PROPORTIONAL CLUSTERING: VERIFICATION AND OPTIMAL SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the problem of proportional clustering. When clustering is used to select representatives, keeping each point close to a center does not ensure that large groups receive a proportional share of the centers. We ask when this proportionality can be checked exactly and when it can be imposed while still optimizing the clustering objective exactly. We consider metric proportional justified representation (mPJR) and its stronger variant mPJR+. In arbitrary metrics, we reduce exact mPJR+ verification to maximum flow; the same test computes the largest number of missing centers and, through a finite search, the smallest dilation of representation distances under which mPJR+ holds. On the real line, we show that proportionality and clustering loss share an interval structure. As a result, a single minimum-cost flow finds an optimal weighted discrete k-median or k-means selection subject to mPJR, and also makes an existing selection proportional with the fewest replacements. In contrast, verifying mPJR is already coNP-complete in four-dimensional Euclidean space. With exact optimization, the cost of proportionality relative to the unconstrained optimum can still be arbitrarily large, even on the line. Experiments on synthetic populations and model confidence scores separate the cost of mPJR from the additional cost of mPJR+ and from the cost of retaining existing centers.

1 INTRODUCTION

What makes a small set of examples *representative* of a large dataset? This question arises when choosing images for annotation (Sener & Savarese, 2018) or inputs for human evaluation of text generation models (Zouhar et al., 2025). Suppose that half of the inputs to a model are similar requests from one family. An evaluation set reflecting this population may reasonably devote half its examples to that family. However, a clustering method, which keeps every input close to a selected example, may give the entire family just one example: once that example is nearby, another may do little to improve the clustering objective. *Proximity* and *proportional representation* thus capture different aspects of representativeness.

A small instance, adapted from the dense-group example of Jerrett & Anshelevich (2025), makes the distinction precise. There are six data points: four at 0, one at 10, and one at 11. We select three centers from $\{-0.1, 0, 0.1, 10, 11\}$. Selecting 0, 10, and 11 gives zero k-median and k-means cost. The four points at 0, however, make up two thirds of the population and receive just one nearby center. They could all reach three candidate centers within distance 0.1, so giving them two nearby representatives is possible. Doing so requires replacing one of the other selected centers and increases the clustering cost. Even a cost-optimal selection can therefore leave a large group with too few representatives.

Metric proportional justified representation (mPJR) formalizes this concern (Aziz et al., 2024; Kellerhals & Peters, 2024). We choose k centers from m candidates to represent n data points. A group containing at least $\ell n/k$ points is entitled to ℓ representatives if there are ℓ candidates within distance r of every group member. The requirement is that at least ℓ selected centers each lie within distance r of at least one group member. It applies to every group and every radius, without requiring group labels in advance. Its stronger variant, mPJR+, makes the same demand whenever the group has even one common unselected candidate within that radius (He et al., 2026). Both conditions can always be met, including when points have positive weights.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existence leaves two questions unanswered. Given a selection produced by an existing method, can we check whether it is proportional? And if we are free to choose the centers, can we find the proportional selection with the smallest clustering loss? If changing an existing selection is costly, we may instead want to replace as few centers as possible. In this work, we ask the following question:

When can we check proportionality exactly, and when can we optimize under it?

He et al. (2026) show that checking mPJR is coNP-complete in general metrics. For mPJR+, they give a polynomial-time test using $O(mn)$ submodular minimizations, as well as an exact test exponential in k and a weaker condition with faster verification. For approval PJR+, Springham (2026) replaces submodular minimization with minimum cuts. Existing mPJR constructions provide proportional selections, but do not determine which such selection minimizes a given clustering objective. We address these computational questions through three main results.

1.1 Our Results

Exact verification by maximum flow. In any metric, we extend the minimum-cut approach for approval PJR+ (Springham, 2026) to weighted mPJR+ at all radii. At most $(m-k)n$ maximum-flow computations suffice. The idea is to compare a group's population share with the number of selected centers it can reach; a minimum cut finds the group with the greatest shortfall. Besides checking proportionality, the algorithm computes the largest number of missing representatives and the smallest distance dilation needed to meet mPJR+, or determines that no finite dilation suffices. These quantities measure how many centers a group lacks and how much farther it must reach.

Optimal proportional clustering on the line. When points and candidates lie on the real line, one minimum-cost flow finds a globally optimal mPJR selection for weighted discrete k-median, weighted discrete k-means, and more general nondecreasing nearest-center losses. The same approach finds a proportional selection that replaces as few existing centers as possible. The key observation is that every proportionality requirement becomes a lower bound on the number of centers selected from an interval. Clustering loss also has an interval formulation, allowing us to optimize it exactly while meeting all these requirements. The network has $m+1$ vertices and $O(nm+nk+m)$ arcs. A separate test checks both mPJR and mPJR+ in $O(n \log n + nk)$ time when $m = \Theta(n)$. The line setting covers selection based on a single numerical attribute, such as a model's confidence score; representation is then measured by differences in scores.

Hardness in four dimensions. Euclidean geometry alone does not make these tasks tractable. We prove that checking mPJR is coNP-complete already in $\mathbb{R}^4$, and in $\mathbb{R}^5$ even when all points lie on one common sphere. Both results hold with integer coordinates. The main difficulty is that a violation may occur at any radius, so the reduction must control all radii at once. As a consequence, finding a proportional selection with the fewest replacements is also computationally hard in these dimensions: deciding whether zero replacements suffice already requires checking mPJR.

The cost of proportionality can be unbounded even under exact optimization, with six points on the line and a positive unconstrained optimum. Experiments on synthetic populations and model confidence scores compare the Spatial Expanding Approval Rule (SEAR) (Aziz et al., 2024), which satisfies mPJR+, with the minimum-loss mPJR+ selection, and measure the cost of keeping centers, running time, and memory.

Table 1 summarizes the results and the condition each one concerns: the general-metric algorithm verifies mPJR+, whereas the line optimizer imposes mPJR. Hardness for mPJR+ concerns a specified level ℓ; the flow algorithm instead checks whether any level is violated.

1.2 Related Work

Proportional clustering draws on approval-based multiwinner voting, especially proportional justified representation (Sánchez-Fernández et al., 2026; Aziz et al., 2018). Brill & Peters (2023) introduced PJR+ under this name and showed that it can be verified in polynomial time, building on the inclusion proportionality condition of Aziz & Lee (2021); Dong & Peters (2026) study its axiomatic justification. Aziz et al. (2024) give proportional clustering constructions, including SEAR, which selects centers as growing balls contain enough remaining population weight. Kellerhals & Peters

Table 1: Exact guarantees. The positive results allow weights. In the last row, the representation level is specified in the input.

Task	**Domain and Condition**	**Result**
Verification	Any metric; mPJR+	Maximum flow; exact missing-center count and minimum dilation
Verification	Line; mPJR and mPJR+	$O(n \log n + nk)$ time when $m = \Theta(n)$
Optimization and fewest replacements	Line; mPJR	One minimum-cost flow; exact proportionality and objective optimization
Verification	$\mathbb{R}^4$; mPJR	coNP-complete; also on a sphere in $\mathbb{R}^5$
Verification at a specified level	$\mathbb{R}^4$; mPJR+	coNP-complete, with one unselected candidate

(2024) develop the connection between clustering and social choice. Jerrett & Anshelevich (2025) combine proportionality with approximation guarantees for the sum of distances to *all* representatives in general metrics; our exact line result includes the usual nearest-center objectives.

Chen et al. (2019) introduced a different proportional clustering condition: no group of at least n/k agents should unanimously prefer a new center to its members' nearest selected centers. They study verification, approximate objective optimization, and the cost of imposing this condition. Li et al. (2021) instead consider deviations that reduce a group's total distance. These conditions concern improvements in nearest-center service, whereas mPJR requires a proportional number of representatives for cohesive groups. Our exact optimization and cost results apply to this latter requirement.

Interval preferences make several voting problems tractable (Elkind & Lackner, 2015; Springham et al., 2026; Becker & Frank, 2026). In particular, Peters (2018) uses total unimodularity to optimize committee objectives under single-peaked preferences, and Elkind et al. (2024) optimize approval welfare subject to justified representation on a one-dimensional domain. Our loss formulation also uses the classical distance-threshold representation of facility-location cost (Cornuéjols et al., 1980). The additional step is to express all mPJR requirements, across every radius and level, as interval constraints that preserve integrality. This gives exact optimization of the clustering objective and minimum-change repair through one minimum-cost flow.

For verification, Springham (2026) applies the classical maximum-closure reduction (Picard, 1976) to approval PJR+. Our metric formulation adds weights, checks all relevant radii, and computes missing-center counts and minimum distance dilation. Suzuki & Vollen (2024) use minimum-cost flow to optimize approval welfare under a proportionality condition for randomized committees. Our optimizer returns a deterministic selection satisfying mPJR at every radius. Finally, He et al. (2026) establish general metric hardness for mPJR and specified-level mPJR+; our hardness result keeps the Euclidean dimension fixed, with integer coordinates and a common-sphere variant. In contrast, realizing arbitrary approval profiles on a sphere with two agent–candidate distances can require dimension at least $\min\{n, m\} - 1$ (Appendix E).

2 PRELIMINARIES

An *instance* consists of a nonempty multiset N of n *agents* and a set M of m *candidate centers* in a metric space with distance d, together with an integer $1 \le k \le m$; distinct agents may occupy the same location. We choose a *selection* $X \subseteq M$ of exactly k centers. Agent i has positive weight w_i; let $w(S) = \sum_{i \in S} w_i$ and $W = w(N)$. The *quota* $q = W/k$ is the population weight corresponding to one center. The unweighted case has $w_i = 1$ for every agent.

For a point u and radius $r \ge 0$, let $B(u, r) = \{v : d(u, v) \le r\}$ be the closed ball. At radius r, a group $S \subseteq N$ *reaches* the candidates $A_r(S) = \bigcup_{i \in S}(B(i, r) \cap M)$, that is, those within distance r of at least one member. For $\ell \in \{1, \dots, k\}$, the group is *eligible at level* ℓ if its weight is at least ℓ

quotas, $w(S) \geq \ell q$. It is *ℓ-cohesive at radius r* if at least ℓ candidates are within distance r of every member: $\left|\bigcap_{i \in S}(B(i,r) \cap M)\right| \geq \ell$.

Definition 2.1 (mPJR)**.** A selection satisfies *metric proportional justified representation (mPJR)* if, for every $r \geq 0$, every $\ell \in \{1, \ldots, k\}$, and every group S that is eligible at level ℓ and ℓ-cohesive at radius r, $|X \cap A_r(S)| \geq \ell$.

The conclusion is collective: each of these ℓ centers need only be near some group member. In the introductory example, the four agents at 0 are eligible at level two and 2-cohesive at radius 0.1, but reach only one selected center at this radius. Under mPJR, a group can claim ℓ representatives only if its members share ℓ nearby candidates. The stronger condition mPJR+ asks whether a group eligible at level ℓ that receives fewer than ℓ representatives can point to even one common candidate that was left out. Such a group is underrepresented, regardless of how many other nearby candidates its members share.

For an unselected candidate $c \in M \setminus X$ and a nonempty group S, define $r(c, S) = \max_{i \in S} d(i, c)$. This is the smallest radius at which every member can reach c.

Definition 2.2 (mPJR+ and distance dilation)**.** For $\gamma \geq 1$, a selection satisfies γ-mPJR+ if, for every $\ell \in \{1, \ldots, k\}$, every unselected candidate c, and every group S eligible at level ℓ,

$$|X \cap A_{\gamma r(c,S)}(S)| \geq \ell. \tag{2.1}$$

The case $\gamma = 1$ is mPJR+ (He et al., 2026).

Increasing γ permits longer representation distances; it does not reduce the group's entitlement. At $\gamma = 1$, the condition implies mPJR: if a group S that is eligible at level ℓ and ℓ-cohesive at radius r reaches fewer than ℓ selected centers, then one of its common candidates c is unselected, and S reaches no more selected centers at the smaller radius $r(c, S) \leq r$. Allowing $\gamma > 1$ weakens the distance requirement and need not imply mPJR.

Unless a single input level is explicitly specified, verification asks whether the condition holds at every level. We use $\mathbf{1}[E]$ for the indicator of a statement E, let $[z]_+ = \max\{z, 0\}$, and take a maximum over an empty set to be zero. Both conditions are attainable with arbitrary positive weights; Appendix C.1 gives a direct proof. The algorithms below operate on weights without replacing agents by copies. Running times count exact arithmetic operations and comparisons; for rational input data, the algorithms also have polynomial bit complexity.

3 Verification in Arbitrary Metrics

The apparent challenge is the number of groups and representation levels to check. Following the verification reformulation of Brill & Peters (2023) and He et al. (2026), these checks can be combined: a group is underrepresented at some level exactly when its weight exceeds the weight corresponding to its reached centers by at least one quota. Multiplying by k gives the quantity $kw(S) - W|X \cap A_r(S)|$, which must be at least W for a violation. This is a maximum-weight closure objective (Picard, 1976): choosing an agent requires counting every selected center it reaches. Springham (2026) uses this structure to verify approval PJR+ by minimum cuts. We give the weighted metric formulation and use it to measure missing centers and the distance dilation needed for representation.

Fix $\gamma \geq 1$, an unselected candidate c, and a radius r. Let $N(c, r) = N \cap B(c, r)$ be the agents who can all reach c at that radius. We compare their population weight with the selected centers they can collectively reach within γr. Build a network with a source s, a sink t, one vertex for each agent in $N(c, r)$, and one for each center in X. The arcs and capacities are

$$s \longrightarrow i : \ kw_i, \quad x \longrightarrow t : \ W, \quad i \longrightarrow x : \ kW + 1 \quad \text{if } d(i, x) \leq \gamma r.$$

Let $\nu_\gamma(c, r)$ be its maximum-flow value. The large capacity on an agent-to-center arc ensures that a minimum cut keeping an agent on the source side must also keep every selected center that the agent reaches. Consequently,

$$kw(N(c, r)) - \nu_\gamma(c, r) = \max_{S \subseteq N(c,r)} (kw(S) - W|X \cap A_{\gamma r}(S)|). \tag{3.1}$$

Indeed, the cut associated with S has capacity $k(w(N(c, r)) - w(S)) + W|X \cap A_{\gamma r}(S)|$. No minimum cut crosses an arc of capacity $kW + 1$, since the cut consisting only of s has capacity at most kW.

