panel is selected uniformly at random from the population. However, in many practical settings, stratified sampling is necessary. Extending the panel complexity analysis to non-uniform lotteries over panels remains an open problem. More fundamentally, our notion of representativeness based on the Wasserstein metric may prove a useful way of framing representativeness in contexts transcending uniform sampling.

Second, we applied our framework to participatory budgeting and facility location, but many other important social choice problems remain to be studied, such as indivisible participatory budgeting Rey and Maly [2023]. Micha and Shah [2020] give a result in the setting of centroid clustering which can be interpreted as a panel complexity of fairness in the multiple facilities extension we study in Section 3.4. Their notion of fairness, proportional fairness, was first defined by Chen, Fain, Lyu, and Munagala [2019] and is inspired by the notion of core. The result of Micha and Shah [2020] applies to Euclidean space with ℓ_2 distances, and uses a VC-dimension bound. Since this technique does not easily extend to other metrics, the panel complexity of fairness in multiple facility location on general metric spaces is an interesting open problem.

Third, we point out that our bounds do not depend on the population size, n. An open question is whether they can be improved when the population size is bounded. Finally, we do not explicitly define the panel decision functions, assuming only that they satisfy certain efficiency or fairness guarantees. Understanding how specific panel decision functions, such as concrete voting rules for participatory budgeting or facility location, impact panel complexity is an interesting direction for future work. We note, however, that our lower bounds on panel complexity hold for any panel decision function.

Acknowledgments and Disclosure of Funding

This project has been partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 866132), by an Amazon Research Award, by the Israel Science Foundation Breakthrough Program (grant No. 2600/24), and by a grant from TAU Center for AI and Data Science (TAD). SF acknowledges financial support for travel from his current institution, AI4I (the Italian Institute of AI).

Competing interests: The authors declare no competing interests.

References

I. Anagnostides, D. Fotakis, and P. Patsilinakos. Metric-distortion bounds under limited information. *J. Artif. Intell. Res.*, 74:1449–1483, 2022.

E. Anshelevich and J. Postl. Randomized social choice functions under metric preferences. *J. Artif. Intell. Res.*, 58:797–827, 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017.

H. Aziz and N. Shah. Participatory budgeting: Models and approaches. *Pathways Between Social Science and Computational Social Science: Theories, Methods, and Interpretations*, pages 215–236, 2021.

H. Aziz, B. E. Lee, S. Morota Chu, and J. Vollen. Proportionally representative clustering. In *WINE*, 2024.

C. Baharav and B. Flanigan. Fair, manipulation-robust, and transparent sortition. In *EC*, pages 756–775. ACM, 2024.

G. Benadè, P. Gölz, and A. D. Procaccia. No stratification without representation. In *EC*, pages 281–314. ACM, 2019.

I. Caragiannis and K. Fehrs. The complexity of learning approval-based multiwinner voting rules. In *AAAI*, pages 4925–4932. AAAI Press, 2022.

I. Caragiannis, E. Micha, and J. Peters. Can a few decide for many? the metric distortion of sortition. In *ICML*. OpenReview.net, 2024.

H. Chan, A. Filos-Ratsikas, B. Li, M. Li, and C. Wang. Mechanism design for facility location problems: A survey. In *IJCAI*, pages 4356–4365. ijcai.org, 2021.

X. Chen, B. Fain, L. Lyu, and K. Munagala. Proportionally fair clustering. In *International Conference on Machine Learning (ICML)*, pages 1032–1041. PMLR, 2019.

Y. Cheng, S. Dughmi, and D. Kempe. Of the people: Voting is more effective with representative candidates. In *EC*, pages 305–322. ACM, 2017.

Y. Cheng, S. Dughmi, and D. Kempe. On the distortion of voting with multiple representative candidates. In *AAAI*, pages 973–980. AAAI Press, 2018.

S. Chewi, J. Niles-Weed, and P. Rigollet. Statistical optimal transport. *CoRR*, abs/2407.18163, 2024.

P. Dey and A. Bhattacharyya. Sample complexity for winner prediction in elections. In *AAMAS*, pages 1421–1430. ACM, 2015.

P. Duetting, M. Feldman, T. Ponitka, and E. Soumalias. The pseudo-dimension of contracts. *CoRR*, abs/2501.14474, 2025.

S. Ebadian and E. Micha. Boosting sortition via proportional representation. In *AAMAS*, 2025.

S. Ebadian, G. Kehne, E. Micha, A. D. Procaccia, and N. Shah. Is sortition both representative and fair? In *NeurIPS*, 2022.

B. Fain, A. Goel, and K. Munagala. The core of the participatory budgeting problem. In *WINE*, volume 10123 of *Lecture Notes in Computer Science*, pages 384–399. Springer, 2016.

B. Fain, A. Goel, K. Munagala, and S. Sakshuwong. Sequential deliberation for social choice. In *WINE*, volume 10660 of *Lecture Notes in Computer Science*, pages 177–190. Springer, 2017.

B. Fain, A. Goel, K. Munagala, and N. Prabhu. Random dictators with a random referee: Constant sample complexity mechanisms for social choice. In *AAAI*, pages 1893–1900. AAAI Press, 2019.

B. Fain, W. Fan, and K. Munagala. Concentration of distortion: The value of extra voters in randomized social choice. In *IJCAI*, pages 110–116. ijcai.org, 2020.

J. S. Fishkin. *Democracy When the People Are Thinking: Revitalizing Our Politics through Public Deliberation*. Oxford University Press, 2018. ISBN 0-19-255189-2.

B. Flanigan, P. Gölz, A. Gupta, and A. D. Procaccia. Neutralizing self-selection bias in sampling for sortition. In *NeurIPS*, 2020.

B. Flanigan, P. Gölz, A. Gupta, B. Hennig, and A. D. Procaccia. Fair algorithms for selecting citizens' assemblies. *Nat.*, 596(7873):548–552, 2021a.

B. Flanigan, G. Kehne, and A. D. Procaccia. Fair sortition made transparent. In *NeurIPS*, pages 25720–25731, 2021b.

B. Flanigan, J. Liang, A. D. Procaccia, and S. Wang. Manipulation-robust selection of citizens' assemblies. In *AAAI*, pages 9696–9703. AAAI Press, 2024.

D. K. Foley. Lindahl's Solution and the Core of an Economy with Public Goods. *Econometrica: Journal of the Econometric Society*, pages 66–72, 1970.

J. Gastil and E. O. Wright. *Legislature by lot: Transformative designs for deliberative governance*. Verso Books, 2019.

A. Goel, A. K. Krishnaswamy, S. Sakshuwong, and T. Aitamurto. Knapsack voting for participatory budgeting. *ACM Trans. Economics and Comput.*, 7(2):8:1–8:27, 2019.

A. Goel, M. Goyal, and K. Munagala. Metric distortion of small-group deliberation. In *To Appear in STOC 2025*, 2025.

C. Gorski. The mathematics of jury size. *Inside Science*, March 2012. URL `https://ww2.aip.org/inside-science/the-mathematics-of-jury-size`.

M. Goyal, S. Sakshuwong, S. Sarmasarkar, and A. Goel. Low sample complexity participatory budgeting. In *ICALP*, volume 261 of *LIPIcs*, pages 70:1–70:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023.

D. Gross and V. Nesme. Note on sampling without replacing from a finite collection of matrices. *CoRR*, abs/1001.2738, 2010.

D. Halpern, A. D. Procaccia, E. Shapiro, and N. Talmon. Federated assemblies. In *AAAI*, pages 13897–13904. AAAI Press, 2025.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

P. Indyk. Sublinear time algorithms for metric space problems. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pages 428–434. ACM, 1999. doi: 10.1145/301250.301366. URL `https://people.csail.mit.edu/indyk/metrics99.ps`.

Y. H. Kalayci, D. Kempe, and V. Kher. Proportional representation in metric spaces and low-distortion committee selection. In *AAAI*, pages 9815–9823. AAAI Press, 2024.

L. Kellerhals and J. Peters. Proportional fairness in clustering: A social choice perspective. *Advances in Neural Information Processing Systems*, 2025.

R. Meir, F. Sandomirskiy, and M. Tennenholtz. Representative committees of peers. *J. Artif. Intell. Res.*, 71:401–429, 2021.

E. Micha and N. Shah. Proportionally fair clustering revisited. In *ICALP*, volume 168 of *LIPIcs*, pages 85:1–85:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

K. Ni, X. Bresson, T. F. Chan, and S. Esedoglu. Local histogram based segmentation using the wasserstein distance. *Int. J. Comput. Vis.*, 84(1):97–111, 2009.

OECD. *Innovative Citizen Participation and New Democratic Institutions: Catching the Deliberative Wave*. OECD Publishing, Paris, 2020. doi: 10.1787/339306da-en.

S. H. Owen and M. S. Daskin. Strategic facility location: A review. *European journal of operational research*, 111(3):423–447, 1998.

B. Polaczyk. Concentration bounds for sampling without replacement and hoeffding statistics. *CoRR*, abs/2301.03096, 2023.

S. Rey and J. Maly. The (computational) social choice take on indivisible participatory budgeting. *CoRR*, abs/2303.00621, 2023.

M. A. Ruzon and C. Tomasi. Edge, junction, and corner detection using color distributions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1281–1295, 2001.

H. Sambale and A. Sinulis. Concentration inequalities on the multislice and for sampling without replacement. *Journal of Theoretical Probability*, 35:2712–2737, 2022.

A. Tamir. The k-centrum multi-facility location problem. *Discrete Applied Mathematics*, 109(3): 293–307, 2001.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

T. Walsh and L. Xia. Lot-based voting rules. In *AAMAS*, pages 603–610. IFAAMAS, 2012.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *CoRR*, abs/1707.00087, 2017.

M. S. Williams. *Voice, Trust, and Memory: Marginalized Groups and the Failings of Liberal Representation*. Princeton University Press, 2000.

I. M. Young. *Inclusion and Democracy*. Oxford University press on demand, 2002.

A Further Related Work

Panel Complexity. Several works have established panel complexity-type results, though none focus on it as their primary objective. Caragiannis, Micha, and Peters [2024] and Goel, Goyal, and Munagala [2025] both provide a panel complexity bound in the facility location setting. Unlike the result of Caragiannis, Micha, and Peters [2024], the result of Goel, Goyal, and Munagala [2025] does not depend on the number of alternatives, but their model differs significantly from our own: panels can be sampled from a continuum of agents an infinite number of times and agents collectively reveal a ranking over all alternatives.

In the setting of centroid clustering, Chen, Fain, Lyu, and Munagala [2019] give a result which can be interpreted as a panel complexity result with respect to fairness and their bound depends on the number of candidate centroids. When every point is a candidate centroid, clustering is equivalent to the setting of Section 3.4. In this setting, Micha and Shah [2020] give a panel complexity bound for ℓ_2 distances. Extending this result to general metric spaces remains open; see Section 4.

Sampling also underlies classical algorithms for metric 1-median. For example, Indyk [1999] uses sampled pairwise cost comparisons and a tournament over the finite input points. Such an algorithm has different information access from a single uniformly sampled citizens' panel choosing its own optimum from a possibly infinite alternative space. Our bounds measure the participants needed for that panel decision, rather than the running time or distance-query complexity of an optimization algorithm.

Sortition. The computer science literature studying sortition consists of two main threads. Starting with the work of Benadè, Gölz, and Procaccia [2019], the first thread [Baharav and Flanigan, 2024, Flanigan et al., 2024, 2020, 2021a,b] focuses on algorithms which select a panel from a pool of random invitees who have agreed to participate. In the presence of selection bias, these selection algorithms then pursue representativeness by complying with quotas while also carefully balancing agents' selection probabilities. Other works refine these algorithms by taking into consideration other desiderata such as manipulation-robustness and transparency [Baharav and Flanigan, 2024, Flanigan et al., 2024].

The other major thread of sortition research [Caragiannis et al., 2024, Cheng et al., 2017, 2018, Ebadian et al., 2022, Ebadian and Micha, 2025, Meir et al., 2021, Goel et al., 2025], which our work falls within, takes the perspective that sortition should sample directly from the entire population [Gastil and Wright, 2019]. Within this line of work, a number of papers have looked at the expected distortion—taking the ratio of expected social cost to the optimal social cost—of a decision made by a sortition-selected panel [Caragiannis et al., 2024, Meir et al., 2021, Anagnostides et al., 2022, Goel et al., 2025]. Rather than quota compliance, works in this strain of literature give novel definitions of representativeness. The definitions of Ebadian, Kehne, Micha, Procaccia, and Shah [2022] and Ebadian and Micha [2025] both frame representativeness in terms of the latent preferences of the population of agents. To do this, they parameterize their definitions by q, assuming that a panel incurs a cost to each agent equal to their distance from the q-th closest panel member. Ebadian, Kehne, Micha, Procaccia, and Shah [2022] then take a distortion-like approach to measuring representativeness. Ebadian and Micha [2025] give a property inspired more by the notion of proportional representation, which can be thought of as belonging to a broader class of proportional representation properties for centroid clustering [Aziz et al., 2024, Kellerhals and Peters, 2025, Kalayci et al., 2024]. Our notion is better-suited for evaluating panel size and does not reason about agents' preferences over panels. We comment further on how our approach to representativeness differs from theirs in Section 2.2.