Only radii in $R(c) = \{d(i, c) : i \in N\}$ are needed: a nonempty group's required radius $r(c, S)$ is one of these distances. Define

$$D_\gamma(X) = \max\left\{0, \max_{c \in M \setminus X} \max_{r \in R(c)} (kw(N(c, r)) - \nu_\gamma(c, r))\right\}. \tag{3.2}$$

A group may be missing more than one center, and the cut values retain this information. Dividing $D_\gamma(X)$ by W expresses the largest shortfall in units of centers; taking the integer part accounts for integer representation levels.

Theorem 3.1. *In any metric, X satisfies weighted γ-mPJR+ if and only if $D_\gamma(X) < W$. Moreover,*

$$\max_{c \in M \setminus X} \max_{\varnothing \neq S \subseteq N} \left[\left\lfloor \frac{w(S)}{q} \right\rfloor - |X \cap A_{\gamma r(c,S)}(S)|\right]_+ = \left\lfloor \frac{D_\gamma(X)}{W} \right\rfloor. \tag{3.3}$$

At most $(m - k)n$ maximum-flow computations suffice, each on at most $n + k + 2$ vertices and $nk + n + k$ arcs. If there is a violation, a minimum cut returns a group attaining the largest number of missing centers.

Appendix A.1 proves that maximizing over the stated radii is equivalent to using each group's own radius $r(c, S)$. The algorithm is strongly polynomial with a strongly polynomial maximum-flow method, such as that of Orlin (2013).

As the radius increases, the existing flow can be continued instead of rebuilt. Appendix A.3 gives the implementation and its unweighted running time. The same calculation checks approval PJR+ with one flow per unselected candidate (Appendix A.4).

We can also measure how far the representation distances must increase. Define the minimum required dilation by $\Gamma^+(X) = \inf\{\gamma \geq 1 : X \text{ satisfies } \gamma\text{-mPJR+}\}$. Unlike the group-size threshold of Springham (2026), this keeps eligibility fixed and increases representation distances. For a positive checked radius $r = d(j, c)$, the network changes with γ only at the ratios $d(i, x)/d(j, c)$. When finite, $\Gamma^+(X)$ is 1 or one of at most $(m - k)n^2k$ ratios, found by sorting and binary search with Theorem 3.1. Only a violation at radius zero makes $\Gamma^+(X)$ infinite. Appendix A.5 gives the ratio set and proof. On the line, we avoid this search.

4 Optimal Proportional Clustering on the Line

Verification tells us whether a given selection is proportional, but leaves open which proportional selection has the smallest clustering loss. To optimize, we need constraints describing all admissible selections. On the line, the order of agents and candidates supplies exactly this description for mPJR: every requirement becomes a lower bound on the number of centers selected from an interval. We first derive these constraints and then show why they are compatible with the clustering objective.

Sort the agent locations as $p_1 \leq \cdots \leq p_n$ and the candidates as $M = \{c_1 < \cdots < c_m\}$. For each level ℓ and each left index a, let $b_\ell(a)$ be the smallest $b \geq a$ such that $\sum_{i=a}^{b} w_i \geq \ell q$, when such an index exists. Call the consecutive agents with indices $I = [a, b_\ell(a)]$ a *level-ℓ eligible block*. These are the shortest eligible blocks with each given left endpoint, so there are at most n per level. In the unweighted case, each contains $\lceil \ell n/k \rceil$ agents.

4.1 From all groups and radii to interval requirements

We first reduce the groups that must be considered. Suppose a group violates mPJR at radius r and level ℓ, and its extreme agents have indices a and b. Every candidate within distance r of both extremes is within that distance of every agent between them. We may therefore fill in the intermediate agents without losing cohesion. Since the original group has a common candidate, its radius-r intervals overlap and their union is exactly $[p_a - r, p_b + r]$. Filling in the agents introduces no newly reached candidate. Finally, keep only the block $[a, b_\ell(a)]$. It still has enough weight, remains cohesive, and cannot reach more selected centers. Thus some eligible block also violates mPJR. The same argument applies to mPJR+ and its dilations; Appendix B gives the details.

It remains to reduce the radii: for each block, we need only the first radius at which cohesion holds. For a block $I = [a, b]$, write its midpoint and half-width as $\mu_I = (p_a + p_b)/2$ and $h_I = (p_b - p_a)/2$.

For a finite set $Z \subset \mathbb{R}$, let $\mathrm{near}_j(\mu; Z)$ be the jth smallest distance from μ to a point of Z, with value $+\infty$ if $|Z| < j$. A candidate c is within distance r of every member of I exactly when $h_I + |c - \mu_I| \le r$. Hence, $\alpha_\ell(I) = h_I + \mathrm{near}_\ell(\mu_I; M)$ is the first radius at which I is ℓ-cohesive. At this radius, the block reaches precisely the candidate interval $J_\ell(I) = M \cap [p_a - \alpha_\ell(I),\ p_b + \alpha_\ell(I)]$. Testing the first cohesion radius suffices, since a larger radius can only add selected centers. We obtain the exact equivalence

$$X \text{ satisfies mPJR} \quad \Longleftrightarrow \quad |X \cap J_\ell(I)| \ge \ell \quad \text{for every level-}\ell \text{ eligible block } I. \tag{4.1}$$

Every interval $J_\ell(I)$ is determined before X is chosen. Thus at most nk lower bounds describe all mPJR selections and can be imposed directly during optimization. Weights decide which blocks are eligible; the required center counts remain integers. For mPJR+, the smallest radius at which a block has a common unselected candidate depends on X, so its verification test gives no such description.

Before optimizing, we can use these blocks to check a given selection more efficiently. Cohesion and representation are determined by nearest-neighbor distances from μ_I, and the ordered block midpoints let us process these queries together. This also gives the minimum mPJR+ dilation directly, avoiding the ratio search needed in a general metric.

Theorem 4.1. *On the real line, weighted mPJR and weighted γ-mPJR+ can each be verified, and the exact value $\Gamma^+(X)$ computed, in $O((n+m)\log(n+m) + k\min\{n\log(m+1),\ n+m\})$ time and $O(n+m)$ space. In particular, the time is $O(n\log n + nk)$ when $m = \Theta(n)$.*

Appendix B gives the necessary and sufficient inequalities, the closed formula for $\Gamma^+(X)$, and the algorithm. This test is useful when one only needs to check a selection; the optimizer below solves a larger problem.

4.2 Minimizing the clustering objective without relaxing fairness

The interval requirements specify which selections are allowed; we now ask which allowed selection has the smallest loss. Let $d(i, X) = \min_{x \in X} |p_i - x|$. Agent i incurs a nondecreasing loss $\phi_i(d(i, X))$, and selecting candidate c_j adds an opening cost a_j. The loss functions are specified at the finitely many agent–candidate distances. The total cost of a selection is $F(X) = \sum_i w_i \phi_i(d(i, X)) + \sum_{c_j \in X} a_j$. Taking $\phi_i(d) = d$ gives weighted discrete k-median; taking $\phi_i(d) = d^2$ gives weighted discrete k-means. In both cases, centers must belong to M.

Theorem 4.2. *On the real line, a k-element selection that satisfies weighted mPJR and minimizes F can be computed in polynomial time. The problem reduces to one minimum-cost flow on $m+1$ vertices and $O(nm + nk + m)$ arcs. The same result holds when each agent's loss is summed over any specified number of its closest selected centers.*

The extension concerns the *sum* of losses to several nearby centers (Appendix C.2). The interval formulation also minimizes the maximum nearest-center distance by a finite search (Corollary C.2).

We use the distance-threshold formulation of facility-location cost (Cornuéjols et al., 1980). On the line, the candidates within a given distance of an agent form an interval, just like the candidate sets in the mPJR requirements. For agent i, list its distinct distances to candidates as $\rho_{i1} < \cdots < \rho_{is_i}$. For $t < s_i$, define $H_{it} = M \cap [p_i - \rho_{it}, p_i + \rho_{it}]$ and $\Delta_{it} = w_i(\phi_i(\rho_{i,t+1}) - \phi_i(\rho_{it})) \ge 0$. The nearest selected center is farther than ρ_{it} exactly when X contains no candidate in H_{it}. Summing the successive loss increments therefore gives us

$$w_i \phi_i(d(i, X)) = w_i \phi_i(\rho_{i1}) + \sum_{t < s_i} \Delta_{it} \left[1 - |X \cap H_{it}|\right]_+ . \tag{4.2}$$

Thus the objective charges for empty intervals, while fairness requires enough selected centers in other intervals (Figure 1). Both families are intervals in the same candidate order.

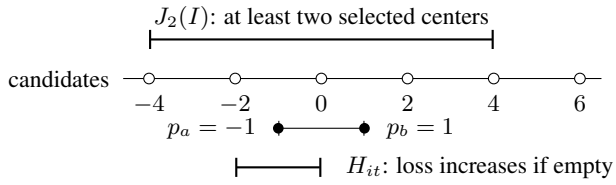

Figure 1: For an eligible level-two block with extreme agents at -1 and 1, the first cohesion radius is 3, giving $J_2(I) = M \cap [-4, 4]$. A distance threshold of 1 for the left agent gives the loss interval $H_{it} = M \cap [-2, 0]$.

Let x_j indicate whether c_j is selected, and let the shortage variable z_{it} represent $[1 - |X \cap H_{it}|]_+$. Omitting the constant in (4.2), solve

$$\begin{aligned} \min \quad & \sum_j a_j x_j + \sum_{i,t} \Delta_{it} z_{it} && (4.3) \\ \text{subject to} \quad & \sum_j x_j = k, \quad 0 \le x_j \le 1, \\ & \sum_{j : c_j \in J_\ell(I)} x_j \ge \ell && \text{for every } (\ell, I), \\ & \sum_{j : c_j \in H_{it}} x_j + z_{it} \ge 1, \quad 0 \le z_{it} \le 1 && \text{for every } (i, t). \end{aligned}$$

As in Peters (2018), the interval structure makes this linear program integral. In every constraint, the nonzero coefficients on x occupy consecutive indices, and each shortage variable occurs in only one constraint besides its bounds. The resulting matrix is *totally unimodular*: every square submatrix has determinant 0, 1, or -1 (Fulkerson & Gross, 1965). Integer right-hand sides then give an integral optimal solution (Hoffman & Kruskal, 1956), hence an optimal mPJR selection. Appendix C.3 gives the proof and establishes feasibility.

For the flow formulation, let $y_j = \sum_{h=1}^{j} x_h$, with $y_0 = 0$. The number of selected centers in $\{c_a, \dots, c_b\}$ becomes $y_b - y_{a-1}$. By linear-programming duality, the problem becomes a minimum-cost flow with one vertex for each index $0, \dots, m$ (Ahuja et al., 1993). Fairness intervals give uncapacitated arcs; loss intervals give arcs with capacities Δ_{it}. Appendix C.4 gives the network and recovers an optimal integral selection from shortest-path distances derived from the optimal flow.

4.3 Changing as few centers as possible

Suppose an existing method has returned X^0 and replacing centers is costly. Set all losses to zero and take $a_j = \mathbf{1}[c_j \notin X^0]$. Then $F(X) = |X \setminus X^0|$, exactly the number of replacements because both selections contain k centers.

Corollary 4.3. *On the line, a weighted mPJR selection requiring the fewest replacements of a given k-element selection is computable by one minimum-cost flow on $m + 1$ vertices and $O(nk + m)$ arcs.*

The formulation can prioritize replacements first and clustering loss second by assigning a replacement cost larger than the full loss range (Appendix C.5). In the introductory example, one replacement suffices, with minimum proportional k-median and k-means cost 1.

How much does proportionality cost? Unbounded increases are known for other clustering fairness conditions (Chen et al., 2019; Esmaeili et al., 2021; Jerrett & Anshelevich, 2025). We give an exact example on the line in which the unconstrained optimum, which also chooses k centers from M, has positive cost. Let $C_p(X) = \sum_i d(i, X)^p$.

Proposition 4.4. *For every fixed $p > 0$, the ratio of the minimum mPJR cost to the positive unconstrained minimum of C_p is unbounded on the line. The same holds for mPJR+, with the two proportional optima equal. Six unit-weight agents, five integer candidate locations, and $k = 3$ suffice.*

Four colocated agents require two nearby centers, leaving only one center for two distant agents. Moving the distant agents farther apart increases the proportional cost while leaving the unconstrained cost unchanged. Appendix C.7 gives the optima exactly.

5 Verification is Hard in Four Dimensions

The interval structure of the line gives both fast verification and exact optimization. Is a fixed number of Euclidean coordinates enough for similar guarantees? The next theorem rules this out already in four dimensions, strengthening the general metric hardness results of He et al. (2026).

Theorem 5.1. *Verifying mPJR is coNP-complete for unweighted Euclidean instances in $\mathbb{R}^4$ with integer coordinates. It remains coNP-complete in $\mathbb{R}^5$ when all agents and candidates lie on one common sphere. Both statements also hold for verifying mPJR+ at a specified input level ℓ, even with only one unselected candidate.*

The last statement does not contradict Theorem 3.1. The flow test asks whether *some* level is violated, which is equivalent to the single inequality $kw(S) - W|X \cap A_r(S)| \geq W$. At a specified level ℓ, which is part of the input, a violation requires both $w(S) \geq \ell q$ and $|X \cap A_r(S)| < \ell$. A group maximizing the difference can violate another level while satisfying the specified one, so the flow test does not answer this question.

At a fixed radius, each agent can be viewed as approving the candidates within that radius. A hard approval profile at one radius is not enough for our theorem: other radii could create unrelated violations. Our reduction ensures that every violation, at any radius, yields a solution to the same graph problem. Appendix D gives the details; the main idea is as follows.

We reduce from Balanced Biclique, also used by He et al. (2026) for approval PJR+ at a specified level: does a bipartite graph contain a complete bipartite subgraph with t vertices on each side (Garey & Johnson, 1979, Problem GT24)? Each vertex becomes a selected candidate, and each edge gives colocated agents near its endpoints. Population sizes ensure that these agents can demand more centers than the endpoints of their edges provide exactly when their edges contain a balanced biclique. Additional candidates and agents exclude unrelated violations at other radii.

The geometric task is to make each edge agent closer to its endpoints than to any other vertex. We adapt the trigonometric moment curve construction (Sanyal et al., 2011) in $\mathbb{R}^4$, with distance gaps that accommodate the additional candidates. Appendix D gives the coordinates and shows that rational approximation and projection onto a sphere preserve the required inequalities.

An exact minimum-change repair algorithm would solve mPJR verification: the minimum number of replacements is zero exactly when the original selection is proportional. Therefore Corollary 4.3 cannot extend to $\mathbb{R}^4$, or to the common-sphere instances in $\mathbb{R}^5$, unless $\mathrm{P} = \mathrm{NP}$.

6 The Cost of Proportionality and the Value of Keeping Centers

We examine the cost of proportionality, the additional cost of mPJR+, and the effect of keeping existing centers. For each loss $C \in \{C_1, C_2\}$, let X^0 be an unconstrained optimum, X^* a minimum-loss mPJR selection, X^{R} a minimum-change mPJR selection, and X^{S} the SEAR selection. Write $C^a = C(X^a)$ and $r^a = |X^a \setminus X^0|$. We break ties in X^* by fewer replacements and in X^{R} by lower loss. Let C^+ be the minimum mPJR+ loss. Comparing C^{S} with C^+, rather than only C^*, tests SEAR against the optimum under its own guarantee.

Populations and confidence scores. The synthetic study uses $n = 80$, $m = 60$, and $k = 6$, with 25 trials per model. Uniform instances sample locations from $[0, 10000]$. Concentrated instances place 64 agents and 20 candidates in $[0, 400]$, and 16 agents and 40 candidates in $[4000, 10000]$. A further 25 paired trials keep candidates fixed while varying the share of agents in $[0, 400]$ from 20% to 80%.

For a one-dimensional selection task, we train logistic regression on the handwritten-digits data (Alpaydin & Kaynak, 1998; Pedregosa et al., 2011) and select $k = 10$ of 540 held-out images, using the maximum predicted class probability as the coordinate. Ten stratified train/test splits give 244–276 distinct candidate scores after rounding to four decimal places. Every distinct score is a

Table 2: Median loss ratios and replacement counts. There are 25 trials per synthetic row and 10 per confidence score row. C^+/C^* measures the extra cost of mPJR+; C^{S}/C^+ compares SEAR with an optimum satisfying the same condition.

Population	Loss	C^*/C^0	C^+/C^*	C^{S}/C^+	C^{R}/C^*	r^{R}, r^*
Uniform	k-median	1.00	1.000	1.45	1.00	0, 0
Uniform	k-means	1.00	1.000	2.37	1.00	0, 0
Concentrated	k-median	1.65	1.000	1.59	1.04	2, 5
Concentrated	k-means	5.83	1.000	2.94	1.18	3, 5
Confidence scores	k-median	1.90	1.006	1.40	1.08	5, 10
Confidence scores	k-means	8.09	1.010	2.08	1.20	5, 9.5

candidate; no separate candidate distribution is imposed. These experiments concern representation of confidence scores, not image diversity or improved prediction accuracy. All populations have unit weights; Appendix F gives the full methods and results.

Proportionality and the choice of centers. On uniform instances, every unconstrained optimum satisfies mPJR; on concentrated instances and confidence scores, none does. The median mPJR cost ratios are 1.65 and 5.83 on concentrated data, and 1.90 and 8.09 on confidence scores, for k-median and k-means respectively (Table 2). The paired study shows the same concentration effect with candidates held fixed (Table 5): a large group in a short interval can require several representatives even when one center already keeps its nearest-center loss small.

The stronger condition contributes much less to these costs. The mPJR and mPJR+ optima agree in 299 of the 300 synthetic objective comparisons; in the remaining one, the increase is 0.55%. On confidence scores, all 20 comparisons have a positive increase, but the medians are only 0.57% and 1.02%. SEAR still costs 1.40 and 2.08 times the mPJR+ optimum. Thus, on these samples, its larger loss is mainly due to how it chooses centers, not to its stronger proportionality requirement. We compute C^+ by an integer formulation checked with the line verifier; this comparison does not give a polynomial-time mPJR+ optimizer.

Keeping centers and computational cost. Minimizing replacements first increases the loss by a median of 4.4% and 17.8% on concentrated data, while reducing replacements from five to two and from five to three. The corresponding confidence score increases are 7.5% and 20.4%. These are distinct costs: proportionality changes which selections are allowed, whereas retaining centers changes which allowed selection is preferred.

With the same HiGHS solver, the flow formulation is also faster than the direct interval LP. At $(n, m, k) = (640, 480, 6)$, the median total times are 1.47 and 27.94 seconds, respectively; at $(1280, 960, 6)$, the flow takes 7.14 seconds, while all three LP runs reach the 30-second solver limit. Pure minimum-change repair takes 0.063 seconds at the largest size. Appendix F.5 reports all timings, memory, tests with varying k, and general-metric verification comparisons. Independent exhaustive tests check the implementations against the definitions on small weighted instances.

7 Conclusion

Maximum flow makes mPJR+ verifiable in any metric and measures underrepresentation by missing centers or by the required dilation. On the line, mPJR and nearest-center loss share an interval formulation, giving exact optimization and minimum-change repair. The four-dimensional hardness result shows that Euclidean geometry alone is not enough. The cost of either proportionality condition can be unbounded, even when their optima agree. Our experiments separate this cost from SEAR's choice of proportional centers and from the cost of retaining existing centers; they do not give a universal cost guarantee or establish how often underrepresentation occurs in applications. The confidence score study uses one dataset and overlapping splits. Efficient mPJR verification in dimensions two and three, and polynomial-time mPJR+ optimization on the line, remain open.

AI USE STATEMENT

We used generative AI tools to assist with writing and revision, literature retrieval and discovery, research ideation and execution, and proving mathematical claims. Specifically, AI assistance supported the development, refinement, and review of mathematical arguments and proofs; improvements to the clarity and organization of the exposition; searches for potentially relevant literature; and the implementation, execution, and interpretation of experiments.

The authors developed the mathematical framework and exercised final judgment over the research direction, the arguments incorporated into the paper, and the presentation of the contributions. We critically evaluated AI-assisted suggestions and reviewed, corrected, and revised the material incorporated into the submission. Our review included checking mathematical statements and proof steps against the definitions and assumptions, examining whether the arguments establish the stated conclusions, and checking consistency between the main text and appendices. We also reviewed the references and their relevance to the claims they support, and checked the experimental implementations, reported results, and their interpretation.

We reviewed and edited the final manuscript in its entirety and made the final decisions about all retained content. We take full responsibility for the correctness, originality, attribution, and integrity of the submission, including its mathematical claims and proofs, references, code, experimental results, and conclusions. This responsibility extends to all content developed with AI assistance and to any errors or omissions that remain.

REFERENCES

Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993. ISBN 9780136175490.

Ethem Alpaydin and Cenk Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. Dataset.

Haris Aziz and Barton E. Lee. Proportionally representative participatory budgeting with ordinal preferences. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 5110–5118, 2021.

Haris Aziz, Edith Elkind, Shenwei Huang, Martin Lackner, Luis Sánchez-Fernández, and Piotr Skowron. On the complexity of extended and proportional justified representation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 902–909, 2018.

Haris Aziz, Barton E. Lee, Sean Morota Chu, and Jeremy Vollen. Proportionally representative clustering. In *Proceedings of the 20th International Conference on Web and Internet Economics (WINE)*, pp. 155–171, 2024.

Daniel Bahrdt and Martin P. Seybold. Rational points on the unit sphere: Approximation complexity and practical constructions. In *Proceedings of the 2017 ACM International Symposium on Symbolic and Algebraic Computation (ISSAC)*, pp. 29–36, 2017.

Patrick Becker and Fabian Frank. On minimal achievable quotas in multiwinner voting. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 828–836, 2026.

Markus Brill and Jannik Peters. Robust and verifiable proportionality axioms for multiwinner voting. In *Proceedings of the 24th ACM Conference on Economics and Computation (EC)*, pp. 301, 2023.

Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. Proportionally fair clustering. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 1032–1041, 2019.

Gérard Cornuéjols, George L. Nemhauser, and Laurence A. Wolsey. A canonical representation of simple plant location problems and its applications. *SIAM Journal on Algebraic and Discrete Methods*, 1(3):261–272, 1980.

Chris Dong and Jannik Peters. An axiomatic analysis of proportionality notions in approval-based multiwinner voting, 2026. arXiv:2605.04612.

Edith Elkind and Martin Lackner. Structure in dichotomous preferences. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2019–2025, 2015.

Edith Elkind, Piotr Faliszewski, Ayumi Igarashi, Pasin Manurangsi, Ulrike Schmidt-Kraepelin, and Warut Suksompong. The price of justified representation. *ACM Transactions on Economics and Computation*, 12(3):11:1–11:27, 2024.

Seyed A. Esmaeili, Brian Brubach, Aravind Srinivasan, and John P. Dickerson. Fair clustering under a bounded cost. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Lester R. Ford, Jr. and Delbert R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404, 1956.

Delbert R. Fulkerson and Oliver A. Gross. Incidence matrices and interval graphs. *Pacific Journal of Mathematics*, 15(3):835–855, 1965.

Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, 2008.

Yu He, Jeremy Vollen, and Edith Elkind. Check, please: Verifiably fair clustering, 2026. arXiv:2605.12317.

Alan J. Hoffman and Joseph B. Kruskal. Integral boundary points of convex polyhedra. In Harold W. Kuhn and Albert W. Tucker (eds.), *Linear Inequalities and Related Systems*, volume 38 of *Annals of Mathematics Studies*, pp. 223–246. Princeton University Press, 1956.

Christopher Jerrett and Elliot Anshelevich. Low cost, fair, and representative committees in a metric space, 2025. arXiv:2510.21039.

Leon Kellerhals and Jannik Peters. Proportional fairness in clustering: A social choice perspective. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 111299–111317, 2024.

Bo Li, Lijun Li, Ankang Sun, Chenhao Wang, and Yingfan Wang. Approximate group fairness for clustering. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 6381–6391, 2021.

Oleg R. Musin. Graphs and spherical two-distance sets. *European Journal of Combinatorics*, 80: 311–325, 2019.

James B. Orlin. Max flows in $O(nm)$ time, or better. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 765–774, 2013.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Dominik Peters. Single-peakedness and total unimodularity: New polynomial-time algorithms for multi-winner elections. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1169–1176, 2018.

Jean-Claude Picard. Maximal closure of a graph and applications to combinatorial problems. *Management Science*, 22(11):1268–1272, 1976.

Raman Sanyal, Frank Sottile, and Bernd Sturmfels. Orbitopes. *Mathematika*, 57(2):275–314, 2011.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

Drew Springham. Faster verification of PJR+ via mincuts, 2026. arXiv:2609.20579.

Drew Springham, Edith Elkind, Bart de Keijzer, and Maria Polukarov. Multiwinner voting with interval preferences under incomplete information. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2026.

Mashbat Suzuki and Jeremy Vollen. Maximum flow is fair: A network flow approach to committee voting. In *Proceedings of the 25th ACM Conference on Economics and Computation (EC)*, pp. 964–983, 2024.

Luis Sánchez-Fernández, Edith Elkind, Martin Lackner, Norberto Fernández García, Jesús A. Fisteus, Pablo Basanta Val, and Piotr Skowron. Proportional justified representation. *Artificial Intelligence*, 353:104503, 2026.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. How to select datapoints for efficient human evaluation of NLG models? *Transactions of the Association for Computational Linguistics*, 13: 1789–1811, 2025.

A Details of verification

A.1 Proof of Theorem 3.1

The cut identity (3.1) optimizes over all groups within a ball about an unselected candidate. We first show that searching these balls is equivalent to using each group's own required radius; this is what makes the test exact rather than a sufficient condition.

Proof. Fix $c \in M \setminus X$. Every nonempty group S occurs in the network at $r(c, S) \in R(c)$. Conversely, if $S \subseteq N(c, r)$ is nonempty, then $r(c, S) \leq r$, and decreasing the radius to $r(c, S)$ cannot increase the number of selected centers reached by S. The empty group contributes zero. Thus (3.1)–(3.2) give

$$D_\gamma(X) = \max\left\{0, \max_{c \in M \setminus X} \max_{\varnothing \neq S \subseteq N} (kw(S) - W|X \cap A_{\gamma r(c,S)}(S)|)\right\}.$$

For an integer a, $\lfloor (kw(S) - Wa)/W \rfloor = \lfloor w(S)/q \rfloor - a$. There are finitely many groups and unselected candidates, so taking the floor commutes with taking their maximum. This proves (3.3), including the case $M = X$ under the empty-maximum convention.

The largest eligible level for a group is $\lfloor w(S)/q \rfloor \leq k$. Consequently, X violates γ-mPJR+ precisely when the left side of (3.3) is positive, or equivalently when $D_\gamma(X) \geq W$. This proves the decision statement.

If $D_\gamma(X) \geq W$, choose a maximizing pair (c, r) and take the source-side agents S of a minimum cut. This group is nonempty, and shrinking the radius to $r(c, S)$ cannot decrease its value. Since $D_\gamma(X)$ is already the global maximum, the value remains equal to $D_\gamma(X)$. The group therefore attains the largest number of missing centers as well. Finally, there are at most $(m - k)n$ pairs (c, r), and each network has at most $n + k + 2$ vertices and $nk + n + k$ arcs. □

A.2 Returning a violated proportionality requirement

The algorithm of Theorem 3.1 returns more than a cut value. Suppose the value in (3.1) is at least W at (c, r), and let S be the source-side agents of a minimum cut. Set $\ell = |X \cap A_{\gamma r(c,S)}(S)| + 1$. Since $r(c, S) \leq r$, the cut inequality gives

$$kw(S) \geq W(|X \cap A_{\gamma r}(S)| + 1) \geq W(|X \cap A_{\gamma r(c,S)}(S)| + 1) = W\ell.$$

Thus S is eligible at level ℓ but reaches only $\ell - 1$ selected centers within distance $\gamma r(c, S)$. Also $\ell \leq k$, since $\ell > k$ would imply $w(S) > W$. Thus the returned group and level violate (2.1). The argument covers zero distances and nonintegral quotas. When the pair (c, r) maximizes (3.2), the returned group also attains (3.3).