We note that while uniform panel selection studied in this work is a standard and well-established model of sortition [Caragiannis et al., 2024, Cheng et al., 2017, 2018, Ebadian et al., 2022, Ebadian and Micha, 2025, Meir et al., 2021, Goel et al., 2025], prior work has also considered non-uniform selection methods—such as the Fair Greedy Capture method proposed by Ebadian and Micha [2025]—which are designed to provide ex-post guarantees that hold for every panel in the support of the lottery. In particular, Caragiannis, Micha, and Peters [2024] analyze both uniform selection and Fair Greedy Capture. For M alternatives, their Theorem 4 gives ex-ante distortion $1+\epsilon$ using $O(\epsilon^{-2}\log(M/\epsilon))$ uniformly selected agents. Their Theorem 11 (Theorem 12 in the original arXiv version) gives the same distortion target for Fair Greedy Capture using $O(\epsilon^{-3}\log(M/\epsilon))$ agents, alongside a constant ex-post distortion guarantee for $k \geq 3$. Fair Greedy Capture uses agents' metric locations to form the panel while retaining equal individual selection probabilities. These results concern welfare of the selected decision; selection fairness and representation of groups on the panel do not themselves establish core fairness of that decision. Our uniform-panel model avoids collecting population-wide preference information, and our core-transfer guarantees address the decision's fairness directly.

Sample Complexity and Social Choice. The literature on sample complexity in metric social choice settings has primarily focused on proving distortion guarantees for a constant number of samples [Fain et al., 2017, 2019, Goyal et al., 2023]. In contrast, Fain, Fan, and Munagala [2020] give results for an arbitrary number of samples k. They give constant bounds for higher moments of distortion. This approach demonstrates the value of larger panels but does not readily offer a way of selecting a panel size. Interestingly, similar to our Theorem 3.2, they give a result showing that their mechanism deviates from a constant approximation to the optimum with probability decaying exponentially in the number of samples. Despite the similarity, their result follows from Markov's inequality, which does not suffice in our case. We note that our results have a similar flavor to those of Dey and Bhattacharyya [2015] and Caragiannis and Fehrs [2022]. However, these works focus on voting settings with ordinal and approval-based preferences, respectively, in addition to a finite number of alternatives.

In this paper, we study the sortition panel complexity of participatory budgeting, facility location, and multiple facility location. In particular, we study a participatory budgeting setting in which a divisible resource is allocated among projects with no fixed costs, commonly referred to as divisible participatory budgeting (for a survey, see Aziz and Shah [2021]). This is in contrast to indivisible participatory budgeting, in which funding decisions are binary [Rey and Maly, 2023]. Facility

location and its multiple facilities extension are both classical problems in operations research [e.g., Owen and Daskin, 1998, Tamir, 2001]) and have been the topic of numerous recent works (see Chan et al. [2021] for a survey).

B Metric Space Preliminaries

The following are basic definitions of metric spaces used throughout this work.

Definition B.1 (Metric Space)**.** *A pair $(\mathcal{X}, d)$, where $d : \mathcal{X} \times \mathcal{X} \to [0, 1]$, is a metric space if:*

1. *$d(x, x) = 0$ for all $x \in \mathcal{X}$ (identity),*
2. *$d(x, y) = d(y, x)$ for all $x, y \in \mathcal{X}$ (symmetry),*
3. *$d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in \mathcal{X}$ (triangle inequality).*

The r-ball centered at $x \in \mathcal{X}$ is defined as $B_r(x) = \{y \in \mathcal{X} \mid d(x, y) \leq r\}$.

We also define a covering of a metric space.

Definition B.2 (ϵ-Covering)**.** *A set $(x_1, \ldots, x_k) \in \mathcal{X}^k$ is an ϵ-covering of $(\mathcal{X}, d)$ if the union of ϵ-balls covers the entire space, i.e., $\mathcal{X} \subseteq \bigcup_{i=1}^{k} B_\epsilon(x_i)$. The covering number $N(\mathcal{X}, d, \epsilon)$ is the minimal size of an ϵ-covering of $(\mathcal{X}, d)$.*

Remark B.3 (Covering Number of Euclidean Space)**.** *For every $p \geq 0$ and $m \geq 1$, the Euclidean space $([0, 1]^m, \|\cdot\|_p)$ is defined as $([0, 1]^m, d)$, where $d(x, y) = \|x - y\|_p$. The covering number of the Euclidean space satisfies $N([0, 1]^m, \|\cdot\|_p, \epsilon) = O((1/\epsilon)^m)$.*

C Additional Material for Section 2

C.1 General Tools for Representative Panels

In this section, we prove general results showing that if k is sufficiently large, a randomly selected panel of size k is ϵ-representative for one or more features with high probability. From a technical perspective, it is useful to note the following basic fact: drawing a panel $S \sim \mathcal{U}_{k,n}$ is equivalent to sampling k individuals *without replacement* from the population $[n]$.

Since concentration bounds are typically easier to derive for independent samples drawn *with replacement*, we begin by establishing a reduction for the expected Wasserstein distance in this setting. The following lemma formalizes the intuition that sampling without replacement more closely reflects the underlying population than sampling with replacement; see Remark C.7 for a more detailed discussion.

Lemma C.1 (Reduction to Independent Sampling)**.** *Let $f : [n] \to \mathcal{X}$ be a feature. Define $\mathcal{R}_{k,n}$ as the probability distribution over multisets of size k drawn from $[n]$ by sampling with replacement. Then, we have:*

$$\mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[W(\phi_f^{[n]}, \phi_f^S)\right] \leq \mathbb{E}_{S\sim\mathcal{R}_{k,n}}\left[W(\phi_f^{[n]}, \phi_f^S)\right].$$

The proof follows from the convexity of Wasserstein distance (Lemma C.3) and the argument of Hoeffding [1963, Theorem 4], as specified in the following lemmas. Although Hoeffding [1963] established the result only for the case $V = \mathbb{R}$, it has been noted by Polaczyk [2023] and Gross and Nesme [2010] that the proof extends to any normed vector space without modification.

Lemma C.2 (Hoeffding [1963])**.** *Let V be a normed vector space, $x_1, \ldots, x_n \in V$, and $\psi : V \to \mathbb{R} \cup \{+\infty\}$ be a convex function. Then, it holds that:*

$$\mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[\psi\left(\sum_{i\in S} x_i\right)\right] \leq \mathbb{E}_{S\sim\mathcal{R}_{k,n}}\left[\psi\left(\sum_{i\in S} x_i\right)\right].$$

To apply Lemma C.2, we first establish the convexity of the Wasserstein distance.

Lemma C.3 (Convexity)**.** *Let ϕ, ψ_1, ψ_2 be discrete probability distributions over a metric space $(\mathcal{X}, d)$, and let $\psi = t \cdot \psi_1 + (1-t) \cdot \psi_2$ be their convex combination for some $t \in (0,1)$. Then, the following holds:*

$$W(\phi, \psi) \leq t \cdot W(\phi, \psi_1) + (1-t) \cdot W(\phi, \psi_2).$$

Proof. Let γ_1 and γ_2 be optimal couplings attaining the minimum in the definition of $W(\phi, \psi_1)$ and $W(\phi, \psi_2)$, respectively. That is,

$$\gamma_1 \in \operatorname*{arg\,min}_{\gamma \in \Gamma(\phi, \psi_1)} \mathbb{E}_{(x,y)\sim\gamma}[d(x,y)] \quad \text{and} \quad \gamma_2 \in \operatorname*{arg\,min}_{\gamma \in \Gamma(\phi, \psi_2)} \mathbb{E}_{(x,y)\sim\gamma}[d(x,y)].$$

Consider a coupling $\gamma = t \cdot \gamma_1 + (1-t) \cdot \gamma_2$. It is easy to see that $\gamma \in \Gamma(\phi, \psi)$ since, given any $A \subseteq \mathcal{X}$, it holds that:

$$\begin{aligned}\gamma(\mathcal{X} \times A) &= t \cdot \phi(A) + (1-t) \cdot \phi(A) = \phi(A),\\ \gamma(A \times \mathcal{X}) &= t \cdot \psi_1(A) + (1-t) \cdot \psi_2(A) = \psi(A).\end{aligned}$$

It follows that:

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\gamma}\left[d(x,y)\right] &= t \cdot \mathbb{E}_{(x,y)\sim\gamma_1}\left[d(x,y)\right] + (1-t) \cdot \mathbb{E}_{(x,y)\sim\gamma_2}\left[d(x,y)\right]\\ &= t \cdot W(\phi, \psi_1) + (1-t) \cdot W(\phi, \psi_2).\end{aligned}$$

which concludes the proof. □

We are now ready to prove Lemma C.1.

Proof. Let $F = \{f_1, \ldots, f_n\}$, and consider the normed vector space $(\mathbb{R}^F, \|\cdot\|_1)$. We interpret any $v \in \mathbb{R}^F$ satisfying $v_j \geq 0$ for all $j \in F$ and $\sum_{j\in F} v_j = 1$ as a probability distribution over F, assigning mass v_j to the element $j \in F$. We define $\psi : \mathbb{R}^F \to \mathbb{R} \cup \{+\infty\}$ by:

$$\psi(v) = \begin{cases} W(\phi_f^{[n]}, v) & \text{if } v_j \geq 0 \text{ for all } j \in F \text{ and } \sum_{j\in F} v_j = 1,\\ +\infty & \text{otherwise.}\end{cases}$$

Next, for each $i \in [n]$ we define the vector $x_i \in \mathbb{R}^F$ by setting for each $j \in F$:

$$x_{i,j} = \begin{cases} 1/k & \text{if } j = f(i),\\ 0 & \text{otherwise.}\end{cases}$$

Then, by Lemma C.2, we obtain:

$$\begin{aligned}\mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[W(\phi_f^{[n]}, \phi_f^S)\right] &= \mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[\psi\left(\sum_{i\in S} x_i\right)\right]\\ &\leq \mathbb{E}_{S\sim\mathcal{R}_{k,n}}\left[\psi\left(\sum_{i\in S} x_i\right)\right]\\ &= \mathbb{E}_{S\sim\mathcal{R}_{k,n}}\left[W(\phi_f^{[n]}, \phi_f^S)\right].\end{aligned}$$

as needed. □

We also provide a counterexample to a stronger version of Lemma C.1, demonstrating that the (first-order) stochastic dominance condition does not hold in general; see Remark C.8.

In addition, we establish the following concentration inequality for the Wasserstein distance. The proof relies on the bounded differences property of the Wasserstein distance by Weed and Bach [2017] and a McDiarmid-style inequality for sampling without replacement by Sambale and Sinulis [2022].

Lemma C.4 (Concentration Inequality)**.** *Let $f : [n] \to \mathcal{X}$ be a feature. Suppose that $d(x,y) \leq 1$ for all $x, y \in \mathcal{X}$. Then, we have:*

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{U}_{k,n}}\left[W(\phi_f^{[n]}, \phi_f^S) \geq \mu + t\right] &\leq \exp\left(-t^2 k/4\right)\\ \textit{where } \mu &= \mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[W(\phi_f^{[n]}, \phi_f^S)\right].\end{aligned}$$

The proof of the lemma relies on the following two results. The first follows directly from the proof of Weed and Bach [2017, Proposition 20].

Lemma C.5 (Weed and Bach [2017]). *Let $f : [n] \to \mathcal{X}$ be a feature. Suppose that $d(x, y) \leq 1$ for all $x, y \in \mathcal{X}$. Let $x_1, \ldots, x_k \in \mathcal{X}$ and $x'_i \in \mathcal{X}$. Then, it holds that:*

$$W(\phi_f^{[n]}, \phi_f^{(x_i, x_{-i})}) - W(\phi_f^{[n]}, \phi_f^{(x'_i, x_{-i})}) \leq 1/k.$$

The second one is the following McDiarmid-style inequality of Sambale and Sinulis [2022, Proposition 2].

Lemma C.6 (Sambale and Sinulis [2022]). *Let $g : [n]^k \to \mathbb{R}$ be a symmetric function; that is, for any permutation $\pi : [k] \to [k]$, we have $g(x_1, \ldots, x_k) = g(x_{\pi(1)}, \ldots, x_{\pi(k)})$. Suppose there exists $c > 0$ such that for all $x_1, \ldots, x_k, x'_i \in [n]$, it holds that $|g(x_i, x_{-i}) - g(x'_i, x_{-i})| \leq c$. Then, we have:*

$$\mathbb{P}_{S \sim \mathcal{U}_{k,n}} \left[g(x_S) \geq \mu + t \right] \leq \exp\left(\frac{-t^2}{4(1 - k/n)c^2 k} \right) \quad \text{where } \mu = \mathbb{E}_{S \sim \mathcal{U}_{k,n}} \left[g(x_S) \right].$$

We are now ready to prove Lemma C.4.

Lemma C.4 (Concentration Inequality). *Let $f : [n] \to \mathcal{X}$ be a feature. Suppose that $d(x, y) \leq 1$ for all $x, y \in \mathcal{X}$. Then, we have:*

$$\mathbb{P}_{S \sim \mathcal{U}_{k,n}} \left[W(\phi_f^{[n]}, \phi_f^S) \geq \mu + t \right] \leq \exp\left(-t^2 k/4 \right)$$
$$\text{where } \mu = \mathbb{E}_{S \sim \mathcal{U}_{k,n}} \left[W(\phi_f^{[n]}, \phi_f^S) \right].$$

Proof. Since $W(\phi_f^{[n]}, \cdot)$ is a symmetric function, and by Lemma C.5, we may apply Lemma C.6 with $c = 1/k$ to obtain:

$$\begin{aligned} \mathbb{P}_{S \sim \mathcal{U}_{k,n}} \left[W(\phi_f^{[n]}, \phi_f^S) \geq \mu + t \right] &\leq \exp\left(\frac{-t^2}{4(1 - k/n)c^2 k} \right) \\ &= \exp\left(\frac{-t^2 k}{4(1 - k/n)} \right) \leq \exp\left(\frac{-t^2 k}{4} \right) \end{aligned}$$

as needed. □

With these two lemmas, deriving panel complexity bounds for ϵ-representativeness of a feature valued in a general metric space reduces to bounding the expected Wasserstein distance for independent samples (drawn with replacement). We apply this approach for the proof of Theorem 2.4, appearing in the main body of the paper.