A.3 Continuing the flow at increasing radii

For a fixed unselected candidate c, an agent i enters the network at radius $d(i, c)$, and its arc to $x \in X$ enters at radius $\max\{d(i, c), d(i, x)/\gamma\}$. Sort these events, insert all events up to each checked radius in $R(c)$, and increase the current flow until it is maximum. All events at the same radius are inserted before that radius is tested.

Proposition A.1. *This procedure returns the same maximum-flow values as rebuilding every network. In the unweighted case, a direct integral implementation takes $O(nk \log(nk) + n^2k^2)$ time per unselected candidate.*

Proof. Adding capacity preserves feasibility. We use the residual-path method of Ford & Fulkerson (1956). A residual arc describes capacity on which the current flow can be increased, or existing flow that can be reversed. Continuing until there is no source-to-sink path of positive residual capacity gives a maximum flow in the enlarged network.

With unit agent weights, all capacities are integers and the final flow value is at most kn. Each successful path search increases the value by at least one, so there are at most kn successful searches over the entire sequence of radii. There is at most one unsuccessful search per checked radius,

and hence at most n such searches. Every search examines $O(nk)$ arcs. Sorting the events takes $O(nk \log(nk))$ time, which gives the stated bound. □

Continuing the flow remains valid with arbitrary agent weights. The displayed path-count bound is specifically for unit weights; the general weighted guarantee is the polynomial-time maximum-flow bound in Theorem 3.1.

A.4 Approval PJR+

For unit-weight approval elections, this maximum-flow test is due to Springham (2026). We state its weighted version and the resulting count of missing committee members to make the connection with Theorem 3.1 explicit.

In an approval election, voter i approves a set $\mathcal{A}_i \subseteq M$. A size-$k$ committee X satisfies weighted PJR+ if, for every level ℓ, every $c \in M \setminus X$, and every group S with $w(S) \geq \ell q$ and $c \in \mathcal{A}_i$ for all $i \in S$, $|X \cap \bigcup_{i \in S} \mathcal{A}_i| \geq \ell$. For each unselected candidate c, build the network of Section 3 using the voters who approve c, and include the arc $i \to x$ exactly when $x \in \mathcal{A}_i$. The cut calculation is unchanged.

Corollary A.2. *Weighted approval PJR+ can be checked by $m - k$ maximum-flow computations. The same computations give the largest number of missing committee members and a group attaining it whenever the number is positive.*

Proof. For each c, the cut value is

$$\max_{S:\, c \in \mathcal{A}_i \ \forall i \in S} \left(kw(S) - W \left| X \cap \bigcup_{i \in S} \mathcal{A}_i \right| \right).$$

Maximize also over c and include zero. Dividing by W and taking the floor gives the exact integer shortfall, as in the proof of Theorem 3.1. □

A.5 The finite search for the minimum dilation

Define the set of possible dilation values by

$$\mathcal{Q}(X) = \{1\} \cup \left\{ \frac{d(i,x)}{d(j,c)} : \begin{array}{l} c \in M \setminus X,\ i,j \in N,\ x \in X, \\ 0 < d(j,c), \quad d(i,c) \leq d(j,c) \leq d(i,x) \end{array} \right\}. \tag{A.1}$$

The first inequality excludes zero denominators. The second keeps only agents i present in the network at radius $d(j,c)$ and ratios of at least one.

Theorem A.3. *If some requirement in* (2.1) *is violated at $r(c,S) = 0$, then $\Gamma^+(X) = +\infty$. Otherwise $\Gamma^+(X)$ is finite, is attained, and belongs to $\mathcal{Q}(X)$. Sorting this set and using binary search takes $O(\log(2 + |\mathcal{Q}(X)|))$ calls to Theorem 3.1, with $|\mathcal{Q}(X)| \leq 1 + (m-k)n^2k$.*

Proof. At radius zero, the condition $d(i,x) \leq \gamma r$ does not depend on any finite γ. Such a violation cannot be removed by dilation.

Assume all zero-radius requirements hold. For a fixed positive checked radius $r = d(j,c)$, the participating agents do not depend on γ. An arc $i \to x$ enters exactly at $\gamma = d(i,x)/r$. All such events with $\gamma \geq 1$ are among the ratios in (A.1). Between consecutive values of $\mathcal{Q}(X)$, all networks are unchanged. At equality, the new arcs are present because the balls are closed.

Increasing γ cannot decrease the number of reached selected centers. For sufficiently large γ, every agent present at each positive checked radius reaches all selected centers. Together with the zero-radius requirements, this shows that X satisfies γ-mPJR+ for all sufficiently large γ. Thus $\Gamma^+(X)$ is finite and attained at 1 or at one of the ratios in (A.1). The monotonicity permits binary search. The size bound follows by counting choices of (c,i,j,x). □

One can first test $\gamma = 1$ and handle $M = X$ immediately. Sorting (A.1) requires $O(|\mathcal{Q}(X)| \log(2 + |\mathcal{Q}(X)|))$ comparisons. The number of maximum-flow computations in the subsequent search is $O((m-k)n \log(2 + |\mathcal{Q}(X)|))$. This is an exact finite algorithm; we do not claim that explicitly constructing all these ratios is the best practical approach.

B LINE CHARACTERIZATIONS AND FAST VERIFICATION

B.1 WHY ELIGIBLE BLOCKS SUFFICE

We use the sorted agent locations, weights, and eligible blocks defined in Section 4.

Lemma B.1. *If a selection violates weighted mPJR or weighted γ-mPJR+ at level ℓ, then it violates the same condition on a level-ℓ eligible block.*

Proof. Let S be a violating group with smallest and largest indices a and b. Since all weights are positive and $w(S) \geq \ell q$, the index $b_\ell(a)$ exists and is at most b. Put $I = [a, b_\ell(a)]$.

For mPJR, keep the radius r at which S violates the condition. Every candidate within r of all agents in S is within r of both extreme locations and therefore of every location between them. Thus I remains ℓ-cohesive. The radius-r intervals about members of S share a point, so their union is $[p_a - r, p_b + r]$. The corresponding union for I is contained in this interval. Consequently I reaches no additional selected center and remains underrepresented.

For γ-mPJR+, keep the same unselected candidate c. We have $r(c, I) \leq r(c, S)$. Every radius-$\gamma r(c, S)$ interval about a member of S contains c, as does every radius-$\gamma r(c, I)$ interval about a member of I. Both unions are intervals. The second is contained in the first because its extreme locations lie between p_a and p_b and its radius is no larger. Again, I reaches no additional selected center. □

B.2 THE RELEVANT RADII

For every level-ℓ eligible block $I = [a, b]$, retain μ_I, h_I, and $\alpha_\ell(I)$ from Section 4, and define

$$\beta_\ell(I) = \max\{0, \mathrm{near}_\ell(\mu_I; X) - h_I\}, \tag{B.1}$$

$$\eta(I) = h_I + \mathrm{near}_1(\mu_I; M \setminus X). \tag{B.2}$$

Set $\eta(I) = +\infty$ when $M = X$. Write $\mathrm{dist}(c, [p_a, p_b]) = \min_{t \in [p_a, p_b]} |c - t|$ for the distance to the interval. The identities

$$\max\{|c - p_a|, |c - p_b|\} = h_I + |c - \mu_I|, \quad \text{and} \quad \mathrm{dist}(c, [p_a, p_b]) = \max\{0, |c - \mu_I| - h_I\}$$

explain all three quantities. The block first has ℓ common candidates at $\alpha_\ell(I)$, and its smallest radius to a common unselected candidate is $\eta(I)$. Whenever $r \geq h_I$, its radius-r intervals overlap and their union is $[p_a - r, p_b + r]$. At such a radius, it reaches at least ℓ selected centers exactly when $r \geq \beta_\ell(I)$.

The restriction $r \geq h_I$ is important: at smaller radii, intervals around discrete agent locations need not cover the gaps between those locations. Every radius used in the proportionality tests satisfies this restriction, since $\alpha_\ell(I) \geq h_I$ and $r(c, I) \geq h_I$.

Theorem B.2. *On the real line, a selection X satisfies weighted mPJR if and only if*

$$\beta_\ell(I) \leq \alpha_\ell(I) \quad \textit{for every level-}\ell\textit{ eligible block } I. \tag{B.3}$$

It satisfies weighted γ-mPJR+ if and only if

$$\beta_\ell(I) \leq \gamma\eta(I) \quad \textit{for every such block.} \tag{B.4}$$

Proof. At radius $\alpha_\ell(I)$, the block is cohesive and its intervals overlap. It reaches enough selected centers precisely when (B.3) holds. Larger radii only add centers, so checking this first radius is sufficient for the block. Lemma B.1 extends the equivalence to all eligible groups.

For mPJR+ and a fixed unselected c, we have $r(c, I) = h_I + |c - \mu_I|$. At dilation γ, the block reaches enough selected centers if and only if $\beta_\ell(I) \leq \gamma(h_I + |c - \mu_I|)$. Requiring this for all unselected candidates is equivalent to using the smallest right-hand side, namely $\gamma\eta(I)$. Apply Lemma B.1 again. If there is no unselected candidate, the condition is vacuous, as is (B.4) under the stated convention. □

This theorem also proves (4.1): at radius $\alpha_\ell(I)$, the selected centers reached by the block are exactly $X \cap J_\ell(I)$.

Corollary B.3. *For weighted line instances,*

$$\Gamma^{+}(X) = \max\left\{1, \max_{\ell, I} \frac{\beta_\ell(I)}{\eta(I)}\right\}, \tag{B.5}$$

where I ranges over the level-ℓ eligible blocks. The ratio is zero when both numerator and denominator are zero or when the denominator is infinite, and is infinite when the denominator is zero and the numerator is positive.

Proof. For a positive finite $\eta(I)$, condition (B.4) is equivalent to $\gamma \geq \beta_\ell(I)/\eta(I)$. The zero and infinite cases have exactly the meanings specified in the statement. Taking all conditions together and imposing $\gamma \geq 1$ proves the formula. □

B.3 Processing nearest-neighbor queries together

The characterizations reduce verification to nearest-neighbor order statistics at the block midpoints. Because these midpoints are sorted, the minimizing candidate window can be updated rather than found anew at every query. The following lemma makes this saving precise.

Lemma B.4. *Let $Z = (z_1, \ldots, z_s)$ be a list of sorted points, so $z_1 \leq \cdots \leq z_s$, and fix $1 \leq j \leq s$. For nondecreasing queries $\mu_1 \leq \cdots \leq \mu_T$, all values $\mathrm{near}_j(\mu_t; Z)$ can be computed in $O(s+T)$ time and $O(1)$ additional space.*

Proof. The j nearest points can be chosen as a window $z_a, \ldots, z_{a+j-1}$ of consecutive points: whenever two chosen points have an unchosen point between them, that point is no farther from μ than the farther of the two and can replace it. Repeating gives a consecutive choice, including when distances tie. Therefore

$$\mathrm{near}_j(\mu; Z) = \min_{1 \leq a \leq s-j+1} g_a(\mu), \quad g_a(\mu) = \max\{\mu - z_a, z_{a+j-1} - \mu\}.$$

For a fixed μ, the first term is nonincreasing in a and the second is nondecreasing. Before the two terms cross, their maximum is the first term; after they cross, it is the second. Thus, as a function of a, $g_a(\mu)$ is first nonincreasing and then nondecreasing.

We next show that the rightmost minimizing index cannot move left as μ increases. Write

$$g_a(\mu) = \frac{z_{a+j-1} - z_a}{2} + \left|\mu - \frac{z_a + z_{a+j-1}}{2}\right|.$$

The midpoints $(z_a + z_{a+j-1})/2$ are nondecreasing in a. If $a < b$, the difference $g_a(\mu) - g_b(\mu)$ is nondecreasing in μ: the difference of the two absolute-value terms is constant to the left of both midpoints, increases between them, and is constant to their right. Hence once window b is no worse than window a, it stays no worse at every later query.

Start with the leftmost window for the first query. For each query, advance the current index while the next value is no larger. The preceding two observations imply that this stops at the rightmost minimizing window. The index advances at most $s - j$ times in total, and each query requires only constant work apart from those advances. This proves the time and space bounds, including ties. □

B.4 Proof of Theorem 4.1

For a fixed level ℓ, generate eligible blocks with two indices. Starting at a, increase the right endpoint until the block weight first reaches ℓq. Output the block, subtract w_a, and continue with $a + 1$. Positive weights imply that the right endpoint never decreases. Thus all blocks at this level are generated in $O(n)$ time. Their midpoints are also nondecreasing.

After sorting the agents and candidates, there are two ways to evaluate the required quantities. An individual nearest-neighbor query takes $O(\log(m+1))$ time. In the notation of Lemma B.4, binary search finds the first window index a for which $\mu - z_a \leq z_{a+j-1} - \mu$. To its left, $g_a(\mu)$ is nonincreasing; from this index onward, it is nondecreasing. Comparing this window with its predecessor gives the minimum. If there is no crossing, use the last window; if the crossing is at the first window, no predecessor is needed. Alternatively, Lemma B.4 evaluates the queries on each of

M, X, and $M \setminus X$ in a combined $O(n+m)$ time per level. The empty unselected list is handled by setting $\eta = +\infty$.

Choose the faster method for the input sizes. Evaluating (B.3), (B.4), and the ratios in (B.5) takes constant additional time per block. Sorting costs $O((n+m)\log(n+m))$, and the k levels cost $O(k \min\{n\log(m+1), n+m\})$. Store only the sorted input and the current level's blocks and queries. This uses $O(n+m)$ space and completes the proof.

C Exact optimization and repair

C.1 Existence with arbitrary positive weights

Before optimizing over proportional selections, we should establish that at least one exists. The following weighted version of the expanding-radius argument for SEAR makes this explicit (Aziz et al., 2024; He et al., 2026). The argument works directly with weights, without replicating agents.

Lemma C.1 (Weighted existence)**.** *Every finite metric instance with positive agent weights has a k-element selection satisfying weighted mPJR+ and hence weighted mPJR.*

Proof. Give agent i remaining weight w_i. In each of k rounds, choose the smallest radius r for which some unselected candidate has at least q remaining weight within distance r. Select such a candidate and decrease the weights of agents in its ball by a total of exactly q, without making any weight negative. A choice always exists: before round $t \le k$, at least one candidate remains and the total remaining weight is $(k-t+1)q \ge q$. A sufficiently large ball about that candidate contains all agents. The chosen radii do not decrease, because weights only decrease and selected candidates are removed. After k rounds, every agent has remaining weight zero.

Let X be the final selection. Fix an unselected candidate c, a group S with $w(S) \ge \ell q$, and $r = r(c,S)$. Consider the rounds with chosen radius at most r. Immediately after these rounds, the remaining weight in $B(c,r)$ is less than q: otherwise the next radius could not exceed r. If these are all k rounds, the same conclusion follows from zero remaining weight. Since $S \subseteq B(c,r)$, more than $(\ell-1)q$ weight has been removed from S in these rounds. Each chosen center removes at most q, so at least ℓ distinct centers remove positive weight from S. Each is within distance r of an agent in S. Thus $|X \cap A_r(S)| \ge \ell$, as required. □

C.2 Loss to several nearby centers

Theorem 4.2 also allows the loss of an agent to depend on several of its closest selected centers. For each agent i, fix an integer $r_i \in \{1,\dots,k\}$ specifying how many selected centers enter its loss, and let $d_{i,(h)}(X)$ be its hth smallest distance to a selected center. Equal distances are counted separately for distinct selected centers. The generalized objective is

$$F(X) = \sum_i w_i \sum_{h=1}^{r_i} \phi_i(d_{i,(h)}(X)) + \sum_{c_j \in X} a_j. \tag{C.1}$$

When every $r_i = 1$, this is the objective F of Section 4. Retain the distance values ρ_{it}, intervals H_{it}, and loss increments Δ_{it} from Section 4. We claim that

$$w_i \sum_{h=1}^{r_i} \phi_i(d_{i,(h)}(X)) = r_i w_i \phi_i(\rho_{i1}) + \sum_{t<s_i} \Delta_{it} \left[r_i - |X \cap H_{it}|\right]_+ . \tag{C.2}$$

To see this, expand each selected-center loss into its successive increments along $\rho_{i1},\dots,\rho_{is_i}$. At threshold ρ_{it}, exactly $[r_i - |X \cap H_{it}|]_+$ of the r_i closest selected centers lie strictly beyond the threshold. Interchanging the two finite sums proves the identity. Because the intervals include candidates at exactly the threshold, the argument also covers tied distances.

In (4.3), replace the loss constraints and bounds by

$$\sum_{j:c_j \in H_{it}} x_j + z_{it} \ge r_i, \quad 0 \le z_{it} \le r_i. \tag{C.3}$$

The objective coefficients and all fairness constraints are unchanged. Since $\Delta_{it} \geq 0$, the shortage can always be chosen to equal the positive part in (C.2). It remains to prove integrality and give the promised flow.

C.3 Integrality and feasibility

We give a direct proof of the classical consecutive-ones matrix property used in the main text (Fulkerson & Gross, 1965). A matrix whose rows are indicator vectors of intervals is totally unimodular. Consider any square submatrix. The ones remain consecutive in each row, even after some columns have been removed. Replace each column after the first by its difference from the preceding column, performing the replacements simultaneously. This is a determinant-one transformation. Each row now has at most one $+1$ and at most one -1.

Such a square matrix has determinant in $\{0, 1, -1\}$. If every row has two nonzero entries, every row sums to zero and the matrix is singular. Otherwise expand the determinant along a row with at most one nonzero and proceed by induction. This proves the claim. Changing row signs, appending a column with a single $+1$ or -1, or appending a unit row preserves total unimodularity by the same determinant expansion argument.

The fairness rows, loss rows on x, and total-size row are interval rows. Equality in the total-size constraint may be written as two signed rows. Variable bounds on x are unit rows. Each z_{it} column has a single nonzero among these constraints, and its bound rows can be added afterwards. Thus the full constraint matrix is totally unimodular. All right-hand sides are integers: k, ℓ, r_i, 0, and 1. The bounded linear program therefore has integral vertices. At every optimal integral vertex, $x_j \in \{0, 1\}$, as required.

Feasibility follows directly from Lemma C.1, including for arbitrary positive real weights. Once a feasible selection is fixed, each shortage can be set to $[r_i - |X \cap H_{it}|]_+$, within its stated bounds.

The formulation has $m + O(nm)$ variables and $O(nk + nm + m)$ constraints. Together with (4.1) and (C.2), integrality proves exact polynomial-time optimization. We next obtain the smaller flow formulation claimed in Theorem 4.2.

C.4 One minimum-cost flow

Put $y_j = \sum_{h=1}^{j} x_h$ and $y_0 = 0$. Every interval count is a difference of two y variables. We identify a candidate interval $\{c_a, \ldots, c_b\}$ with its index interval $[a, b]$ in the construction below.

Define the required net inflow at each of the vertices $0, \ldots, m$ by

$$b_0 = a_1, \quad b_j = a_{j+1} - a_j \ (1 \leq j < m), \quad b_m = -a_m. \tag{C.4}$$

These quantities sum to zero, and

$$\sum_{j=1}^{m} a_j x_j = -\sum_{j=0}^{m} b_j y_j.$$

A positive b_j is demand and a negative b_j is supply. Add the arcs in Table 3. Parallel arcs are retained, and arcs with zero capacity may be omitted.

Table 3: The minimum-cost flow for exact proportional selection. An infinite capacity means that no upper bound is imposed.

Arc and associated requirement	Cost	Capacity
$j - 1 \to j$, for each $1 \leq j \leq m$	1	$+\infty$
$j \to j - 1$, for each $1 \leq j \leq m$	0	$+\infty$
$0 \to m$	k	$+\infty$
$m \to 0$	$-k$	$+\infty$
$b \to a - 1$, for $J_\ell(I) = [a, b]$	$-\ell$	$+\infty$
$b \to a - 1$, for $H_{it} = [a, b]$	$-r_i$	Δ_{it}

If f is a minimum-cost flow with the balances (C.4), then

$$\min_X F(X) = \sum_i r_i w_i \phi_i(\rho_{i1}) - \text{cost}(f), \tag{C.5}$$

where the minimum on the left is over mPJR selections of size k.

Deriving the network. We derive the network by linear-programming duality; its flows are dual variables, not indicators of selected centers. First drop the upper bounds $z_{it} \le r_i$: since $x_j \ge 0$, the smallest feasible nonnegative shortage never exceeds r_i. Its objective coefficient is nonnegative, so removing this bound does not change the optimum. The constraints on the remaining variables are

$$\begin{aligned} y_j - y_{j-1} &\le 1, & y_{j-1} - y_j &\le 0, \\ y_m - y_0 &\le k, & y_0 - y_m &\le -k, \\ y_{a-1} - y_b &\le -\ell & &\text{for a fairness interval } [a,b], \\ y_{a-1} - y_b - z_{it} &\le -r_i & &\text{for a loss interval } [a,b]. \end{aligned}$$

We may treat all y_j as free, since translating them by a common constant changes neither the constraints nor the objective.

For a difference constraint $y_v - y_u \le c$, introduce a multiplier $f_{uv} \ge 0$ and add $f_{uv}(y_v - y_u - c)$ to the minimization objective. Minimizing over the free y variables forces net inflow b_j at each vertex. For a loss constraint, the coefficient of $z_{it} \ge 0$ becomes $\Delta_{it} - f_{uv}$, so minimization also forces $f_{uv} \le \Delta_{it}$. The resulting dual maximizes minus the total arc cost. These are exactly the balances, costs, and capacities of Table 3, proving (C.5) by linear-programming duality. The first four types of arcs impose the bounds $0 \le x_j \le 1$ and the equality $\sum_j x_j = k$.

Recovering selected centers. We use the standard shortest-path potentials for minimum-cost flow (Ahuja et al., 1993). The costs of all arcs in Table 3 are integers, even when weights, loss values, and balances are not. The residual network contains every direction in which flow can be increased or reversed, with the corresponding cost. An optimal flow has no negative-cost cycle in this network, since sending a positive amount around such a cycle would preserve all balances and decrease the cost. Add a temporary source with zero-cost arcs to all vertices, and let π_j be the resulting shortest-path distances. They are finite because the temporary source reaches every vertex and there is no negative-cost cycle. They are integers because arc costs are integers. Every residual arc $u \to v$ of cost c satisfies $\pi_v - \pi_u \le c$. The uncapacitated arcs always have forward residual capacity. The first four types therefore give

$$x_j = \pi_j - \pi_{j-1} \in \{0,1\}, \quad \sum_j x_j = k,$$

and the fairness arcs give every required lower bound on interval counts.

For completeness, the recovered selection is optimal, not just feasible. An uncapacitated arc with positive flow also has a reverse residual arc, making its difference constraint tight. Consider a loss arc of capacity $\Delta_{it} > 0$ for an interval containing u selected centers. If its flow is smaller than Δ_{it}, its forward residual arc gives $u \ge r_i$, so the positive-part shortage is zero. If its flow is positive, its reverse residual arc gives $u \le r_i$, so the loss constraint with shortage $[r_i - u]_+$ is tight. Thus the shortage is positive only when the arc is at capacity, and a positive arc flow makes its constraint tight. Together with the flow balances, these are the complementary slackness conditions: a positive multiplier makes its constraint tight, and a positive shortage makes its dual capacity bound tight. A zero-capacity loss arc has zero objective coefficient and can use its positive-part shortage without any extra condition. Hence the recovered selection attains (C.5).

There are $m + 1$ flow vertices, $2m + 2$ basic arcs, at most nk fairness arcs, and at most $n(m - 1)$ loss arcs. Preprocessing consists of forming the eligible blocks, their candidate intervals, and the sorted distance lists. It is polynomial in the input size. This completes the proof of Theorem 4.2, including the general objective (C.1).

Finite capacities in an implementation. If a solver requires finite capacities, let T be the total supply plus the sum of all finite capacities in Table 3. The uncapacitated arcs may be replaced by

capacity $T+1$. To see this, take a finite optimal flow and decompose it into simple supply-to-demand paths and cycles. Such an optimum exists: the linear-programming dual above is feasible and has a finite attained optimum. Positive-cost cycles cannot occur in an optimum, and zero-cost cycles can be removed. Every negative-cost cycle contains a finite-capacity loss arc, because the uncapacitated difference constraints are feasible and therefore have no negative-cost cycle. The total weight of the paths is the total supply. Assign each remaining negative-cost cycle to one finite-capacity arc it uses; the total weight assigned to an arc is at most that arc's flow, hence at most its capacity. Thus the total weight of these cycles is at most the sum of the finite capacities. Hence every arc carries at most T, so the finite substitute preserves the optimal value, and any optimal flow it returns is also optimal in the original network. For shortest-path recovery, use the residual network of that original problem: include the forward direction of every originally uncapacitated arc, regardless of its flow. The accompanying optimizer uses this finite-capacity implementation and checks both flow feasibility and optimality with integer arithmetic (Appendix F.1).

C.5 Prioritizing either replacements or loss

For $X^0 \subseteq M$ with $|X^0| = k$, set all losses to zero and $a_j = \mathbf{1}[c_j \notin X^0]$. The objective becomes $|X \setminus X^0|$. Since $|X| = |X^0|$, this equals $|X^0 \setminus X|$ and counts replacements rather than counting removals and additions twice. All loss arcs disappear, leaving $m+1$ vertices and $O(nk+m)$ arcs. This proves Corollary 4.3.

To minimize replacements first and clustering loss second, multiply this opening cost by any number larger than the range of the secondary objective. For the loss part of (C.1), the choice $1 + \sum_i w_i r_i(\phi_i(\rho_{is_i}) - \phi_i(\rho_{i1}))$ suffices. An additional replacement then costs more than any possible improvement in the secondary loss. This is a choice of objective coefficients within the same flow, not an additional constraint on the feasible selections.

The opposite priority is also useful: minimize loss first, then replacements. If the clustering loss $C(X)$ is integer-valued, minimizing $(k+1)C(X) + |X \setminus X^0|$ achieves this priority, because changing the loss by one outweighs the entire possible replacement count. The experiments use this choice to avoid unnecessary replacements among proportional selections with the same optimal loss.

C.6 Minimizing the maximum distance

A different clustering objective asks that no agent be too far from its nearest selected center. The interval formulation also solves this problem exactly.

Corollary C.2 (Proportional discrete k-center)**.** *On the line, a weighted mPJR selection minimizing* $\max_i d(i, X)$ *can be found with* $O(\log(2+nm))$ *minimum-cost-flow computations. Each network has* $m+1$ *vertices and* $O(nk+n+m)$ *arcs.*

Proof. Fix a proposed maximum distance R, and put $H_i(R) = M \cap [p_i - R, p_i + R]$. Minimize the number of agents whose nearest selected center is farther than R. Agent i contributes $[1 - |X \cap H_i(R)|]_+$, an interval shortage with coefficient one. An empty $H_i(R)$ contributes the constant one; $H_i(R) = M$ contributes zero. Equivalently, take $\phi_i(t) = \mathbf{1}[t > R]/w_i$ and $a_j = 0$ in Theorem 4.2. The flow then has at most one loss arc per agent. The minimum number of agents farther than R is zero exactly when some mPJR selection has a center within distance R of every agent.