C.2 Proof of Theorem 2.4

Theorem 2.4 (Panel Complexity of Representativeness). *Consider $\mathcal{X} = [0, 1]$ equipped with the Euclidean distance and let $f_1, \ldots, f_\ell$ be features such that $f_i : [n] \to \mathcal{X}$ for all $i \in [\ell]$. Let the panel size be:*

$$k = O\left(\left(\frac{1}{\epsilon} \right)^2 \left(\log \ell + \log\left(\frac{1}{\delta} \right) \right) \right).$$

Then it holds that:

$$\mathbb{P}_{S \sim \mathcal{U}_{k,n}} \left[S \text{ is } \epsilon\text{-representative of } f_j \text{ for all } j \in [\ell] \right] \geq 1 - \delta.$$

Proof. For any feature $f : [n] \to [0, 1]$, we denote:

$$\mu_f = \mathbb{E}_{S \sim \mathcal{U}_{k,n}} \left[W(\phi_f^{[n]}, \phi_f^S) \right] \quad \text{and} \quad \mu'_f = \mathbb{E}_{S \sim \mathcal{R}_{k,n}} \left[W(\phi_f^{[n]}, \phi_f^S) \right].$$

For any feature $f : [n] \to [0, 1]$, by [Chewi et al., 2024, Proposition 2.1], we have $\mu'_f \leq \epsilon/2$ for $k = O(1/\epsilon^2)$. By Lemma C.1, it follows that $\mu_f \leq \epsilon/2$. Therefore, by Lemma C.4, we conclude that for any $j \in [\ell]$, it holds that:

$$\begin{aligned}
&\mathbb{P}_{S\sim\mathcal{U}_{k,n}}\left[S \text{ is not } \epsilon\text{-representative of } f_j\right] \\
&\leq \mathbb{P}_{S\sim\mathcal{U}_{k,n}}\left[W(\phi_{f_j}^{[n]}, \phi_{f_j}^{S}) \geq \mu_{f_j} + \epsilon/2\right] \\
&\leq \exp\left(\frac{-(\epsilon/2)^2 k}{4}\right) \leq \delta/\ell.
\end{aligned}$$

The result follows by the union bound. □

C.3 Additional Remarks

Remark C.7 (Comparing Sampling With and Without Replacement)**.** *Intuitively, sampling without replacement yields a more representative panel than sampling with replacement.*

Consider first the case where the population size n is large. By the union bound, the probability of selecting the same individual more than once when sampling with replacement is at most $\binom{k}{2}/n$. Thus, as n grows large, the two sampling schemes become nearly equivalent.

In the other extreme case of a small population with $n = 2$, sampling without replacement guarantees that a panel of size $k = 2$ includes both agents. In contrast, sampling with replacement requires a panel of size at least $\log(1/\delta) + 1$ to ensure, with probability at least $1 - \delta$, that both individuals are included. This illustrates that sampling with replacement can lead to higher panel complexity to achieve certain desirable guarantees.

Lemma C.1 extends this intuition beyond these extreme cases.

Remark C.8 (Counterexample to Stochastic Dominance)**.** *Using the notation of Lemma C.1, we show that there exists a population of n agents and a feature $f : [n] \to [0, 1]$ such that:*

$$\mathbb{P}_{S\sim\mathcal{R}_{k,n}}\left[W(\phi_f^{[n]}, \phi_f^S) \leq \delta\right] > \mathbb{P}_{S\sim\mathcal{U}_{k,n}}\left[W(\phi_f^{[n]}, \phi_f^S) \leq \delta\right] \quad \textit{for some } \delta > 0, \tag{2}$$

meaning that the (first-order) stochastic dominance condition may fail.

Consider a population of size $n = 5$ with feature values given by the vector $f = (0, 1/2, 1/2, 1/2, 1)$. Let the panel size be $k = 2$. We will demonstrate that:

$$\begin{aligned}
&\mathbb{P}_{S\sim\mathcal{R}_{k,n}}\left[W(\phi_f^{[n]}, \phi_f^S) \leq 0.2\right] \geq 9/25 = 0.36, \quad \textit{while} \\
&\mathbb{P}_{S\sim\mathcal{U}_{k,n}}\left[W(\phi_f^{[n]}, \phi_f^S) \leq 0.2\right] = 3/10 = 0.3.
\end{aligned} \tag{3}$$

This contradicts the inequality in Equation (2)*. We first compute the Wasserstein distance for different panels based on the feature values of their members:*

- *If $f(S) = \{0, 1/2\}$, then $W(\phi_f^S, \phi_f^{[n]}) = (1/2) \cdot (1/2 - 1/5) + (1/2) \cdot (1/5) = 1/4 = 0.25$.*
- *If $f(S) = \{1/2, 1/2\}$, then $W(\phi_f^S, \phi_f^{[n]}) = (1/2) \cdot (1/5) + (1/2) \cdot (1/5) = 1/5 = 0.2$.*
- *If $f(S) = \{0, 1\}$, then $W(\phi_f^S, \phi_f^{[n]}) = (1/2) \cdot (1/2 - 1/5) + (1/2) \cdot (1/2 - 1/5) = 0.3$.*

The case $f(S) = \{1/2, 1\}$ is symmetric to $f(S) = \{0, 1/2\}$.

Note that $W(\phi_f^S, \phi_f^{[n]}) \leq 0.2$ holds only when $f(S) = \{1/2, 1/2\}$. Thus, Equation (3) *follows since the probability of drawing two agents with feature value $1/2$ when sampling with replacement is $(3/5)^2 = 9/25$, and when sampling without replacement it is $\binom{3}{2}/\binom{5}{2} = 3/10$.*

Note that the calculations above yield only a lower bound on $\mathbb{P}_{S\sim\mathcal{R}_{k,n}}[W(\phi_f^{[n]}, \phi_f^S) \leq 0.2]$, since we did not consider panels with $f(S) = \{0, 0\}$ or $f(S) = \{1, 1\}$. However, these combinations occur with zero probability when sampling without replacement.

C.4 Empirical Mean of Representative Panels

Our notion of representativeness ensures that the (true) mean of a feature f over the population $[n]$ is close to its (empirical) mean over any ϵ-representative panel S.

Lemma C.9 (Empirical Mean of Representative Panels)**.** *Let $f : [n] \to [0,1]$ be a feature and $S \subseteq [n]$ be an ϵ-representative panel of f. Then, it holds that $|(1/n) \cdot \sum_{i \in [n]} f(i) - (1/|S|) \cdot \sum_{i \in S} f(i)| \leq \epsilon$.*

Proof. Let $\phi_f^{[n]}$ and ϕ_f^S denote the population and panel distributions with respect to f, respectively, and let $f(A) \subseteq [0,1]$ represent the image of a subset $A \subseteq [n]$. The result follows from the Kantorovich–Rubinstein duality [Villani, 2008], which provides an alternative expression for the Wasserstein distance:

$$W(\phi_f^{[n]}, \phi_f^S) = \sup_{||g||_{\mathsf{Lip}} \leq 1} \left\{ \sum_{y \in f([n])} g(y) \cdot \phi_f^{[n]}(y) - \sum_{z \in f(S)} g(z) \cdot \phi_f^S(z) \right\},$$

where the supremum is taken over all Lipschitz functions $g : [0,1] \to \mathbb{R}$ with Lipschitz constant at most 1, i.e., $||g||_{\mathsf{Lip}} \leq 1$. The functions $g_1(y) = y$ and $g_2(y) = -y$ clearly satisfy this condition. Together with the assumption of ϵ-representativeness, this implies:

$$\epsilon \geq W(\phi_f^{[n]}, \phi_f^S) \geq \sum_{y \in f([n])} y \cdot \phi_f^{[n]}(y) - \sum_{z \in f(S)} z \cdot \phi_f^S(z) = \frac{1}{n} \sum_{i \in [n]} f(i) - \frac{1}{|S|} \sum_{j \in S} f(j),$$

$$\epsilon \geq W(\phi_f^{[n]}, \phi_f^S) \geq \sum_{z \in f(S)} z \cdot \phi_f^S(z) - \sum_{y \in f([n])} y \cdot \phi_f^{[n]}(y) = \frac{1}{|S|} \sum_{j \in S} f(j) - \frac{1}{n} \sum_{i \in [n]} f(i).$$

The result follows. □

We note that one could alternatively consider a weaker notion of representativeness than Definition 2.2, by defining a panel to be ϵ-representative of a real-valued feature $f : [n] \to [0,1]$ if the assertion of Lemma C.9 holds. This weaker notion would suffice to establish some of our panel complexity guarantees in the following sections. Moreover, under this weaker definition, the panel complexity guarantee of Theorem 2.4 can be proved directly using Hoeffding's inequality for sampling without replacement.

However, our more general definition is significantly more powerful, as it enables us to extend the panel complexity guarantees for facility location to the problem of locating multiple facilities, as discussed in Section 3.4. For this result, the weaker notion is insufficient. This suggests that our more general framework may be an essential tool for the analysis of settings beyond those considered in this paper. Furthermore, the stronger notion—defined via the Wasserstein distance—more naturally captures representativeness and we believe it to be of independent interest, regardless of its role in proving our panel complexity results.

D Missing Proofs of Section 3

In this section, we report all proofs and details missing from the discussion of the results in Section 3.

D.1 Proofs of Section 3.2

Corollary 3.3 (Multiplicative PAC Guarantee Suffices)**.** *Let $\langle \mathcal{X}, \mathcal{C}, d, (x_1, \ldots, x_n) \rangle$ be a facility location instance. For any $0 < \epsilon < 1$, suppose the panel size satisfies $k \geq O(\log(1/\epsilon))$, and it holds that:*

$$\mathbb{P}_{S \sim \mathcal{U}_{k,n}}[\textsc{Social-Cost}(\overline{q}(S)) \leq (1+\epsilon) \cdot \textsc{Social-Opt}] \geq 1 - \epsilon.$$

Then it follows that:

$$\mathbb{E}_{S \sim \mathcal{U}_{k,n}}[\textsc{Social-Cost}(\overline{q}(S))] \leq (1 + 5\epsilon) \cdot \textsc{Social-Opt}.$$

Proof of Corollary 3.3. We analyze the following formulation of the expected panel cost:

$$\mathbb{E}_{S\sim\mathcal{U}_{k,n}}[\textsc{Social-Cost}(\overline{q}(S))] = \int_0^\infty \mathbb{P}_{S\sim\mathcal{U}_{k,n}}[\textsc{Social-Cost}(\overline{q}(S)) \geq x]\,\mathrm{d}x.$$

We split the analysis of this integral into different intervals of $x \in [0,\infty)$. We denote $\textsc{Opt} = \textsc{Social-Opt}$. First, since probabilities are always bounded by 1:

$$\int_0^{(1+\epsilon)\cdot\textsc{Opt}} \mathbb{P}_{S\sim\mathcal{U}_{k,n}}[\textsc{Social-Cost}(\overline{q}(S)) \geq x]\,\mathrm{d}x \leq (1+\epsilon)\cdot\textsc{Opt}.$$

Second, by the PAC-style guarantee with multiplicative error:

$$\int_{(1+\epsilon)\cdot\textsc{Opt}}^{4\cdot\textsc{Opt}} \mathbb{P}_{S\sim\mathcal{U}_{k,n}}[\textsc{Social-Cost}(\overline{q}(S)) \geq x]\,\mathrm{d}x \leq \int_{(1+\epsilon)\cdot\textsc{Opt}}^{4\cdot\textsc{Opt}} \epsilon\,\mathrm{d}x \leq 3\epsilon\cdot\textsc{Opt}.$$

Third, for $x > 4\cdot\textsc{Opt}$, we prepare to apply Theorem 3.2. To do so, observe that by triangle inequality

$$d(q^\star,\overline{q}(S)) + \textsc{Social-Cost}(q^\star) = d(q^\star,\overline{q}(S)) + \textsc{Opt} \geq \textsc{Social-Cost}(\overline{q}(S)),$$

implying for any $X > 0$,

$$\mathbb{P}_{S\sim\mathcal{U}_{k,n}}[\textsc{Social-Cost}(\overline{q}(S)) \geq (1+X)\cdot\textsc{Opt}] \leq \mathbb{P}_{S\sim\mathcal{U}_{k,n}}[d(q^\star,\overline{q}(S)) \geq X\cdot\textsc{Opt}].$$