This property is monotone in R. The optimal maximum distance is one of the at most nm agent–candidate distances, and the largest such distance is feasible by Lemma C.1. Sort these values and use binary search, retaining the selection at the smallest feasible value. This proves the stated number and size of flow computations. □

C.7 An exact example with arbitrarily large cost increase

We refine the dense-group example in Jerrett & Anshelevich (2025, Section 2) and prove Proposition 4.4 by calculating both optima for a fixed-size mPJR instance with positive unconstrained cost. Fix an integer $T \geq 3$. Place four agents at 0, one at $10T+1$, and one at $11T+1$. Let $M = \{-1, 0, 1, 10T, 11T\}$ and $k = 3$, so $q = 2$. Each distant agent is at distance at least one from

every candidate. Therefore every selection has cost at least two, and $\{0, 10T, 11T\}$ attains two. This is the positive unconstrained optimum for every $p > 0$.

The four agents at 0 are eligible at level two and have three common candidates at radius one. Every mPJR selection must consequently include at least two of $\{-1, 0, 1\}$ and at most one of $\{10T, 11T\}$. If it includes $11T$ but not $10T$, the two distant agents contribute $(T-1)^p + 1$. If it includes $10T$ but not $11T$, they contribute $1 + (T+1)^p$. With neither distant candidate, their contribution is still larger. Hence every mPJR selection costs at least $(T-1)^p + 1$.

The selection $X = \{0, 1, 11T\}$ attains this bound and satisfies the stronger mPJR+ condition. Both unselected candidates, -1 and $10T$, are at distance at least one from every agent, so every required radius $r(c, S)$ is at least one. A level-one eligible group contains either an agent at 0 or both distant agents; it therefore reaches 0 or $11T$ within distance one. A level-two eligible group has at least four agents, so it contains an agent at 0 and reaches both 0 and 1 within distance one. The only level-three eligible group is the whole population, which also reaches $11T$ from the agent at $11T + 1$. Thus every mPJR+ requirement holds. Since mPJR+ implies mPJR, the same selection attains both proportional minima.

Thus both proportional minima are exactly $(T-1)^p + 1$, and their ratio to the unconstrained optimum is

$$\frac{(T-1)^p + 1}{2} \longrightarrow +\infty \quad \text{as } T \longrightarrow +\infty.$$

This proves the proposition without relying on an unconstrained optimum of zero cost.

D Proof of fixed-dimensional hardness

We prove Theorem 5.1 in two steps. First we give a reduction that requires only a particular strict ordering of distances. We then realize that ordering in four dimensions and place the resulting points on a sphere in five dimensions. Keeping these steps separate allows the proportionality argument and the geometry to be checked independently.

D.1 A reduction that is valid at every radius

We reduce from Balanced Biclique: given a bipartite graph and an integer t, determine whether the graph contains a complete bipartite subgraph with t vertices in each part (Garey & Johnson, 1979, Problem GT24). We may assume $t \geq 3$ and that the graph has at least $2t$ vertices. The excluded cases can be decided directly, and the second assumption bounds t by the graph size.

Set $B = 2t + 1$ and $q = t^2$. The later construction adds selected centers and agents to control what happens at larger radii. We first adjust the graph's density so that the required number of those agents is nonnegative and does not itself create underrepresentation. If the original graph has v_0 vertices, add a disjoint complete bipartite graph with parts of sizes $t - 1$ and $s = 2(v_0 + t)$. This component cannot contain a balanced biclique of size t, and such a biclique cannot span distinct components. The answer to the graph problem is therefore unchanged.

Let v and E be the number of vertices and the edge set of the enlarged graph. The construction ensures

$$B|E| \geq vq. \tag{D.1}$$

Indeed, ignoring all original edges, the difference between the left and right sides is at least

$$((2t+1)(t-1) - t^2)s - t^2(v_0 + t - 1).$$

For $t \geq 3$, the coefficient of s is $t^2 - t - 1 \geq t^2/2$. With $s = 2(v_0 + t)$, the displayed difference is positive.

Create one selected candidate for each graph vertex. For each graph edge, create B colocated agents, called *edge agents*. Define

$$K = \left\lceil \frac{B|E|}{q} \right\rceil, \quad k = v + K, \quad F = kq - B|E|. \tag{D.2}$$

Add K additional selected candidates. Also add F agents, all at a point z. The chosen selection X consists of the vertex candidates and the K additional selected candidates. Thus $|X| = k$, the

total number of agents is $n = kq$, and the quota W/k is indeed $q = t^2$. Each of the two populations is small enough to be represented by the K additional selected candidates once those candidates become reachable:

$$B|E| \le Kq, \quad 0 \le F \le Kq. \tag{D.3}$$

The first bound follows from the definition of K. For the second, $Kq \ge B|E|$ gives $F \ge 0$, and (D.1) gives $F = vq + Kq - B|E| \le Kq$.

For mPJR, add B distinct unselected candidates. For the specified-level mPJR+ problem, add only one unselected candidate and set the tested level to $\ell = B$. This is a valid level because (D.1) implies $K \ge v$ and hence $k \ge 2v \ge 4t > B$.

Required distances. For each edge agent, distinguish its two endpoint candidates from the other vertex candidates. We require the following *global* ordering: every distance in an earlier row is smaller than every distance in a later row,

$$\begin{array}{rl} & \text{edge agent to one of its endpoint candidates} \\ < & \text{edge agent to an unselected candidate} \\ < & \text{edge agent to an additional selected candidate} \\ < & \text{edge agent to a nonendpoint vertex candidate.} \end{array} \tag{D.4}$$

Write $D = \max_{p \text{ an edge agent},\, x \in X} d(p, x)$. The point z must satisfy

$$\max_{x \text{ an additional selected candidate}} d(z, x) < \min_{c \text{ unselected}} d(z, c), \tag{D.5}$$

$$\min_{p \text{ an edge agent}} d(z, p) > 2D. \tag{D.6}$$

The next subsection realizes all these strict inequalities in $\mathbb{R}^4$.

A balanced biclique gives an mPJR violation. Take all Bq agents associated with the $q = t^2$ edges of a balanced biclique. Choose a radius at least the largest edge-agent-to-unselected-candidate distance and smaller than every edge-agent-to-additional-selected-candidate distance. This group has B common candidates and is eligible at level B. It reaches exactly its $2t = B - 1$ selected endpoint candidates. Thus it violates mPJR.

Every mPJR violation gives a balanced biclique. Suppose S violates mPJR at radius r and level ℓ. We first determine which agents and radii can occur.

If S includes both an edge agent and an agent at z, cohesion gives a candidate within distance r of both. The triangle inequality and (D.6) imply $r > D$. The edge agent then reaches every selected candidate, contradicting underrepresentation.

If S consists only of agents at z, at least one common candidate must be unselected; otherwise its common selected candidates already provide the required number. By (D.5), reaching an unselected candidate makes all K additional selected candidates reachable. But $|S| \le F \le Kq$, so eligibility implies $\ell \le K$. Again there is no violation.

It follows that S consists only of edge agents. If r is at least the largest distance from an edge agent to an additional selected candidate, then S reaches all K such candidates. Since $|S| \le B|E| \le Kq$, this is enough. At any remaining violating radius, some unselected candidate is common to S; otherwise the common selected candidates alone would satisfy mPJR. The global ordering (D.4) then makes the endpoints of every edge with an agent in S reachable. No nonendpoint vertex candidate is reachable, because the radius is below the largest edge-agent distance to an additional selected candidate, which is below every nonendpoint distance. Some, but not necessarily all, additional selected candidates may already be reached.

Let b be the number of distinct endpoint vertices among the edges with an agent in S, let $h \in \{0, 1, 2\}$ be the number of endpoints common to all these edges, and let g be the number of additional selected candidates reached by S. The group reaches exactly $b + g$ selected candidates. Its number of common candidates is at most $B + h + g$. Put $j = \ell - g$. Underrepresentation and cohesion imply

$$b < j \le B + h, \quad \ell q \le |S| \le B \left\lfloor \frac{b^2}{4} \right\rfloor. \tag{D.7}$$

For the upper bound, a bipartite graph with b vertices and part sizes u and $b-u$ has at most $u(b-u) \le \lfloor b^2/4 \rfloor$ edges. Each edge contributes at most B agents, even if S contains only some of its copies. Subtracting g in the definition of j allows the same count to handle radii at which additional centers have already been reached. We now consider the possible values of j.

If $j < B$, then

$$|S| \le B \left\lfloor \frac{(j-1)^2}{4} \right\rfloor < jq \le \ell q,$$

a contradiction. For the strict inequality, $(j-1)^2/j$ is strictly increasing on positive integers. Since $j < B = 2t+1$,

$$\frac{(j-1)^2}{j} < \frac{(B-1)^2}{B} = \frac{4q}{B},$$

which gives $B(j-1)^2/4 < jq$.

If $j = B$, then

$$Bq \le \ell q \le |S| \le B \left\lfloor \frac{b^2}{4} \right\rfloor \le Bt^2 = Bq.$$

All inequalities are equalities. Since $\ell = B + g$, we get $g = 0$. Since $b \le 2t$ and $\lfloor (2t-1)^2/4 \rfloor < t^2$, we get $b = 2t$. These edges attain the maximum of t^2 edges on these vertices. The two parts each have t vertices, and every edge between them is present. This is a balanced biclique.

Finally, if $j > B$, then $h \ge 1$. All edges with an agent in S share a vertex and form a star, so

$$|S| \le B(b-1) \le B(j-2) < jq \le \ell q.$$

The strict inequality uses $q = t^2 \ge 2t+1 = B$ for $t \ge 3$. This is impossible. We have proved that every violation, at any radius and any level, implies a balanced biclique.

Specified-level mPJR+. Use the same population and geometry, but retain only one unselected candidate c and test level B. A balanced biclique gives the same violating group at radius $r(c, S)$. Conversely, the preceding population and distance arguments exclude groups at z and mixed groups. For a group S of edge agents, $r(c, S)$ lies between the endpoint distances and the additional-selected-candidate distances, so the selected centers that S reaches are exactly the endpoints of the edges with an agent in S. An eligible violating group has at least Bq agents and at most $B - 1 = 2t$ such endpoints. The bound $|S| \le B\lfloor b^2/4 \rfloor \le Bt^2 = Bq$ forces equality and gives a balanced biclique. No additional cohesion requirement is used for mPJR+.

D.2 Realizing the distance ordering in four dimensions

All norms in this construction are Euclidean. Let the enlarged graph have v vertices and set

$$\varepsilon = \frac{1}{100v^3}, \quad y(\theta) = (\cos\theta, \sin\theta, \tfrac{1}{2}\cos 2\theta, \tfrac{1}{2}\sin 2\theta), \quad x(\theta) = \tfrac{4}{5}y(\theta).$$

Then $\|y(\theta)\|^2 = 5/4$ and $\|x(\theta)\|^2 = 4/5$. Enumerate the left vertices and assign their angles $\alpha_i = 2\arctan(i\varepsilon)$. Enumerate the right vertices and assign $\beta_j = \pi + 2\arctan((v+j)\varepsilon)$. The vertex candidate at angle θ is $x(\theta)$. These coordinates are rational because

$$\cos(2\arctan u) = \frac{1-u^2}{1+u^2}, \quad \sin(2\arctan u) = \frac{2u}{1+u^2}.$$

The candidates in each part lie on a short arc, and the two arcs are almost opposite. No assigned cross pair is exactly antipodal.

Separating a prescribed pair by a nonnegative trigonometric polynomial is classical (Sanyal et al., 2011, Section 5); we additionally need uniform distance bounds over all edges. For an edge with endpoint angles α, β, the product below vanishes at those endpoints and is positive at every other assigned vertex angle. Expanding it gives the point at which to place the edge agents. Define

$$\begin{aligned} f(\theta) &= (1-\cos(\theta-\alpha))(1-\cos(\theta-\beta)), \\ A &= (\cos\alpha + \cos\beta,\ \sin\alpha + \sin\beta, -\cos(\alpha+\beta),\ -\sin(\alpha+\beta)), \\ a &= 1 + \tfrac{1}{2}\cos(\alpha-\beta), \quad p = A/\|A\|, \quad \delta = a - \|A\|/2. \end{aligned}$$

Place the edge agents at p. A direct expansion gives

$$f(\theta) = a - A \cdot y(\theta), \quad \|A\|^2 = 3 + 2\cos(\alpha - \beta).$$

Since $\|p\| = 1$ and $\|x(\theta)\|^2 = 4/5$, it follows that

$$\|p - x(\theta)\|^2 - 1 = \frac{8}{5\,\|A\|}(f(\theta) - \delta). \tag{D.8}$$

We now bound δ and the nonzero values of f. These bounds place every endpoint distance below 1 and every nonendpoint distance above 1, with enough separation to add the other candidates.

Write $\omega = \beta - \alpha - \pi$ and $u = 1 - \cos\omega$. The derivative of $2\arctan t$ lies between one and two over the parameter range used above. Therefore

$$\varepsilon \le \omega \le 4v\varepsilon < 1, \quad \frac{\varepsilon^2}{4} \le u \le 8v^2\varepsilon^2 < 1.$$

Also $a = (1+u)/2$ and $\|A\| = \sqrt{1+2u}$, so rationalizing the difference gives

$$\delta = \frac{u^2}{2(1+u+\sqrt{1+2u})}, \quad \frac{\varepsilon^4}{192} \le \delta \le 16v^4\varepsilon^4. \tag{D.9}$$

At an endpoint angle, $f = 0$, so (D.8) is negative. At any other assigned vertex, its angle on the same arc differs from the corresponding endpoint angle by at least ε. The other endpoint lies on the opposite arc. Consequently,

$$f(\theta) \ge \frac{\varepsilon^2}{4}, \quad f(\theta) - \delta \ge \frac{\varepsilon^2}{8} > 0.$$

Here the first factor is at least $\varepsilon^2/4$ by $1 - \cos\psi \ge \psi^2/4$ for $|\psi| \le 1$, and the opposite-arc factor is at least one. The second inequality follows from $16v^4\varepsilon^4 \le \varepsilon^2/8$ for the chosen ε.