This gives us

$$\begin{aligned}
&\int_{4\cdot\textsc{Opt}}^\infty \mathbb{P}_{S\sim\mathcal{U}_{k,n}}[\textsc{Social-Cost}(\overline{q}(S)) \geq x]\,\mathrm{d}x \\
&\quad= \int_{4\cdot\textsc{Opt}}^\infty \mathbb{P}_{S\sim\mathcal{U}_{k,n}}[\textsc{Social-Cost}(\overline{q}(S)) \geq (1+(x/\textsc{Opt}-1))\cdot\textsc{Opt}]\,\mathrm{d}x \\
&\quad\leq \int_{4\cdot\textsc{Opt}}^\infty \mathbb{P}_{S\sim\mathcal{U}_{k,n}}[d(q^\star,\overline{q}(S)) \geq (x/\textsc{Opt}-1)\cdot\textsc{Opt}]\,\mathrm{d}x.
\end{aligned}$$

For $x \geq 4\cdot\textsc{Opt}$, we wish to apply Theorem 3.2 with parameters $T = x/\textsc{Opt} - 1$ and $\delta = 3\epsilon/T^2$. First, observe that $T > 2$ since in this case $T \geq 4\cdot\textsc{Opt}/\textsc{Opt} - 1 = 3$. We then compute the required panel size as:

$$\frac{2\cdot\log(1/\delta)}{\log\left(T^2/(4(T-1))\right)} = \frac{\log\left(T^2/(3\epsilon)\right)}{\log(T/8)} = \frac{2\log(T)+\log(1/(3\epsilon))}{\log(T)-\log(8)} = O\left(\log\left(\frac{1}{\epsilon}\right)\right).$$

Thus, Theorem 3.2 gives:

$$\begin{aligned}
\int_{4\cdot\textsc{Opt}}^\infty \mathbb{P}_{S\sim\mathcal{U}_{k,n}}[d(q^\star,\overline{q}(S)) \geq (x/\textsc{Opt}-1)\cdot\textsc{Opt}]\,\mathrm{d}x &\leq \int_{4\cdot\textsc{Opt}}^\infty \frac{3\epsilon}{(x/\textsc{Opt}-1)^2}\,\mathrm{d}x \\
&= 3\epsilon\cdot\frac{\textsc{Opt}^2}{\textsc{Opt}-x}\Big|_{4\cdot\textsc{Opt}}^\infty = 3\epsilon\cdot\frac{\textsc{Opt}}{3} = \epsilon\cdot\textsc{Opt},
\end{aligned}$$

as desired. □

Lemma 3.5. *Let $I = \langle\mathcal{X},\mathcal{C},d,(x_1,\ldots,x_n)\rangle$ be a facility location instance. Let $q_1,q_2 \in \mathcal{X}$ and $T \geq 0$ be such that $d(q_1,q_2) = T\cdot\textsc{Social-Cost}(q_1)$. Then, we have $\textsc{Social-Cost}(q_2) > (T-1)\cdot\textsc{Social-Cost}(q_1)$.*

Proof. Denote by $S_1 = \{i \in [n] : d(q_1,q_2) \geq d(q_1,x_i)\}$, $S_2 = [n]\setminus S_2$ and $K = \textsc{Social-Cost}(q_1)$. We have:

$$\begin{aligned}
\sum_{i\in[n]} d(q_2,x_i) &\geq \sum_{i\in S_1}(d(q_1,q_2)-d(q_1,x_i)) + \sum_{i\in S_2}(d(q_1,x_i)-d(q_1,q_2)) \\
&\qquad\text{(by triangle inequality)} \\
&= (|S_1|-|S_2|)\cdot d(q_1,q_2) - \left(\sum_{i\in S_1} d(q_1,x_i) - \sum_{i\in S_2} d(q_1,x_i)\right)
\end{aligned}$$

$$
\begin{aligned}
&= (|S_1| - |S_2|) \cdot d(q_1, q_2) - n \cdot K + 2 \cdot \sum_{i \in S_2} d(q_1, x_i) \\
&\geq (|S_1| - |S_2|) \cdot d(q_1, q_2) - n \cdot K + 2 \cdot |S_2| \cdot d(q_1, q_2) \\
&\quad \text{(by definition of } S_2\text{)} \\
&= (|S_1| + |S_2|) \cdot d(q_1, q_2) - n \cdot K \\
&= (T-1) \cdot n \cdot K \\
&\quad \text{(since } d(q_1, q_2) = T \cdot K\text{)}.
\end{aligned}
$$

The result follows. □

Lemma 3.7 (Reduction to the Real Line)**.** *Let $T > 2$ and $I = \langle X, \mathcal{C}_1, d_X, (x_1, \ldots, x_n)\rangle$ be a facility location instance with population optimum $q^\star$. Then, there exists an instance $J = \langle Y, \mathcal{C}_2, d_Y, (y_1, \ldots, y_n)\rangle$, where $Y = [0, \infty)$ and $d_Y(z_1, z_2) = |z_1 - z_2|$, such that:*

(i) for any $S \subseteq [n]$, if $d(\bar{q}(S, I), q^\star) \geq T \cdot$ SOCIAL-OPT(I), then $d_Y(\bar{q}(S, J), q^\star) \geq T \cdot$ SOCIAL-OPT(J),
(ii) $\mathcal{C}_2 = \{0\} \cup [T \cdot$ SOCIAL-OPT$(J), \infty)$, and
(iii) the population optimum in J is 0.

Proof. Let $R = \langle X, \mathcal{C}_3, d_X, (x_1, \ldots, x_n)\rangle$ where $\mathcal{C}_3 = \{q^\star\} \cup \{x \in \mathcal{C} \mid d_X(q^\star, x) \geq T \cdot$ SOCIAL-OPT$(I)\}$. Take a subset $S \subseteq [n]$ such that $d_X(q^\star, \bar{q}(S, I)) \geq T \cdot$ SOCIAL-OPT(I). Since $\mathcal{C}_1 \subseteq \mathcal{C}$, and $\bar{q}(S, I) \in \arg\min_{q \in \mathcal{C}}$ PANEL-COST(q, S, I), we must have $\bar{q}(S, I_1) = \bar{q}(S, I)$. Now consider the metric space $Y = [0, \infty)$ with distance given by $d_Y(z_1, z_2) = |z_1 - z_2|$. Further, define the map $\pi : \mathcal{X} \to Y$ as $\pi(x) = d_X(x, q^\star)$ for all $x \in \mathcal{X}$. We verify that π satisfies the conditions of a metric map centered at $q^\star$. First, for any $x_1, x_2 \in \mathcal{X}$, the triangle inequality gives:

$$d_Y(\pi(x_1), \pi(x_2)) = |d_X(q^\star, x_1) - d_X(q^\star, x_2)| \leq d_X(x_1, x_2),$$

which establishes Property (i) of Definition 3.4. Next, for any $x \in \mathcal{X}$:

$$d_Y(\pi(q^\star), \pi(x)) = |d_X(q^\star, q^\star) - d_X(q^\star, x)| = d_X(q^\star, x).$$

verifying Property (ii) of Definition 3.4. Therefore, π is a valid metric map centered at $q^\star$. The desired result then follows directly from Lemma 3.6. □

Lemma 3.8 (Reduction to a Finite Interval)**.** *Let $T > 2$ and $I = \langle X, \mathcal{C}_1, d, (x_1, \ldots, x_n)\rangle$ be a facility location instance with $X = [0, \infty)$, $d(x, y) = |x - y|$, $\mathcal{C}_1 = \{0\} \cup [T \cdot$ SOCIAL-OPT$(I), \infty)$, and population optimum at 0. Then, there exists $J = \langle Y, \mathcal{C}_2, d, (y_1, \ldots, y_n)\rangle$, where $Y = [0, T]$, such that:*

(i) for any $S \subseteq [n]$, if $d(\bar{q}(S, I), q^\star) \geq T \cdot$ SOCIAL-OPT(I), then $d_Y(\bar{q}(S, J), q^\star) \geq T \cdot$ SOCIAL-OPT(J),
(ii) $\mathcal{C}_2 = \{0, T\}$,
(iii) the population optimum in J is 0, and
(iv) SOCIAL-OPT$(J) \leq 1$.

Proof. By scaling all x_i by a factor of SOCIAL-OPT(I), we may assume that SOCIAL-OPT$(I) = 1$. Moreover, by relabeling the agents, we may assume that $x_1 \leq x_2 \leq \ldots \leq x_n$. Let t be the smallest index such that $x_{t+1} \geq T$. We define instance J by setting $\mathcal{C}_2 = \{0, T\}$ and $y_i = \min(x_i, T)$.

First, we argue that the population optimum in J is 0. Since the population optimum in I is 0, we have:

$$\sum_{i \in [n]} |y_i - T| = \sum_{i=1}^{n} |T - x_i| - \sum_{i=t+1}^{n} |T - x_i| \geq \sum_{i=1}^{n} |x_i - 0| - \sum_{i=t+1}^{n} |T - x_i| = \sum_{i=1}^{n} |y_i - 0|$$

which proves the claim. It also follows that SOCIAL-OPT$(J) = \sum_{i \in [n]} y_i \leq \sum_{i \in [n]} x_i =$ SOCIAL-OPT$(I) = 1$.

Let $S \subseteq [n]$ be a subset such that $\bar{q}(S, I) \neq 0$. Define $q' = \bar{q}(S, I)$. Consider the function $f_S : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ defined by $f_S(z) = \sum_{i \in S} |x_i - z|$. Clearly, $f_S(z)$ is unimodal and decreasing at $z = 0$. By definition of $\bar{q}(S, I)$, we have that $\bar{q}(S, I) \in \arg\min_{z \in \mathcal{C}_1} f_S(z)$. If $q' > T$, then

$q' = \arg\min_{z \in \mathbb{R}_{\geq 0}} f_S(z)$. By unimodality, this implies $f_S(T) \leq f_S(0)$. Moreover, observe that the reduction in social cost from I to J is the same for both 0 and T, i.e., we have $\sum_{i \in S} x_i - \sum_{i \in S} y_i = \sum_{i \in S} |x_i - T| - \sum_{i \in S} |y_i - T|$. Thus, if $\bar{q}(S, I) \neq 0$, then T has lower social cost than 0 in instance I, and therefore also in instance J. Hence, if $\bar{q}(S, I) \neq 0$, it follows that $\bar{q}(S, J) = T$, and so $d_Y(\bar{q}(S, J), q^\star) \geq T \cdot \textsc{Social-Opt}(J)$. □

Finally, we are ready to prove the main theorem of Section 3.2 by a standard concentration argument. We note that, in order to obtain a valid bound for any $T > 2$, it is crucial to use the strongest version of Hoeffding's inequality, rather than the simplified version more commonly applied.

Theorem 3.2 (Panel Complexity Upper Bound)**.** *Let $\langle \mathcal{X}, \mathcal{C}, d, (x_1, \ldots, x_n) \rangle$ be a facility location instance. Fix any $T > 2$ and $\delta \in (0, 1)$. Let the panel size be:*

$$k = \frac{2 \cdot \log(1/\delta)}{\log\left(T^2/(4(T-1))\right)}.$$

Then it holds that:

$$\mathbb{P}_{S \sim \mathcal{U}_{k,n}}[d(q^\star, \bar{q}(S)) \leq T \cdot \textsc{Social-Opt}] \geq 1 - \delta.$$

Proof. By Lemmas 3.7 and 3.8, we only need to consider an instance $I = \langle X, \mathcal{C}, d, (y_1, \ldots, y_n) \rangle$ where $X = [0, T]$, $d(x, y) = |x - y|$, $\mathcal{C} = \{0, T\}$, $\textsc{Social-Opt}(I) \leq 1$, and the population optimum is $q^\star = 0$. Denote $\textsc{Opt} = \textsc{Social-Opt}(I)$. Observe that for any panel $S \subseteq [n]$, we have:

$$\bar{q}(S) = 0 \quad \iff \quad \sum_{i \in S} x_i \leq \sum_{i \in S} (T - x_i) \quad \iff \quad \sum_{i \in S} x_i \leq (T/2) \cdot k$$

It follows that:

$$\begin{aligned}
\mathbb{P}_{S \sim \mathcal{U}_{k,n}}\left[d(\bar{q}(S), q^\star) \geq T\right] &= \mathbb{P}_{S \sim \mathcal{U}_{k,n}}\left[\sum_{i \in S} x_i > (T/2) \cdot k\right] \\
&= \mathbb{P}_{S \sim \mathcal{U}_{k,n}}\left[\frac{1}{k} \sum_{i \in S} \frac{x_i}{T} - \frac{\textsc{Opt}}{T} > \frac{1}{2} - \frac{\textsc{Opt}}{T}\right] \\
&\leq \left(\left(\frac{\textsc{Opt}/T}{1/2}\right)^{1/2} \cdot \left(\frac{1 - \textsc{Opt}/T}{1/2}\right)^{1/2}\right)^k \qquad (4) \\
&\leq \left(\left(\frac{1/T}{1/2}\right)^{1/2} \cdot \left(\frac{1 - 1/T}{1/2}\right)^{1/2}\right)^k \qquad (5) \\
&= \left(\frac{4(T-1)}{T^2}\right)^{k/2}.
\end{aligned}$$

Inequality (4) follows from the fact that $x_i/T \in [0, 1]$ and $(1/n) \cdot \sum_{i \in [n]} x_i/T = \textsc{Opt}/T$, allowing us to apply Hoeffding's inequality [Hoeffding, 1963, Equation 2.1], which holds for sampling without replacement. Inequality (5) holds because the function $x(1 - x)$ is increasing on the interval $[0, 1/2]$, and thus $(\textsc{Opt}/T) \cdot (1 - \textsc{Opt}/T) \leq (1/T) \cdot (1 - 1/T)$, given that $\textsc{Opt} \leq 1$ and $T > 2$. Finally, the result follows by choosing k such that:

$$\delta = \left(\frac{4(T-1)}{T^2}\right)^{k/2} \quad \iff \quad k = \frac{2 \cdot \log(1/\delta)}{\log(T^2/(4(T-1)))}$$

which concludes the proof. □

D.2 Proofs of Section 3.4

Lemma 3.11. *Let $\langle \mathcal{X}, \mathcal{C}, d, \ell, (x_1, \ldots, x_n) \rangle$ be a multiple facility location instance. Define $f(i) = x_i$ as the feature representing each agent's location. Let $\phi_f^{[n]}$ denote the population distribution of this feature and ϕ_f^S the panel distribution. Then, we have:*

$$\mathbb{E}_{S \sim \mathcal{U}_{k,n}}\left[\textsc{Social-Cost}(\overline{y}(S))\right] \leq \textsc{Social-Opt} + \mathbb{E}_{S \sim \mathcal{U}_{k,n}}\left[W(\phi_f^{[n]}, \phi_f^S)\right]$$

Proof. Let S be any panel of size k and $\overline{y}(S)$ the set of facilities minimizing the social cost for S. In the following calculations, we will consider S fixed and write $\overline{y}$ in place of $\overline{y}(S)$. Let us denote, as usual, with δ_x the Dirac mass at point $x \in \mathcal{X}$. Our first remark is that, for any probability distribution $\lambda = \sum_{j\in[\ell]} \lambda_j \delta_{y_j}$ supported on the ℓ positions of the facilities in $\overline{y}$, the following holds: $\textsc{Social-Cost}(\overline{y}) \leq W(\phi_f^{[n]}, \lambda)$.