Since $\|A\| < 2$, all endpoint and nonendpoint squared distances differ from 1 by at least $g_0 = \frac{\varepsilon^4}{1000}$. Furthermore, the third coordinate of every edge-agent point exceeds $1/2$. Its numerator in $A/\|A\|$ is $\cos(\alpha + \beta - \pi)$, where $|\alpha + \beta - \pi| \le 6v\varepsilon$, and the denominator is at most $\sqrt{1 + 16v^2\varepsilon^2}$. Since $v\varepsilon \le 1/100$, the numerator is at least $1 - 18/10000$ and the denominator is at most $\sqrt{1 + 16/10000}$; their ratio exceeds $1/2$. This common positive third coordinate lets us move the remaining candidates a small distance along one axis and order their distances uniformly over all edge agents.

Let $\tau = g_0/100$. For U unselected candidates, where $U = B$ or $U = 1$ according to the problem, put

$$\xi_u = \tau\left(1 + \frac{u}{U+1}\right) \quad (1 \le u \le U), \quad \zeta_j = \tau\left(6 + \frac{j}{K+1}\right) \quad (1 \le j \le K).$$

Place the unselected candidates at $-\xi_u e_3$ and the additional selected candidates at $-\zeta_j e_3$, where e_3 is the third coordinate unit vector. The candidates are all distinct. For any edge-agent point p and parameter $\lambda \ge 0$,

$$\|p + \lambda e_3\|^2 = 1 + 2\lambda p_3 + \lambda^2, \quad \tfrac{1}{2} < p_3 \le 1.$$

For every edge-agent point p and all indices u, j, we have

$$1 < \|p + \xi_u e_3\|^2 \le 1 + 4\tau + 4\tau^2,$$
$$1 + 6\tau + 36\tau^2 \le \|p + \zeta_j e_3\|^2 \le 1 + 14\tau + 49\tau^2 < 1 + g_0.$$

Since $1 + 4\tau + 4\tau^2 < 1 + 6\tau + 36\tau^2$, these uniform bounds and (D.8) prove every part of the global ordering (D.4).

Finally, put $z = -100e_3$. Its distance to each additional selected candidate is smaller than its distance to every unselected candidate, since $\zeta_j > \xi_u$. All selected candidates have norm less than one, so $D < 2$, whereas the distance from z to every edge agent exceeds 99. This proves (D.5) and (D.6).

D.3 INTEGER COORDINATES

Only the normalizations $A/\|A\|$ may be irrational. The strict distance separations above allow them to be approximated without changing the reduction. To make this explicit, the gap between the unselected and additional-selected distance ranges is at least $\tau/2$: their squared-distance gap exceeds 2τ and the sum of the distances is less than four. The other consecutive ranges have larger gaps. Approximate each of the four coordinates of each edge-agent point by a dyadic rational to within $\tau/1000$. Its Euclidean displacement is at most $\tau/500$. A gap between two distance ranges decreases by at most twice this displacement, even when its two distances involve different edge-agent points, so all the required orderings remain strict. The inequalities involving z also remain valid. No common norm is needed at this stage.

The required precision uses $O(\log v)$ bits, since τ is an inverse polynomial in v. Square-root approximations can be computed by integer bisection. The rational candidate parameters have polynomial encoding length as well. Multiply every coordinate by a common positive denominator. Its bit length is at most the sum of the denominator bit lengths, hence polynomial in the input size. This produces integer coordinates in $\mathbb{R}^4$. Uniform scaling preserves every comparison used in the reduction.

D.4 A COMMON SPHERE IN FIVE DIMENSIONS

Start with the bounded rational construction, before clearing denominators. All points have norm at most 101. We use a scaled inverse stereographic projection, a standard rational parametrization of the sphere (Bahrdt & Seybold, 2017). For $L > 0$, define

$$\Psi_L(p) = \left(\frac{2L^2 p}{L^2 + \|p\|^2}, \quad L\frac{\|p\|^2 - L^2}{L^2 + \|p\|^2} \right) \in \mathbb{R}^5. \tag{D.10}$$

Its norm is L, its coordinates are rational when p and L are rational, and the map is injective. Direct calculation gives

$$\|\Psi_L(p) - \Psi_L(p')\| = \frac{2L^2 \|p - p'\|}{\sqrt{L^2 + \|p\|^2}\sqrt{L^2 + \|p'\|^2}}. \tag{D.11}$$

The multiplicative factor on $2\|p - p'\|$ lies between $L^2/(L^2 + 101^2)$ and 1. Since $\|p - p'\| \le 202$, the difference from $2\|p - p'\|$ is at most $4 \cdot 101^3/L^2$. Thus increasing L makes this map uniformly close to doubling all distances, while putting every point on the same sphere.

Choose an integer L such that this error is less than $\tau/1000$. The rational approximation step retained a gap of at least $\tau/4$ between consecutive distance ranges. Therefore (D.4) and (D.5) are preserved. The separation in (D.6) is much larger and is preserved as well. Such an L has polynomial encoding length. Clearing all coordinate denominators now gives integer points on one common sphere in $\mathbb{R}^5$, while injectivity preserves distinct candidates.

D.5 COMPLEXITY OF THE REDUCTION

The added graph has polynomial size. Since t is bounded by the original number of vertices, the numbers B, K, F, k, and n in (D.2) are polynomially bounded. All coordinate computations and bit lengths are polynomial. We have therefore reduced Balanced Biclique to the existence of an mPJR violation, and likewise to a specified-level mPJR+ violation, under each stated geometric restriction.

Both verification problems belong to coNP. A violating group, its level, and a relevant squared radius can be checked in polynomial time from integer coordinates. For mPJR, take the largest agent–candidate distance no greater than a violating radius. Such a distance exists because the group has a common candidate. The closed-ball candidate sets are unchanged, so the same group still violates mPJR. It therefore suffices to specify a pair of input points and use their squared distance. For specified-level mPJR+, provide the unselected candidate and compute the maximum squared distance to it. All required weight, cohesion, and selected-center comparisons are then exact. This completes the proof of Theorem 5.1.

E HARDNESS WITH TWO AGENT–CANDIDATE DISTANCES

Theorem 5.1 keeps the dimension constant but allows several agent–candidate distances. We give a complementary result: hardness also holds when there are only two such distances and all points lie on one sphere, at the cost of a dimension that grows with the instance. The constructions below realize arbitrary approval profiles; this stronger requirement explains why they need more dimensions than the reduction in Appendix D.

Spherical two-distance representations of graphs are well studied (Musin, 2019). Here the restriction applies only to agent–candidate distances; distances between agents or between candidates are unrestricted.

An approval profile consists of a set of approved candidates $\mathcal{A}_i \subseteq M$ for each agent i. Approval PJR requires $|X \cap \bigcup_{i\in S} \mathcal{A}_i| \geq \ell$ whenever $|S| \geq \ell n/k$ and $|\bigcap_{i\in S} \mathcal{A}_i| \geq \ell$. Its verification problem is coNP-complete (Aziz et al., 2018). Write $\Sigma \in \{-1,+1\}^{n\times m}$ for the profile's sign matrix, with $\Sigma_{ij} = +1$ when $c_j \in \mathcal{A}_i$. Let σ_i be row i and σ^j column j.

Theorem E.1 (Two-distance Euclidean hardness)**.** *Verifying mPJR is coNP-complete for integer points on one sphere in dimension* $\min\{m+1, n+3\}$*, even when agent–candidate distances take only two values. The same holds for verifying mPJR+ at an input-specified level. Moreover, some approval profiles require dimension at least* $\min\{n, m\} - 1$ *in every common-sphere realization with one distance for approvals and another for disapprovals.*

An $(m+1)$-dimensional realization. Let e_j denote the jth standard basis vector in $\mathbb{R}^m$. Place candidate c_j and agent i at

$$p(c_j) = (e_j, 2m), \quad p(i) = (2\sigma_i, 2m-1).$$

Both have squared norm $4m^2+1$. In the first m coordinates, the squared distance is $4(m-1)+1$ for an approved pair and $4(m-1)+9$ for a disapproved pair. The last coordinate adds one. Thus the two squared distances are $4m-2$ and $4m+6$. All coordinates are integers of $O(\log(m+1))$ bits, and the candidate points are distinct.

An $(n+3)$-dimensional realization. To obtain a bound in terms of the number of agents, we represent candidates by the columns of Σ. Distinct candidates may have identical columns, so two extra coordinates distinguish them without changing their distances to agents.

Set

$$L = \prod_{j=1}^{m}(1+j^2), \quad R = 2L, \quad u_j = \frac{R(1-j^2)}{1+j^2}, \quad v_j = \frac{2Rj}{1+j^2}.$$

The pairs (u_j, v_j) are distinct integer points on the circle of radius R. Define $V = 2n + R^2/2$ and place the candidates and agents at

$$p(c_j) = (2\sigma^j, u_j, v_j, V-1), \quad p(i) = (e_i, 0, 0, V),$$

where e_i is now a basis vector in $\mathbb{R}^n$. The candidate squared norm is $4n + R^2 + (V-1)^2$, equal to the agent squared norm $1+V^2$ because $2V = 4n+R^2$. The last three coordinates contribute R^2+1 to every agent–candidate squared distance. The first n coordinates contribute $4n-3$ for an approval and $4n+5$ for a disapproval. The two squared distances are therefore

$$4n + R^2 - 2 \quad \text{and} \quad 4n + R^2 + 6.$$

All points are integral, the candidate points are distinct, and all points lie on the same sphere. Since $\log L = O(m\log(m+1))$, every coordinate has polynomial bit length and the construction takes polynomial time.

Transferring the verification problems. Choose the smaller-dimensional realization, and let $a < b$ denote its two agent–candidate distances. For every agent i,

$$B(p(i), r) \cap p(M) = \begin{cases} \varnothing, & r < a, \\ p(\mathcal{A}_i), & a \leq r < b, \\ p(M), & r \geq b. \end{cases}$$

Thus an mPJR violation cannot occur below a, since there is no common candidate, or at or above b, since every selected candidate is reached. Between a and b, mPJR is exactly approval PJR. This proves coNP-hardness of mPJR verification.

For the specified-level mPJR+ problem, reduce from approval PJR+ verification at an input-specified level, which is coNP-complete (He et al., 2026). Fix an unselected candidate c. If all members of a group S approve c, then $r(c, S) = a$ and $X \cap A_a(S) = X \cap \bigcup_{i \in S} \mathcal{A}_i$. Otherwise $r(c, S) = b$, at which all selected centers are reached. Hence the approval and metric conditions are equivalent at every specified level. Membership in coNP follows by checking a violating group, its level, and the relevant squared radius or unselected candidate. This proves both complexity statements.

Why arbitrary two-distance realizations need growing dimension. Suppose an approval sign matrix Σ has such a realization in $\mathbb{R}^d$. Translate the common sphere to the origin, and let its squared radius be ρ^2. If each agent is at distance a from its approved candidates and at distance $b \neq a$ from the others, the matrix G of agent–candidate inner products satisfies

$$G = \left(\rho^2 - \frac{a^2 + b^2}{4}\right) J + \frac{b^2 - a^2}{4}\Sigma,$$

where J is the all-ones matrix. As an inner-product matrix of points in $\mathbb{R}^d$, G has rank at most d. The coefficient of Σ is nonzero, so

$$\operatorname{rank}(\Sigma) \leq \operatorname{rank}(G) + \operatorname{rank}(J) \leq d + 1.$$

For every n, m, there is a sign matrix of rank $\min\{n, m\}$. To see this, put a full-rank square sign matrix of that size in the profile: $2I - J$ works in every size except two, for which the 2×2 Hadamard matrix works. Fill any remaining entries arbitrarily. Thus some profiles require $d \geq \min\{n, m\} - 1$, proving the last part of Theorem E.1.

The upper and lower bounds differ by at most four. This comparison concerns arbitrary approval profiles with exactly two agent–candidate distances and a common sphere. It does not apply to the different, constant-dimensional construction used for Theorem 5.1.

F Experimental methods and additional results

We give implementation details, sampling procedures, checks, and full results for Section 6. The supplied files contain every location instance, selected candidate indices, per-instance losses, solver outcomes, and measured times. Tables are generated from these outputs rather than entered by hand.

F.1 Implementation and comparison methods

The flow optimizer. The implementation forms eligible blocks using two moving indices and processes nearest-neighbor queries together, as in Appendix B. It builds the network in Table 3. Identical fairness intervals retain the largest required count; identical loss intervals add their coefficients. These operations leave the feasible selections and objective unchanged. The primary implementation solves the resulting minimum-cost flow with HiGHS dual simplex through SciPy (Virtanen et al., 2020). An alternative implementation uses NetworkX network simplex (Hagberg et al., 2008).

All experimental coordinates, weights, losses, and opening costs are integers. After the numerical solver returns a flow, the implementation rounds its entries and checks every capacity and balance with integer arithmetic. It then uses integer shortest-path distances in the original residual network, including the forward directions of the originally uncapacitated arcs. The absence of a negative-cost cycle proves optimality of the checked flow. We additionally check that the recovered selection is binary, has size k, satisfies every fairness interval, and attains the exact objective equality (C.5). A run is not accepted unless all checks pass. This uses the recovery argument in Appendix C.4; it is not a claim that the numerical solver itself uses exact arithmetic.

The direct LP comparison uses the same HiGHS dual-simplex solver on the interval formulation in prefix variables y_j, with the shortage variables retained. This keeps each constraint sparse and avoids an unnecessarily dense baseline. We check the returned integral selection and evaluate its loss directly. Every completed LP run has the same optimum as the flow formulation. The reported software versions are Python 3.13.5, NumPy 2.3.5, SciPy 1.17.0, NetworkX 3.6.1, and scikit-learn 1.8.0.

Objective priorities. For $C_p(X) = \sum_i d(i, X)^p$, with $p \in \{1, 2\}$, the unconstrained solve omits the fairness arcs and returns X^0. The proportional solve minimizes $(k+1)C_p(X) + |X \setminus X^0|$. Since the loss is integer-valued, this selects X^* with minimum loss first and minimum replacements second. The repair solve minimizes

$$P|X \setminus X^0| + C_p(X), \qquad P = 1 + \sum_i \max_{c \in M} |p_i - c|^p.$$

The coefficient P exceeds the entire possible loss range, so X^{R} minimizes replacements first and loss second. Reported ratios use the unscaled loss, not these combined objectives. No further tie-breaking is imposed on the main X^0; Appendix F.4 examines other unconstrained optima.