Indeed, $\textsc{Cost}_i(\overline{y})$ is equal to $\min_{j\in[\ell]} d(x_i, \overline{y}_j)$, and, as a consequence, $\textsc{Cost}_i(\overline{y}) \leq \sum_j \lambda_j \cdot d(x_i, y_j)$ since this is a convex combination of positive values. Moreover, if γ is a coupling of ϕ, λ, by definition we have that: $1/n = \sum_j \gamma(x_i, y_j)$. Since this holds for every coupling γ, it holds for the minimal one as well, implying the above inequality. Let us define the following function: $\overline{\alpha} : [n] \times S \to \mathcal{X}^2$ such that $\overline{\alpha}(i, j) = (x_i, \alpha(z_j))$, where $\alpha(z_j)$ is a facility that realizes the minimum distance from z_j, namely $d(z_j, \alpha(z_j)) = \textsc{Cost}_{z_j}(\overline{y})$ (we wrote z_j in place of the corresponding agent, to avoid confusion with the indexes). Now, call $\alpha(\phi_f^S)$ the distribution defined by pushing forward the uniform distribution over the k members of panel S through α. If γ is the coupling realizing the minimum in the definition of $W(\phi_f^{[n]}, \phi_f^S)$, then the pushforward measure $\overline{\alpha}(\gamma)$ is a valid coupling between ϕ and $\alpha(\phi_f^S)$ and the following holds:

$$
\begin{aligned}
\textsc{Social-Cost}(\overline{y}) &\leq W\left(\phi_f^{[n]}, \alpha(\phi_f^S)\right) \\
&\leq \sum_{i\in[n]} \sum_{j\in[\ell]} d(x_i, \overline{y}_j) \cdot \overline{\alpha}(\gamma)(x_i, \overline{y}_j) \\
&= \sum_{i\in[n]} \sum_{j\in[k]} d(x_i, \alpha(z_j)) \cdot \gamma(x_i, z_j) \\
&\leq \sum_{i\in[n]} \sum_{j\in[k]} d(x_i, z_j) \cdot \gamma(x_i, z_j) + \sum_{i\in[n]} \sum_{j\in[k]} d(z_j, \alpha(z_j)) \cdot \gamma(x_i, z_j) \\
&= W(\phi_f^{[n]}, \phi_f^S) + \sum_{i\in[n]} \sum_{j\in[k]} d(z_j, \alpha(z_j)) \cdot \gamma(x_i, z_j) \\
&= W(\phi_f^{[n]}, \phi_f^S) + \frac{1}{k} \sum_{j\in[k]} d(z_j, \alpha(z_j)) \\
&= W(\phi_f^{[n]}, \phi_f^S) + \textsc{Panel-Opt}(S),
\end{aligned}
$$

where $\textsc{Panel-Opt}(S) = (1/k) \cdot \sum_{j\in S} \textsc{Cost}_j(\overline{y})$. Now, by taking the expectation of the quantities above over the choice of the panel S, we get:

$$
\mathbb{E}_{S\sim\mathcal{U}_{k,n}} \left[\textsc{Social-Cost}(\overline{y}(S))\right] \leq \mathbb{E}_{S\sim\mathcal{U}_{k,n}} \left[W(\phi_f^{[n]}, \phi_f^S)\right] + \mathbb{E}_{S\sim\mathcal{U}_{k,n}} \left[\textsc{Panel-Opt}(S)\right]
$$

It remains to prove that the expected value of $\textsc{Panel-Opt}(S)$ is less than $\textsc{Social-Opt}$. Indeed, by comparing the panel cost of $\overline{y}$ with that of the social optimum y^*, we can write the following:

$$
\begin{aligned}
\mathbb{E}_{S\sim\mathcal{U}_{k,n}} [\textsc{Panel-Opt}(S)] &\leq \mathbb{E}_{S\sim\mathcal{U}_{k,n}} \left[\textsc{Panel-Cost}(y^\star, S)\right] \\
&= \frac{1}{k} \sum_{i\in[n]} \left(\binom{n-1}{k-1} \Big/ \binom{n}{k} \right) \cdot \textsc{Cost}_i(y^\star) = \textsc{Social-Opt},
\end{aligned}
$$

which concludes the proof. □

Next, we will prove the panel complexity upper bound, based on the lemma above.

Theorem 3.12 (Panel Complexity Upper Bound)**.** *Let $\langle \mathcal{X}, \mathcal{C}, d, \ell, (x_1, \ldots, x_n)\rangle$ be a multiple facility location instance with $\mathcal{X} = [0, 1]$. Let the panel size be $k = O(1/\epsilon^2)$. Then it holds that:*

$$
\mathbb{E}_{S\sim\mathcal{U}_{k,n}} \left[\textsc{Social-Cost}(\overline{y}(S))\right] \leq \textsc{Social-Opt} + \epsilon.
$$

Proof. Let us again consider the feature $f(i) = x_i$ for $i \in [n]$. We adopt the following notation:

$$
\mu_f = \mathbb{E}_{S\sim\mathcal{U}_{k,n}} \left[W(\phi_f^{[n]}, \phi_f^S)\right] \quad \text{and} \quad \mu'_f = \mathbb{E}_{S\sim\mathcal{R}_{k,n}} \left[W(\phi_f^{[n]}, \phi_f^S)\right].
$$

Applying the result of Chewi, Niles-Weed, and Rigollet [2024, Proposition 2.1], we can choose $k = O(1/\epsilon^2)$ such that $\mu'_f \leq \epsilon$. Then, by Lemma C.1, it follows that $\mu_f \leq \epsilon$. Finally, by Lemma 3.11, we obtain:

$$\mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[\textsc{Social-Cost}(\overline{y}(S))\right] \leq \textsc{Social-Opt} + \mu_f \leq \textsc{Social-Opt} + \epsilon$$

as needed. □

E Participatory Budgeting via Sortition

In this section, we study the participatory budgeting problem. We first define the model, then establish panel complexity guarantees related to efficiency in terms of social welfare, and finally analyze fairness in terms of the core.

E.1 Model of Participatory Budgeting

In this section, we describe the (divisible) participatory budgeting model [see, e.g., Goel et al., 2019, Fain et al., 2016]. An instance of the participatory budgeting problem is a tuple $\langle m, B, (\textsc{Cost}_1, \ldots, \textsc{Cost}_n)\rangle$ where $m \geq 2$ denotes the number of projects, $B \in \mathbb{R}_{>0}$ denotes the available budget, and $\textsc{Cost}_i : [0,1]^m \to [0,1]$ is the cost function of the i-th agent, mapping budget allocations to the agent's disutility. Our goal is to select a budget allocation $x = (x_1, \ldots, x_m) \in [0,1]^m$ such that $\sum_{i=1}^m x_i \leq B$. We denote by $\mathcal{X}_B = \{x \in [0,1]^m : \sum_{j=1}^m x_j \leq B\}$ the space of all feasible budget allocations. We make two important assumptions on the agent's cost functions:

(i) Monotonicity: $\textsc{Cost}_i(x) \geq \textsc{Cost}_i(y)$ for any $x, y \in [0,1]^m$ such that $x_j \leq y_j$ for every $j \in [m]$, and
(ii) Lipschitz continuity: $|\textsc{Cost}_i(x) - \textsc{Cost}_i(y)| \leq \|x-y\|_1$ where $\|x-y\|_1 = \sum_{j=1}^m |x_j - y_j|$.

Given a panel of size k with Lipschitz cost functions, a *panel decision function* $\overline{x}$ selects a feasible budget allocation, i.e., $\overline{x} : \mathcal{L}^k \to \mathcal{X}_B$ where $\mathcal{L} = \mathrm{Lip}([0,1]^m, [0,1])$ denotes the space of all Lipschitz functions from $[0,1]^m$ to $[0,1]$. For simplicity of notation, we will denote the output of the panel decision function $\overline{x}$, given the cost functions of a panel of voters S, by $\overline{x}(S) := \overline{x}((\textsc{Cost}_i)_{i\in S})$.

E.2 Welfare Maximization in Participatory Budgeting

In this section, we establish fundamental limits on the efficiency of participatory budgeting when decisions are made by a randomly selected panel rather than the full population.

We first introduce the following notation:

Definition E.1 (Social Cost for Participatory Budgeting)**.** *We define:*

(i) the social cost of $x \in [0,1]^m$ is denoted by $\textsc{Social-Cost}(x) = (1/n) \cdot \sum_{i=1}^n \textsc{Cost}_i(x)$,
(ii) the optimal social cost is denoted by $\textsc{Social-Opt} = \min_{x\in\mathcal{X}_B} \textsc{Social-Cost}(x)$,
(iii) the panel cost of $x \in [0,1]^m$ is denoted by $\textsc{Panel-Cost}(x, S) = (1/k) \cdot \sum_{i\in S} \textsc{Cost}_i(x)$, *and*
(iv) the optimal panel cost is denoted by $\textsc{Panel-Opt}(S) = \min_{x\in\mathcal{X}_B} \textsc{Panel-Cost}(x, S)$.

The following theorem provides an upper bound on the panel size required to ensure a near-optimal budget allocation. Specifically, it shows that if the panel size is sufficiently large, then selecting the best allocation based on panel preferences guarantees that the expected social cost is close to the optimal population-wide allocation.

Theorem E.2 (Panel Complexity Upper Bound)**.** *Fix any instance $\langle m, B, (\textsc{Cost}_1, \ldots, \textsc{Cost}_n)\rangle$ of the participatory budgeting problem. Let the panel size be:*

$$k = O\left(\left(\frac{1}{\epsilon}\right)^2 \left(m \log\left(\frac{1}{\epsilon}\right)\right)\right).$$

Suppose that panel $S \subseteq [n]$ of size k selects a budget allocation $\overline{x}(S)$ satisfying:

$$\textsc{Panel-Cost}(\overline{x}(S), S) \leq \rho \cdot \textsc{Panel-Opt}(S) + \tau.$$

for some $\rho \geq 1$ and $\tau \geq 0$, where ρ is a constant. Then it holds that:

$$\mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[\textsc{Social-Cost}(\overline{x}(S))\right] \leq \rho \cdot \textsc{Social-Opt} + \tau + \epsilon.$$

Proof. There exists an $\epsilon/(12\rho)$-covering $(x_1, \ldots, x_w)$ of $\mathcal{X}_B$ with size $w = O((1/\epsilon)^m)$; see Definition B.2 for a formal definition of a covering of a metric space. We define w features $f_j : N \to [0,1]$ for $j \in \{1, \ldots, w\}$ by setting $f_j(i) = \textsc{Cost}_i(x_j)$. By Theorem 2.4, for $k = O((1/\epsilon)^2 \cdot (\log w + \log(1/\epsilon)) = O((1/\epsilon)^2 \cdot (m \log(1/\epsilon)))$, with probability at least $1 - \epsilon/2$, the panel $S \sim \mathcal{U}_{k,n}$ is $(\epsilon/12)$-representative for all features $f_1, \ldots, f_w$. Fix any $(\epsilon/12)$-representative panel S.

We first claim that for any budget allocation $x \in [0,1]^m$, it holds that:

$$\left|\textsc{Social-Cost}(x) - \textsc{Panel-Cost}(x, S)\right| \leq \epsilon/(4\rho). \tag{6}$$

Indeed, by the definition of ϵ-covering, there exists some $1 \leq j \leq w$ such that $\|x - x_j\|_1 \leq \epsilon$. Thus:

$$\begin{aligned}
&\left|(1/n) \cdot \sum_{i=1}^{n} \textsc{Cost}_i(x) - (1/|S|) \cdot \sum_{i \in S} \textsc{Cost}_i(x)\right| \\
&\leq \left|(1/n) \cdot \sum_{i=1}^{n} \textsc{Cost}_i(x_j) - (1/|S|) \cdot \sum_{i \in S} \textsc{Cost}_i(x_j)\right| + \epsilon/(6\rho) \\
&\qquad \text{(since } \textsc{Cost}_i \text{ is Lipschitz)} \\
&= \left|(1/n) \cdot \sum_{i=1}^{n} f_j(i) - (1/|S|) \cdot \sum_{i \in S} f_j(i)\right| + \epsilon/(6\rho) \\
&\qquad \text{(by definition of } f_j(i)) \\
&\leq \epsilon/(4\rho) \\
&\qquad \text{(since } S \text{ is } (\epsilon/12)\text{-representative)}
\end{aligned}$$

Now, letting $x^\star \in \arg\min_{x \in \mathcal{X}_B} \textsc{Social-Cost}(x)$, we observe:

$$\begin{aligned}
\textsc{Social-Cost}(\overline{x}(S)) &\leq \textsc{Panel-Cost}(\overline{x}(S), S) + \epsilon/(4\rho) && \text{(by Equation (6))} \\
&\leq \rho \cdot \textsc{Panel-Cost}(x^\star, S) + \tau + \epsilon/(4\rho) && \text{(by the assumption on } \overline{x}(S)) \\
&\leq \rho \cdot \textsc{Social-Opt} + \tau + \epsilon/4 + \epsilon/(4\rho) && \text{(by Equation (6))} \\
&\leq \rho \cdot \textsc{Social-Opt} + \tau + \epsilon/2
\end{aligned}$$

Finally, we conclude:

$$\begin{aligned}
&\mathbb{E}_{S \sim \mathcal{U}_{k,n}}\left[\textsc{Social-Cost}(\overline{x}(S)) \cdot \mathbb{1}\left[S \text{ is } (\epsilon/12)\text{-representative}\right]\right] \\
&\qquad \leq \rho \cdot \textsc{Social-Opt} + \epsilon/2 + \tau, \\
&\mathbb{E}_{S \sim \mathcal{U}_{k,n}}\left[\textsc{Social-Cost}(\overline{x}(S)) \cdot \mathbb{1}\left[S \text{ is not } (\epsilon/12)\text{-representative}\right]\right] \\
&\qquad \leq \epsilon/2 \\
\implies \quad &\mathbb{E}_{S \sim \mathcal{U}_{k,n}}\left[\textsc{Social-Cost}(\overline{x}(S))\right] \leq \rho \cdot \textsc{Social-Opt} + \epsilon + \tau
\end{aligned}$$

which completes the proof. □

For the matching lower bound for the above theorem, see Theorem F.6.

E.3 Core Fairness in Participatory Budgeting

We now turn to the question of the panel complexity of fairness in participatory budgeting. In participatory budgeting, the *core* is a group fairness property that enforces quite a strong requirement while also being guaranteed to exist under some natural conditions [Fain et al., 2016]. Informally, a budget allocation belongs to the core when no group of agents could, given their proportional share of the budget, select an alternative budget allocation under which each agent in the group incurs a strictly lower cost. We define approximate core by introducing an additive error to the share of the budget given to a group of agents and both an additive and multiplicative error to the cost reduction requirement.

Definition E.3 (Approximate Core Fairness for Participatory Budgeting)**.** *Given a budget allocation $x \in \mathcal{X}_B$ and a subset of agents $S \subseteq [n]$, a subset $T \subseteq S$ is said to form an (η, τ, ρ)-blocking coalition*

for some $\eta, \tau \geq 0$ and $\rho \geq 1$ if there exists an alternative budget allocation $x' \in \mathcal{X}_B$ such that:

$$\sum_{j=1}^{m} x'_j/B + \eta \leq |T|/|S| \quad \textit{and} \quad \rho \cdot \textsc{Cost}_i(x') + \tau < \textsc{Cost}_i(x) \textit{ for all } i \in T.$$

A budget allocation $x \in [0,1]^m$ is in the (η, τ, ρ, S)-core if no (η, τ, ρ)-blocking coalition exists for x. The set of all budget allocations in the (η, τ, ρ, S)-core for a panel S is denoted by $\textsc{Panel-Core}_S(\eta, \tau, \rho)$*, while the set of all budget allocations in the $(\eta, \tau, \rho, [n])$-core for the entire population is denoted by* $\textsc{Social-Core}(\eta, \tau, \rho)$.

Hence, budget allocations in $\textsc{Social-Core}(0, 0, 1)$ are in the exact core. Budget allocations in the exact core are known to exist under mild assumptions on the agents' cost functions [Foley, 1970]. For instance, core existence is guaranteed under cost functions that have continuous derivatives and are strictly monotonic and strictly convex. In any case, the main theorem of this section, which we state and prove next, states a panel complexity result for the approximate core that applies to all participatory budgeting settings in which the core can be approximated to a constant factor.

Theorem E.4 (Panel Complexity Upper Bound)**.** *Fix any instance $\langle m, B, (\textsc{Cost}_1, \ldots, \textsc{Cost}_n)\rangle$ of the participatory budgeting problem. Let the panel size be:*

$$k = \Omega\left(\left(\frac{1}{\epsilon}\right)^2 \left(m \log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right).$$

Suppose that panel $S \subseteq [n]$ of size k selects a budget allocation satisfying:

$$\overline{x}(S) \in \textsc{Panel-Core}_S(\eta, \tau, \rho)$$

for some $\eta, \tau \geq 0$ and $\rho \geq 1$, where ρ is a constant.[3] *Then it holds that:*

$$\mathbb{P}_{S \sim \mathcal{U}_{k,n}}\left[\overline{x}(S) \in \textsc{Social-Core}(\eta + \epsilon, \tau + \epsilon, \rho)\right] \geq 1 - \delta.$$

Proof. Consider an $\epsilon/\min(4\rho, 2B)$-covering of $\mathcal{X}_B$; see Definition B.2 for a formal definition of a covering of a metric space. We have some $x_1, \ldots, x_K \in \mathcal{X}_B$ where $K = (1/\epsilon)^{O(m)}$ so that for every $x \in \mathcal{X}_B$ there is some $a \in [K]$ such that $\|x - x_a\|_1 \leq \epsilon/(4\rho)$ and $\|x - x_a\|_1 \leq \epsilon/(2B)$. For every $a, b \in [K]$ and $i \in [n]$, we define a feature:

$$f_{a,b}(i) = \mathbb{1}\left[\textsc{Cost}_i(x_a) > \rho \cdot \textsc{Cost}_i(x_b) + \tau + \epsilon/(4\rho)\right].$$

Since there are $K^2 = (1/\epsilon)^{O(m)}$ features, it is by Theorem 2.4 that, if the panel size is $k = O((1/\epsilon)^2 \cdot (\log(K^2) + \log(1/\delta))) = O((1/\epsilon)^2 \cdot (m \log(1/\epsilon) + \log(1/\delta)))$, then panel $S \subseteq [n]$ is $(\epsilon/2)$-representative of the features defined above with probability $1 - \delta$. Formally, this means that the following holds for each $a, b \in [K]$ with probability at least $1 - \delta$:

$$\left|\frac{1}{|S|}\sum_{i \in S} f_{a,b}(i) - \frac{1}{n}\sum_{i=1}^{n} f_{a,b}(i)\right| \leq \epsilon/2.$$

Due to this observation, we need only show that, for any $(\epsilon/2)$-representative panel $S \subseteq [n]$, it holds that $\overline{x}(S) \in \textsc{Panel-Core}_S(\eta, \tau, \rho) \implies \overline{x}(S) \in \textsc{Social-Core}(\eta + \epsilon, \tau + \epsilon, \rho)$. Let us fix an $(\epsilon/2)$-representative panel $S \subseteq [n]$. Suppose that an alternative $x \in [0,1]^m$ is not in the $(\eta + \epsilon, \tau + \epsilon, \rho)$-core for the whole population, i.e., $x \notin \textsc{Social-Core}(\eta + \epsilon, \tau + \epsilon, \rho)$. By the definition of approximate core, we know that there is some coalition $T \subseteq [n]$ and an alternative $y \in [0,1]^m$ such that:

$$\sum_{j=1}^{m} y_j/B + \eta + \epsilon \leq |T|/n \quad \text{and} \quad \rho \cdot \textsc{Cost}_i(y) + \tau + \epsilon < \textsc{Cost}_i(x) \text{ for all } i \in T.$$

Let $a \in [K]$ be such that $\|x - x_a\|_1 \leq \epsilon/(4\rho)$ and let $b \in [K]$ be such that $\|y - x_b\|_1 \leq \epsilon/(4\rho)$. Let $y' = x_b$. We will argue that $x \notin \textsc{Panel-Core}_S(\eta, \tau, \rho)$ by showing that there exists a (η, τ, ρ)-blocking coalition for y'. That is, we will show $\sum_{j=1}^{m} y'_j + \eta \leq |T'|/|S|$ where $T' = \{i \in S : \rho \cdot \textsc{Cost}_i(y') + \tau < \textsc{Cost}_i(x)\}$. Note that for every $i \in T$:

$$\rho \cdot \textsc{Cost}_i(x_b) + \tau \leq \rho\left[\textsc{Cost}_i(y) + \epsilon/(4\rho)\right] + \tau \qquad \text{(since } \|y - x_b\| \leq \epsilon/(4\rho)\text{)}$$

[3] If we were to extend the argument to non-constant ρ, the panel complexity would depend on ρ.

$$
\begin{aligned}
&< \textsc{Cost}_i(x) - \epsilon + \epsilon/4 && \text{(by the core violation)}\\
&\le \textsc{Cost}_i(x_a) - \epsilon + \epsilon/4 + \epsilon/(4\rho) && \text{(since } \|x - x_a\| \le \epsilon/(4\rho)\text{)}\\
&\le \textsc{Cost}_i(x_a) - \epsilon/(4\rho)
\end{aligned}
$$

which means that $f_{a,b}(i) = 1$ for all $i \in T$ and consequently $(1/n)\sum_{i=1}^{n} f_{a,b}(i) \ge |T|/n$.

At last, we see that:

$$
\begin{aligned}
\frac{|T'|}{|S|} &= \frac{1}{|S|}\sum_{i\in S} \mathbb{1}[\rho \cdot \textsc{Cost}_i(y') + \tau < \textsc{Cost}_i(x)] && \text{(by the definition of } T'\text{)}\\
&\ge \frac{1}{|S|}\sum_{i\in S} \mathbb{1}[\rho \cdot \textsc{Cost}_i(y') + \tau + \epsilon/(4\rho) < \textsc{Cost}_i(x_a)] && \text{(since } \|x - x_a\| \le \epsilon/(4\rho)\text{)}\\
&= \frac{1}{|S|}\sum_{i\in S} f_{a,b}(i) && \text{(by the definition of } f_{a,b}(i)\text{)}\\
&\ge \frac{1}{n}\sum_{i=1}^{n} f_{a,b}(i) - \epsilon/2 && \text{(since } S \text{ is } (\epsilon/2)\text{-representative)}\\
&\ge \frac{|T|}{n} - \epsilon/2 && \text{(since } f_{a,b}(i) = 1 \text{ for all } i \in T\text{)}\\
&\ge \sum_{j=1}^{m} y_j/B + \eta + \epsilon - \epsilon/2 && \text{(by the core violation)}\\
&\ge \sum_{j=1}^{m} y'_j/B + \eta + \epsilon - \epsilon/2 - \epsilon/2 && \text{(since } \|y - y'\| \le \epsilon/(2B)\text{)}\\
&\ge \sum_{j=1}^{m} y'_j/B + \eta
\end{aligned}
$$

as needed. □

F Lower Bounds

In this section, we develop the necessary tools to establish our lower bounds on panel complexity.

F.1 Camouflaged Populations

We begin by defining camouflaged populations—populations that are nearly uniform with respect to a certain feature. As we show below, the difficulty of distinguishing camouflaged populations with a small panel makes them a powerful tool for lower bounding panel complexity.

Definition F.1 (Camouflaged Populations)**.** *Fix integers $h, w, r \ge 2$. For each vector $z \in \{-1, +1\}^h$, we define a population of size $n = 2hwr$ and a feature $f(z, h, w, r) : [n] \to \{1, \ldots, 2h\}$ as follows. We first choose any partition $(A_1, \ldots, A_{2h})$ of $[n]$ such that for each $j \in [h]$ it holds that $|A_{2j}| = r \cdot (w + z_j)$ and $|A_{2j-1}| = r \cdot (w - z_j)$. Then, we define $f_i(z, h, w, r) = j$ for all $i \in A_j$ and $j \in [2h]$.*

For a panel $S = \{i_1, \ldots, i_k\}$, we denote $f_S(z, h, w, r) = (f_{i_1}(z, h, w, r), \ldots, f_{i_k}(z, h, w, r))$. Note that a $(1 + z_j/w)/(2h)$ fraction of the population is assigned value $2j$, while a $(1 - z_j/w)/(2h)$ fraction is assigned value $2j - 1$, totaling a $1/h$ fraction with values in $\{2j - 1, 2j\}$ for each $j \in [h]$.

The following lemma shows that a panel of size k smaller than $\Omega(h \cdot w^2)$ drawn from a camouflaged population *cannot* recover $(3/4) \cdot h$ entries of z with probability $6/7$. The proof follows from a reduction to a standard sample complexity lower bound for sampling with replacement, as shown in [Duetting et al., 2025, Lemma 5.10].

Lemma F.2 (Panel Complexity Lower Bound)**.** *Fix integers $h, w \ge 2$. There exists an integer $r \ge 1$ such that, if there exists some function $g : \{1, 2, \ldots, 2h\}^k \to \{-1, +1\}^h$ for some $k \ge 1$ satisfying:*

$$
\mathbb{P}_{S\sim\mathcal{U}_{k,n}}\left[\|g(f_S(z, h, w, r)) - z\|_1 \le h/4\right] \ge 6/7 \quad \textit{for all } z \in \{-1, +1\}^h,
$$

then it must be that $k = \Omega(h \cdot w^2)$.

Proof. Let us define $\mathcal{R}_{k,n}$ as the probability distribution corresponding to sampling *with* replacement from the population. We use the following result of Duetting, Feldman, Ponitka, and Soumalias [2025, Lemma 5.10], which guarantees if it holds that:

$$\mathbb{P}_{(i_1,\ldots,i_k)\sim\mathcal{R}_{k,n}}\left[\|g(f_{i_1,\ldots,i_k}(z,h,w,r)) - z\|_1 \le h/4\right] \ge 3/4 \quad \text{for all } z \in \{-1,+1\}^h,$$

then it must be that $k = \Omega(h \cdot w^2)$. A key distinction between this statement and that of the lemma is how the panel is formed. In the lemma, the panel consists of k uniformly sampled agents without repetition, ensuring that set S contains distinct elements. In contrast, Duetting, Feldman, Ponitka, and Soumalias [2025] considers drawing k agents uniformly at random, allowing for possible repetitions.

We choose an integer parameter r large enough such that $k^2/(2n) = k^2/(4hwr) \le 1/8$.

Next, we define the event that no repetition occurs:

$$E = \{(i_1,\ldots,i_k) \in [n]^k \,:\, i_a \ne i_b \text{ for all } a,b \in [k]\}.$$

We bound the probability of repetition using the union bound:

$$\mathbb{P}_{(i_1,\ldots,i_k)\sim\mathcal{R}_{k,n}}[E] \ge 1 - \sum_{a\ne b} \mathbb{P}_{(i_1,\ldots,i_k)\sim\mathcal{R}_{k,n}}[i_a = i_b] = 1 - \binom{k}{2}\cdot\frac{1}{n} \ge 1 - \frac{k^2}{2n} \ge \frac{7}{8}.$$

Thus, for all $z \in \{-1,+1\}^\ell$, we have:

$$\begin{aligned}
&\mathbb{P}_{(i_1,\ldots,i_k)\sim\mathcal{R}_{k,n}}\left[\|g(f_{i_1,\ldots,i_k}(z,h,w,r)) - z\|_1 \le h/4\right] \\
&\qquad \ge \mathbb{P}_{(i_1,\ldots,i_k)\sim\mathcal{R}_{k,n}}\left[\|g(f_{i_1,\ldots,i_k}(z,h,w,r)) - z\|_1 \le h/4 \text{ and } E \text{ holds}\right] \\
&\qquad \ge \mathbb{P}_{(i_1,\ldots,i_k)\sim\mathcal{R}_{k,n}}\left[\|g(f_{i_1,\ldots,i_k}(z,h,w,r)) - z\|_1 \le h/4 \,|\, E\right] \cdot \mathbb{P}_{(i_1,\ldots,i_k)\sim\mathcal{R}_{k,n}}[E] \\
&\qquad \ge \mathbb{P}_{(i_1,\ldots,i_k)\sim\mathcal{R}_{k,n}}\left[\|g(f_{i_1,\ldots,i_k}(z,h,w,r)) - z\|_1 \le h/4 \,|\, E\right] \cdot (7/8) \\
&\qquad = \mathbb{P}_{S\sim\mathcal{U}_{k,n}}\left[\|g(f_S(z,h,w,r)) - z\|_1 \le h/4\right] \cdot (7/8) \\
&\qquad \ge (6/7)\cdot(7/8) \\
&\qquad \ge 3/4.
\end{aligned}$$

This concludes the proof. □

F.2 Facility Location Results

Theorem F.3 (Panel Complexity Lower Bound)**.** *For any $k > 0$, there exists a facility location instance $\langle \mathcal{X}, \mathcal{C}, d, (x_1,\ldots,x_n)\rangle$ such that for any panel decision function $\widetilde{q} : \mathcal{X}^k \to \mathcal{C}$, it holds that:*

$$\mathbb{P}_{S\sim\mathcal{U}_{k,n}}[d(\widetilde{q}(S), q^\star) \ge 2 \cdot \textsc{Social-Opt}] \ge 1/4.$$

Proof. Consider a population of $n = 2k+1$ agents located at $x_1,\ldots,x_k = 0$ and $x_{k+1},\ldots,x_{2k} = 1$. We leave the location of x_{2k+1} unspecified for now. Let $\mathcal{C} = \mathcal{X} = [0,1]$ and $d(x,y) = |x-y|$. Notice that the unique optimal location is $q^\star = x_{2k+1}$, regardless of x_{2k+1}. We compute:

$$\textsc{Opt} \le (T/2 + (1 - T/2))\cdot(1/2) = 1/2.$$

Notice that the probability that a panel $S \sim \mathcal{U}_{k,n}$ contains agent $2k+1$ is at most $1/2$. Conditioned on the event that the panel does not contain agent $2k+1$, we have that $\widetilde{x}(S)$ selects either 0 with probability at least $1/2$ or 1 with probability at least $1/2$. Assume without loss of generality that the latter case holds, and let $x_{2k+1} = 0$. Now, with probability at least $1/4$ over the draw of a panel $S \sim \mathcal{U}_{k,n}$, the panel decision function $\widetilde{q}$ selects the facility at 1. For each such panel S we have:

$$d(\widetilde{q}(S), q^\star) = d(1,0) = 1 \ge 2 \cdot \textsc{Opt},$$

since $\textsc{Opt} \le 1/2$ as shown above. This concludes the proof. □

Theorem F.4 (Panel Complexity Lower Bound)**.** *Fix $\mathcal{X} = [0,1]^t$ with $d = \ell_\infty$. Let $\mathcal{C} = [1/4, 3/4]^t$ and $0 < \epsilon < 1/64$. Suppose there exists a panel decision function $\widetilde{q} : \mathcal{X}^k \to \mathcal{C}$ for some $k \geq 1$ such that, for any facility location instance $\langle \mathcal{X}, \mathcal{C}, d, (x_1, \ldots, x_n)\rangle$, the following holds:*

$$\mathbb{E}_{S \sim \mathcal{U}_{k,n}} \left[\textsc{Social-Cost}(\widetilde{q}(S))\right] \leq (1+\epsilon) \cdot \textsc{Social-Opt}.$$

Then it must be that $k = \Omega((1/\epsilon)^2 \cdot t)$.

Proof. Our goal is to construct a function g that satisfies the assumption of Lemma F.2, establishing a necessary lower bound on k. Let $w = \lfloor (1/64) \cdot (1/\epsilon) \rfloor$. Note that $\epsilon \leq (1/64) \cdot (1/w)$ and that $1/w = \Omega(1/\epsilon)$. Let $r > 1$ be an integer parameter large enough to satisfy the conditions of Lemma F.2.

For every $z \in \{-1,+1\}^t$, consider a camouflaged population of size $n = 2twr$ with a feature $f(z,t,w,r)$ as in Definition F.1. We will now define a facility location instances $\mathcal{I}(z) = \langle \mathcal{X}, \mathcal{C}, d, (x_1^{(z)}, \ldots, x_n^{(z)})\rangle$ as follows. Let $c = (1/2, 1/2, \ldots, 1/2) \in \mathcal{X}$ and let $e_j = (0, \ldots, 0, 1/2, 0, \ldots, 0)$ for $j \in [t]$ where the $1/2$ entry is in the j-th position. For every agent $i \in [n]$, we set:

$$x_i^{(z)} = \begin{cases} c + e_j & \text{if } f_i(z,t,r,w) = 2j, \\ c - e_j & \text{if } f_i(z,t,r,w) = 2j-1. \end{cases}$$

Note that exactly a $(1 + z_j/w)/(2t)$-fraction of the population has location $c + e_j$, while exactly a $(1 - z_j/w)/(2t)$-fraction has location $c - e_j$. It follows that:

$$\textsc{Social-Cost}(q) = \sum_{j=1}^{t} \left(\frac{1 + z_j/w}{2t}\right) \cdot \|c + e_j - q\|_\infty + \left(\frac{1 - z_j/w}{2t}\right) \cdot \|c - e_j - q\|_\infty.$$

To obtain an upper bound on the optimal social cost, fix any $z \in \{-1,+1\}^t$ and consider the location:

$$q^\star = c + \sum_{j \in [t]} (1/2) \cdot z_j \cdot e_j \in \mathcal{C}.$$

We obtain:

$$\begin{aligned} \textsc{Social-Opt}(\mathcal{I}(z)) &\leq \textsc{Social-Cost}(q^\star) \\ &= \sum_{j=1}^{t} \left(\frac{1 + 1/w}{2t}\right) \cdot (1/4) + \left(\frac{1 - 1/w}{2t}\right) \cdot (3/4) \\ &= \frac{1}{2} - \frac{1/w}{4}. \end{aligned}$$

Now, fix an arbitrary candidate location $q \in \mathcal{C}$. We aim to derive a concrete lower bound on $\textsc{Social-Cost}(q)$. First, for any $j \in [t]$, since $1/4 \leq q_j \leq 3/4$, we have:

$$\left(\frac{1 + z_j/w}{2t}\right) \cdot \|c + e_j - q\|_\infty + \left(\frac{1 - z_j/w}{2t}\right) \cdot \|c - e_j - q\|_\infty \geq \frac{1}{2t} - \frac{1/w}{4t}.$$

Furthermore, if $q_j \leq 1/2$ and $z_j = +1$, then:

$$\begin{aligned} &\left(\frac{1 + z_j/w}{2t}\right) \cdot \|c + e_j - q\|_\infty + \left(\frac{1 - z_j/w}{2t}\right) \cdot \|c - e_j - q\|_\infty \\ &\geq \left(\frac{1 + 1/w}{2t}\right) \cdot (1 - q_j) + \left(\frac{1 - 1/w}{2t}\right) \cdot q_j \\ &= \frac{1}{2t} + \frac{(1/w) \cdot (1 - 2q_j)}{2t} \\ &\geq \frac{1}{2t}. \end{aligned}$$

Similarly, if $q_j \geq 1/2$ and $z_j = -1$, then:

$$\left(\frac{1 + z_j/w}{2t}\right) \cdot \|c + e_j - q\|_\infty + \left(\frac{1 - z_j/w}{2t}\right) \cdot \|c - e_j - q\|_\infty$$

$$
\begin{aligned}
&\geq \left(\frac{1+1/w}{2t}\right) \cdot q_j + \left(\frac{1-1/w}{2t}\right) \cdot (1-q_j) \\
&= \frac{1}{2t} + \frac{(1/w)\cdot(2q_j-1)}{2t} \\
&\geq \frac{1}{2t}.
\end{aligned}
$$

Define $g : \mathcal{X}^k \to \{-1,+1\}^t$ as follows:

$$
g_j(x_1,\ldots,x_k) = \begin{cases} +1 & \text{if } \widetilde{q}_j(x_1,\ldots,x_k) > 1/2, \\ -1 & \text{otherwise.} \end{cases}
$$

Note that g is independent of z. We use $g(f_S(z,t,w,r))$ for $S \subseteq [n]$ to denote $g((x_i^{(z)})_{i\in S})$.

Let $S \subseteq [n]$ be such that $\|g(f_S(z,t,w,r)) - z\|_1 > (1/4)\cdot t$. From the analysis above, we get:

$$
\begin{aligned}
\textsc{Social-Cost}(\widetilde{q}(S)) - \textsc{Social-Opt}(\mathcal{I}(z)) &\geq \left(\frac{t}{4}\right)\cdot\left(\frac{1/w}{4t}\right) = \frac{1}{16w} \\
&\geq \frac{1}{8w}\cdot \textsc{Social-Opt}(\mathcal{I}(z))
\end{aligned}
$$

Combining this with our assumption on $\overline{q}(S)$, and the fact that $\epsilon \leq (1/64)\cdot(1/w)$, we obtain:

$$
\begin{aligned}
&\frac{1}{64w}\cdot \textsc{Social-Opt}(\mathcal{I}(z)) \\
&\geq \epsilon \cdot \textsc{Social-Opt}(\mathcal{I}(z)) \\
&\geq \mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[\textsc{Social-Cost}(\widetilde{q}(S)) - \textsc{Social-Opt}(\mathcal{I}(z))\right] \\
&> \frac{1}{8w}\cdot \textsc{Social-Opt}(\mathcal{I}(z)) \cdot \mathbb{P}_{S\sim\mathcal{U}_{k,n}}\left[\|g(f_S(z,t,w,r)) - z\|_1 > (1/4)\cdot t\right]
\end{aligned}
$$

It follows that:

$$
\mathbb{P}_{S\sim\mathcal{U}_{k,n}}\left[\|g(f_S(z,t,w,r)) - z\|_1 \leq (1/4)\cdot t\right] \geq 7/8 \quad \text{for all } z \in \{-1,+1\}^t.
$$

Hence, g satisfies the assumption of Lemma F.2, which concludes the proof. □

Theorem F.5 (Impossibility for Multiplicative Error Guarantee)**.** *There exists a metric space $\mathcal{X}$, a set of alternatives $\mathcal{C}$, a distance function d, and a number of facilities ℓ such that for any $k < n$ and any panel decision function $\widetilde{q} : \mathcal{X}^k \to \mathcal{C}^\ell$, there exist locations $x_1,\ldots,x_n$ such that, for any $\rho > 1$:*

$$
\mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[\textsc{Social-Cost}(\widetilde{q}(S))\right] > \rho \cdot \textsc{Social-Opt}.
$$

Proof. Consider $\mathcal{X} = [0,1]$, $\mathcal{C} = \{0, 1/2, 1\}$, and $\ell = 2$. Fix any panel decision function $\widetilde{q} : \mathcal{X}^k \to \mathcal{C}^\ell$. Let $(q_1, q_2) = \widetilde{q}(0,0,\ldots,0)$. Without loss of generality, assume that $q_1 \neq 1$ and $q_2 \neq 1$.

Consider a population of agents located at $x_1 = \cdots = x_{n-1} = 0$ and $x_n = 1$. Note that $\textsc{Opt} = 0$ since placing one facility at 0 and another at 1 achieves zero cost. Since the panel decision function selects $(0, 1/2)$ with nonzero probability, the result follows. □

F.3 Participatory Budgeting Results

The following theorem gives a matching lower bound for the panel complexity of welfare maximization in participatory budgeting (Theorem E.2).

Theorem F.6 (Panel Complexity Lower Bound)**.** *Fix any number of projects $m \geq 2$, error $0 < \epsilon < 1/64$, and budget $B = \lfloor m/2 \rfloor$. Suppose there exists a panel decision function $\widetilde{x} : \mathcal{L}^k \to \mathcal{X}_B$ for some $k \geq 1$ such that, for any participatory budgeting instance $\langle m, B, (\textsc{Cost}_1,\ldots,\textsc{Cost}_n)\rangle$, the following holds:*

$$
\mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[\textsc{Social-Cost}(\widetilde{x}(S))\right] \leq \textsc{Social-Opt} + \epsilon.
$$

Then it must be that $k = \Omega(m\cdot(1/\epsilon)^2)$.

Proof. Our goal is to construct a function g that satisfies the assumption of Lemma F.2, establishing a necessary lower bound on k. Let $h = \lfloor m/2 \rfloor$ and $w = \lfloor (1/64) \cdot (1/\epsilon) \rfloor$. Note that $\epsilon \leq (1/64) \cdot (1/w)$ and that $1/w = \Omega(1/\epsilon)$. Let $r > 1$ be an integer parameter large enough to satisfy the conditions of Lemma F.2.

For each $z \in \{-1, +1\}^h$, consider the camouflaged population of size $n = 2hwr$ with a feature $f(z, h, w, r)$ as defined in Definition F.1. We will now define an instance $\mathcal{I}(z) = \langle m, B, \textsc{Cost}_1^{(z)}, \ldots, \textsc{Cost}_n^{(z)} \rangle$ for every vector $z \in \{-1, +1\}^h$ by letting $\textsc{Cost}_i^{(z)}(x) = 1 - x_j$ for each agent $i \in [n]$ such that $j = f_i(z, h, w, r)$. Note that exactly a $(1 + z_j/w)/(2h)$-fraction of the population have cost function $1 - x_{2j}$, while exactly a $(1 - z_j/w)/(2h)$-fraction have cost function $1 - x_{2j-1}$. Moreover, if m is odd, then none of those cost functions depend on the x_m, i.e., the m-th project has no value to anyone.

Next, we compare the expected social cost of any budget allocation with SOCIAL-OPT. For every budget allocation $x \in [0, 1]^m$ satisfying $\sum_{j=1}^{2\ell} x_j = B = h$, the social cost of x in instance $\mathcal{I}(z)$ is given by:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \textsc{Cost}_i^{(z)}(x) &= \sum_{j=1}^{h} \frac{(1 + z_j/w) \cdot (1 - x_{2j}) + (1 - z_j/w) \cdot (1 - x_{2j-1})}{2h} \\
&= \sum_{j=1}^{h} \frac{(2 - x_{2j} - x_{2j-1}) - (z_j/w) \cdot (x_{2j} - x_{2j-1})}{2h} \\
&= \sum_{j=1}^{h} \frac{2 - x_{2j} - x_{2j-1}}{2h} - \sum_{j=1}^{h} \frac{(z_j/w) \cdot (x_{2j} - x_{2j-1})}{2h} \\
&= \frac{1}{2} - \sum_{j=1}^{h} \frac{(z_j/w) \cdot (x_{2j} - x_{2j-1})}{2h}.
\end{aligned}$$

The optimal budget allocation for instance $\mathcal{I}(z)$ is to set $x_{2j-1} = 1$ and $x_{2j} = 0$ if $z_j = -1$, and $x_{2j-1} = 0$ and $x_{2j} = 1$ if $z_j = 1$. This allocation is budget-feasible since $B = h$. Thus,

$$\textsc{Social-Opt}(\mathcal{I}(z)) = \frac{1}{n} \sum_{i=1}^{n} \textsc{Cost}_i^{(z)}(x) = \frac{1}{2} - \sum_{j=1}^{\ell} \frac{1/w}{2h} = \frac{1}{2} - \frac{1}{2w},$$

for every $z \in \{-1, +1\}^h$. Henceforth, we will write $\textsc{Social-Opt} = \textsc{Social-Opt}(\mathcal{I}(z))$.

Since cost functions in the instances constructed above are monotone and they do not depend on x_m if m is odd, we may assume without loss of generality that $\sum_{j=1}^{2h} \widetilde{x}(c_1, \ldots, c_k) = B = h$ for every $c_1, \ldots, c_k \in \mathcal{L}$. By our assumption on $\widetilde{x} : \mathcal{L}^k \to \mathcal{X}_B$, for every $z \in \{-1, +1\}^h$, we have:

$$\begin{aligned}
\mathbb{E}_{S \sim \mathcal{U}_{k,n}} \left[\frac{1}{2} - \sum_{j=1}^{\ell} \frac{(z_j/w) \cdot (\widetilde{x}_{2j}(S) - \widetilde{x}_{2j-1}(S))}{2h} \right] &= \mathbb{E}_{S \sim \mathcal{U}_{k,n}} \left[\textsc{Social-Cost}(\widetilde{x}(S)) \right] \\
&\leq \textsc{Social-Opt} + \epsilon \\
&\leq \frac{1}{2} - \frac{1}{2w} + \frac{1}{64w} \\
&\leq \frac{1}{2} - \frac{31}{64w},
\end{aligned}$$

since $\epsilon \leq (1/64) \cdot (1/w)$. Multiplying both sides by $2w$ and rearranging terms gives:

$$\mathbb{E}_{S \sim \mathcal{U}_{k,n}} \left[\frac{1}{h} \sum_{j=1}^{h} z_j \cdot (\widetilde{x}_{2j}(S) - \widetilde{x}_{2j-1}(S)) \right] \geq \frac{31}{32}. \tag{7}$$

From this, we argue that:

$$\mathbb{P}_{S \sim \mathcal{U}_{k,n}} \left[\frac{1}{h} \sum_{j=1}^{h} \mathbb{1} \left[z_j \cdot (\widetilde{x}_{2j}(S) - \widetilde{x}_{2j-1}(S)) > 0 \right] \geq \frac{3}{4} \right] \geq \frac{7}{8} \quad \text{for all } z \in \{-1, +1\}^h, \tag{8}$$

since if this probability were smaller than $7/8$, then we would get:

$$\mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[\frac{1}{h}\sum_{j=1}^{h} z_j\cdot(\widetilde{x}_{2j}(S)-\widetilde{x}_{2j-1}(S))\right] < \frac{7}{8}\cdot 1+\frac{1}{8}\cdot\frac{3}{4}=\frac{31}{32},$$

which would contradict the previous bound in Statement (7).

Define $g:\mathcal{L}^k\to\{-1,+1\}^h$ as follows:

$$g_j(c)=\begin{cases}+1 & \text{if } \widetilde{x}_{2j}(c)-\widetilde{x}_{2j-1}(c)>0,\\ -1 & \text{otherwise.}\end{cases}$$

Note that g is independent of z. We use $g(f_S(z,h,w,r))$ for $S\subseteq[n]$ to denote $g((\textsc{Cost}_i^{(z)})_{i\in S})$.

Using Statement (8), we will argue that $g(f_S(z,h,w,r))$ and z agree on at least $(3/4)\cdot h$ indices with probability at least $7/8$, which can be expressed as:

$$\mathbb{P}_{S\sim\mathcal{U}_{k,n}}\left[\|g(f_S(z,h,w,r))-z\|_1\le 1/4\right]\ge 7/8 \quad \text{for all } z\in\{-1,+1\}^h.$$

To see why this holds, note that:

$$\begin{aligned}
& z_j\cdot(\widetilde{x}_{2j}(S)-\widetilde{x}_{2j-1}(S))>0\\
\implies\ & (z_j=+1 \text{ and } \widetilde{x}_{2j}(S)-\widetilde{x}_{2j-1}(S)>0) \text{ or } (z_j=-1 \text{ and } \widetilde{x}_{2j}(S)-\widetilde{x}_{2j-1}(S)\le 0)\\
\iff\ & (z_j=+1 \text{ and } g_j(f_S(z,h,w,r))=+1) \text{ or } (z_j=-1 \text{ and } g_j(f_S(z,h,w,r))=-1)\\
\iff\ & g_j(f_S(z,h,w,r))=z_j.
\end{aligned}$$

Hence, the function g satisfies the assumption of Lemma F.2, which concludes the proof. □

The next result shows that, even with an arbitrarily large panel, no panel decision function can guarantee a purely multiplicative approximation to the optimal allocation.

Theorem F.7 (Impossibility for Multiplicative Error Guarantee)**.** *Fix any number of projects $m\ge 2$, panel size $k\ge 1$, and budget $0<B<m$. Then, for any panel decision function $\widetilde{x}:\mathcal{L}^k\to\mathcal{X}_B$, there exists a participatory budgeting instance $\langle m,B,(\textsc{Cost}_1,\dots,\textsc{Cost}_n)\rangle$ such that, for any $\rho>1$, the following holds:*

$$\mathbb{E}_{S\sim\mathcal{U}_{k,n}}\left[\textsc{Social-Cost}(\widetilde{x}(S))\right] > \rho\cdot\textsc{Social-Opt}.$$

Proof. Let $B_1,B_2\ge 0$ satisfy $B_1+B_2=B$ and $B_1+B_2/m\le 1$. When $B\ge 1$, choosing $B_1=(m-B)/(m-1)$ and $B_2=(m\cdot B-m)/(m-1)$ ensures that $B_1+B_2=(m-B+m\cdot B-m)/(m-1)=B$ and $B_1+B_2/m=(m-B+B-1)/(m-1)=1$. When $B<1$, setting $B_1=B$ and $B_2=0$ satisfies the conditions. For any $x\in[0,1]^m$, define three cost functions as:

$$\begin{aligned}
\textsc{Cost}^{(0)}(x)&=(1/B_2)\cdot\sum_{i=1}^{m}\max(B_2/m-x_i,0),\\
\textsc{Cost}^{(1)}(x)&=B_1+B_2/m-x_1,\\
\textsc{Cost}^{(2)}(x)&=B_1+B_2/m-x_2.
\end{aligned}$$

These are valid cost functions in our model since $B_2\le m$ and $B_1+B_2/m\le 1$.

Consider the following two participatory budgeting instances: $\langle m,B,(\textsc{Cost}^{(0)},\dots,\textsc{Cost}^{(0)},\textsc{Cost}^{(1)})\rangle$ and $\langle m,B,(\textsc{Cost}^{(0)},\dots,\textsc{Cost}^{(0)},\textsc{Cost}^{(2)})\rangle$. These instances differ only in the cost function assigned to the n-th agent. In the first instance, an allocation x with $x_i=B_2/m$ for $i\ne 1$ and $x_1=B_1+B_2/m$ results in zero cost for all agents. This is a feasible budget allocation since $B_1+B_2=B$. Similarly, in the second instance, setting $x_i=B_2/m$ for $i\ne 2$ and $x_2=B_1+B_2/m$ also yields a cost of zero for all agents. Thus, the optimal cost in both instances is $\textsc{Social-Opt}=0$.

Now, let $y=\widetilde{x}(\textsc{Cost}^{(0)},\dots,\textsc{Cost}^{(0)})$. Since $\sum_{j=1}^{m}y_j\le B$ and $B_1+B_2=B$, at least one of the following must hold: $y_1<B_1+B_2/m$, $y_2<B_1+B_2/m$, or $y_j<B_2/m$ for some $j\ge 3$.

In the first case, we obtain

$$\mathbb{E}_{S\sim\mathcal{U}_{k,n}}[\textsc{Social-Cost}(\widetilde{x}(S))]\ge\mathbb{P}_{S\sim\mathcal{U}_{k,n}}[S\subseteq\{1,\dots,n-1\}]\cdot(B_1+B_2/m-y_1)>0.$$

The cases where $y_2<B_1+B_2/m$ or $y_j<B_2/m$ follow analogously. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the discussion of the practical limitations of uniform sampling in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide proofs and assumptions for all statements (sometimes deferred to the Appendix).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: The paper does not include experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the importance of the panel complexity problem in various practical social choice settings in the introduction. Since our main contributions are conceptual and theoretical, our discussion of societal impact is minimal.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.