SEAR. We implement the weighted expanding-radius rule in Lemma C.1. At each round, choose the smallest radius at which an unselected candidate has a quota of remaining weight in its ball. Among candidates attaining this radius, prefer the greatest remaining weight in the ball, then the lowest candidate index. Remove exactly one quota, considering agents by increasing distance from the chosen center, then increasing agent index. Multiplying all agent weights by k makes the quota W and permits integer arithmetic. SEAR does not use the loss and returns the same selection for both objectives on a location instance. Every returned selection is independently checked for mPJR+.

F.2 An optimal comparison under mPJR+

To distinguish the cost of the stronger condition from SEAR's choice of centers, we compute the minimum mPJR+ loss C^+. First solve the mPJR problem. If its optimum satisfies mPJR+, it is optimal for both conditions. Otherwise, use an integer formulation with the same loss and mPJR constraints. For a level-ℓ eligible block I and candidate c_j, the additional requirement is

$$\sum_{c_h \in A_{r(c_j, I)}(I)} x_h + (\ell - 1)x_j \geq \ell. \tag{F.1}$$

If c_j is unselected, this is exactly the required center count. If c_j is selected, it belongs to the reached interval, so the inequality holds automatically. Lemma B.1 shows that imposing these requirements for all eligible blocks and candidates is equivalent to mPJR+.

We add a requirement violated by the current selection, solve the integer model again, and repeat until the line verifier accepts the result. Each added inequality holds for every mPJR+ selection. Therefore, a minimum-loss solution of the current model that satisfies mPJR+ also has minimum loss over all mPJR+ selections. Prefix variables are integer and shortages are continuous; HiGHS is asked for zero relative optimality gap. Every run reported here ends with an optimal solver status and a selection passing the exact verifier. The time limit is 120 seconds for each full comparison, and no comparison reaches it. This is an exact experimental formulation, not a polynomial-time result.

In 299 of the 300 synthetic objective comparisons, the mPJR optimum already satisfies mPJR+. The only exception is paired trial 12 with 64 agents in $[0, 400]$ and $p = 1$: the optima are 10041 and 10096. For confidence scores, all 20 mPJR optima require changes to meet mPJR+; the integer comparison takes between two and nine solves. Table 4 gives the resulting cost differences and comparisons under the same condition.

F.3 Populations, scores, and variation across trials

Synthetic populations. All location sampling is uniform over the indicated integer interval. Agent draws allow repetitions; candidate draws within an interval are distinct. The uniform and concentrated instances use seed 220917: first the 25 uniform instances, then the 25 concentrated ones, with the counts and intervals in Section 6.

The paired study uses seed 381927. In each of 25 trials, draw 80 agents in $[0, 400]$ and 80 in $[4000, 10000]$, together with 20 candidates in the first interval and 40 in the second. For each $h \in \{16, 32, 48, 64\}$, take the first h agents drawn in $[0, 400]$ and the first $80 - h$ drawn in $[4000, 10000]$. Candidates and underlying draws are unchanged across the four populations. Thus the paired study contains 100 location instances, and both synthetic studies together contain 150. Each is evaluated with both losses. The paired comparisons are not independent samples and do not establish a general monotonicity property.

Table 4: Comparisons with the optimal mPJR+ cost: median [25th percentile, 75th percentile]. The last column counts equal mPJR and mPJR+ optima. Synthetic rows have 25 trials and confidence rows have 10.

Population	p	C^+/C^*	C^S/C^+	Equal optima
Uniform	1	1.000 [1.000, 1.000]	1.45 [1.27, 1.57]	25/25
Uniform	2	1.000 [1.000, 1.000]	2.37 [1.64, 2.78]	25/25
Concentrated	1	1.000 [1.000, 1.000]	1.59 [1.40, 1.86]	25/25
Concentrated	2	1.000 [1.000, 1.000]	2.94 [2.29, 4.06]	25/25
Paired: 20%	1	1.000 [1.000, 1.000]	1.38 [1.21, 1.49]	25/25
Paired: 20%	2	1.000 [1.000, 1.000]	1.85 [1.47, 2.62]	25/25
Paired: 40%	1	1.000 [1.000, 1.000]	1.45 [1.30, 1.57]	25/25
Paired: 40%	2	1.000 [1.000, 1.000]	2.11 [1.81, 2.88]	25/25
Paired: 60%	1	1.000 [1.000, 1.000]	1.51 [1.26, 1.72]	25/25
Paired: 60%	2	1.000 [1.000, 1.000]	2.40 [1.57, 2.96]	25/25
Paired: 80%	1	1.000 [1.000, 1.000]	1.47 [1.38, 1.71]	24/25
Paired: 80%	2	1.000 [1.000, 1.000]	2.53 [2.06, 2.91]	25/25
Confidence scores	1	1.006 [1.003, 1.015]	1.40 [1.34, 1.47]	0/10
Confidence scores	2	1.010 [1.004, 1.052]	2.08 [1.75, 2.40]	0/10

Table 5: Paired concentration study. Median cost ratios [25th percentile, 75th percentile], over 25 trials. The number of nonproportional unconstrained selections is the same for both losses.

Agents in $[0, 400]$	Nonproportional X^0	k-median: C^*/C^0	k-means: C^*/C^0
20%	0/25	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
40%	25/25	1.12 [1.07, 1.14]	1.45 [1.38, 1.54]
60%	25/25	1.29 [1.27, 1.32]	2.86 [2.66, 3.11]
80%	25/25	1.58 [1.48, 1.75]	5.59 [4.91, 6.44]

Confidence score selection. We use the 1797 handwritten-digit images supplied by `load_digits` in scikit-learn, a copy of the test portion of the UCI Optical Recognition of Handwritten Digits data (Alpaydin & Kaynak, 1998; Pedregosa et al., 2011). For seeds $17000, \ldots, 17009$, a stratified $70/30$ split yields 1257 training images and 540 held-out images. Standardization is fitted only on the training set. Logistic regression uses inverse regularization strength $C = 1$, the L-BFGS solver, at most 2000 iterations, and tolerance 10^{-8}. No run reports nonconvergence. The coordinate of a held-out image is its largest predicted class probability, rounded to four decimal places and multiplied by 10000. Uniform scaling changes absolute losses but not the selections, proportionality, or within-objective loss ratios.

Each held-out image is an agent. Candidates are all distinct rounded scores; if several images have the same score, the image with the smallest original dataset index is used as that candidate. This gives 244–276 candidates, from which we select $k = 10$. Coincident scores remain separate agents. We retain the train/test indices, unrounded probabilities, rounded scores, and candidate-to-image mappings. The trained classifier is used only to produce coordinates; neither clustering loss nor proportionality measures prediction accuracy. The ten splits overlap, so the results describe variation within this dataset rather than ten independent populations. The results concern these rounded instances; they give no approximation guarantee for unrounded scores.

Reported summaries. Ratios are computed within an instance before taking a median or quartile. Quartiles use linear interpolation at positions $(s-1)/4$ and $3(s-1)/4$ in a sorted sample of size s, indexed from zero. All loss denominators are positive. Tables 6 and 7 report quartiles. The full range can be wider: for example, the concentrated-data mPJR cost ratio ranges from 1.26 to 2.45 for k-median and from 3.28 to 11.67 for k-means. A median ratio therefore does not describe every instance.

Table 6: Loss ratios: median [25th percentile, 75th percentile]. "Paired" refers to the fixed-candidate concentration study. Synthetic rows have 25 trials and confidence rows have 10.

Population	p	C^*/C^0	C^{R}/C^*
Uniform	1	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
Uniform	2	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
Concentrated	1	1.65 [1.45, 1.79]	1.04 [1.03, 1.08]
Concentrated	2	5.83 [5.11, 6.44]	1.18 [1.11, 1.26]
Paired: 20%	1	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
Paired: 20%	2	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
Paired: 40%	1	1.12 [1.07, 1.14]	1.07 [1.05, 1.08]
Paired: 40%	2	1.45 [1.38, 1.54]	1.26 [1.17, 1.32]
Paired: 60%	1	1.29 [1.27, 1.32]	1.04 [1.02, 1.07]
Paired: 60%	2	2.86 [2.66, 3.11]	1.26 [1.15, 1.39]
Paired: 80%	1	1.58 [1.48, 1.75]	1.04 [1.02, 1.07]
Paired: 80%	2	5.59 [4.91, 6.44]	1.11 [1.08, 1.24]
Confidence scores	1	1.90 [1.84, 2.01]	1.08 [1.03, 1.16]
Confidence scores	2	8.09 [7.32, 8.87]	1.20 [1.15, 1.27]

Table 7: Replacement counts: median [25th percentile, 75th percentile]. r^{R} minimizes replacements first and loss second; r^* minimizes loss first and replacements second, relative to the same X^0.

Population	p	Replacements first: r^{R}	Loss first: r^*
Uniform	1	0 [0, 0]	0 [0, 0]
Uniform	2	0 [0, 0]	0 [0, 0]
Concentrated	1	2 [2, 2]	5 [4, 5]
Concentrated	2	3 [2, 3]	5 [5, 6]
Paired: 20%	1	0 [0, 0]	0 [0, 0]
Paired: 20%	2	0 [0, 0]	0 [0, 0]
Paired: 40%	1	1 [1, 1]	3 [3, 3]
Paired: 40%	2	1 [1, 1]	4 [3, 4]
Paired: 60%	1	1 [1, 2]	4 [3, 4]
Paired: 60%	2	2 [2, 2]	5 [4, 5]
Paired: 80%	1	2 [2, 2]	4 [4, 5]
Paired: 80%	2	3 [3, 3]	5 [5, 6]
Confidence scores	1	5 [4, 5]	10 [9, 10]
Confidence scores	2	5 [5, 6]	9.5 [9, 10]

F.4 Independent checks and alternative unconstrained optima

Checking the implementations. The test suite directly enumerates groups, representation levels, relevant radii, and selections on small instances, rather than using the interval description as its reference. It includes positive integer weights, repeated agent locations, tied distances, zero radii, selecting all candidates, negative opening costs, and both objective priorities. Batched nearest-neighbor queries are compared with direct sorting of distances. On random Manhattan-metric instances, the rebuilt and continued flow verifiers and the subset method are compared with direct group enumeration, including distance dilation. The mPJR+ integer formulation is compared with exhaustive selection, including an instance where its optimum differs from the mPJR optimum. The cost construction in Proposition 4.4 is checked for several values of T and both experimental losses. The saved test summary gives all counts; in total, the suite checks 2,782 selections, 2,160 flow optima, and 700 metric-verification comparisons. These finite checks complement the proofs; they do not replace them.

Dependence on the starting optimum. Replacement counts refer to a particular X^0. For each of the 100 objective comparisons on uniform and concentrated instances, we test four unconstrained choices: the original HiGHS solution, the NetworkX solution, and optima that prefer lower or higher candidate indices. For the latter two, an opening-cost preference is scaled so that it cannot change

the primary clustering optimum. We check that all four primary losses agree, then recompute both proportional priorities relative to each distinct starting selection. Table 8 reports when the replacement counts vary. The four methods produce different starting selections in 23 of the 100 comparisons. The minimum replacement count is unchanged in all comparisons; the loss-first count changes by one in two concentrated k-median instances. The least loss at the minimum replacement count changes in 11 concentrated k-median instances. This study examines four choices, not the full range over all unconstrained optima.

Table 8: Sensitivity to four choices of unconstrained optimum, over 25 instances per row. "Different X^0" counts trials with distinct starting selections. The other columns count trials in which the minimum replacement count r^{R} or the loss-first count r^* changes.

Population	p	Different X^0	Different r^{R}	Different r^*
Uniform	1	4	0	0
Uniform	2	0	0	0
Concentrated	1	19	0	2
Concentrated	2	0	0	0

F.5 Running time and memory

Exact line optimization and repair. We vary $n \in \{80, 160, 320, 640, 1280\}$ with $m = 3n/4$ and $k = 6$, using three trials per size. Agents are sampled with replacement and candidates without replacement from the integers in $[0, 10000]$, with seed $644710 + 100n +$ trial. Each instance also contains a uniformly sampled existing k-selection. We time mPJR-constrained k-median, pure minimum-change repair of that selection, and the full line verifier. Pure repair omits loss intervals, as in Corollary 4.3; it does not include secondary clustering loss.

Each method runs in a fresh process after the same untimed small solver warm-up. Elapsed time includes preprocessing, network or LP construction, solving, and result checks; imports and warm-up are excluded. Peak process memory, measured in MiB, includes the common imports. The flow and direct interval LP use the same HiGHS solver. The LP has a 30-second solver limit; incomplete runs are retained and are not averaged with completed runs. Tables 9 and 10 give times and memory, including preprocessing and recovery for the flow. All measurements were made sequentially on a Linux environment with an AMD EPYC 9V74 processor. The supplied metadata gives the software versions and thread environment; timings are specific to this implementation and machine.

Table 9: End-to-end time in seconds, median over three trials, with $m = 3n/4$ and $k = 6$. "Limit" means that all three direct LP runs reached the 30-second solver limit. Repair minimizes replacements without secondary loss. Peak memory is for the flow process and includes imports.

n	Flow	Interval LP	Repair	Verification	Flow MiB
80	0.021	0.021	0.005	0.0009	131
160	0.073	0.173	0.008	0.0014	139
320	0.276	1.85	0.014	0.0027	164
640	1.47	27.94	0.031	0.0057	270
1280	7.14	Limit	0.063	0.0111	671

A separate paired study fixes $(n, m) = (320, 240)$ and varies $k \in \{3, 6, 12, 24\}$. For each trial, the coordinates are reused across all four values of k; seeds are $681002 +$ trial. The existing selections are prefixes of a single random ordering of 24 candidates. Table 11 reports all three methods. These experiments separate the cost of full optimization, which constructs the loss intervals, from the smaller problems of checking or repairing a selection.

General-metric verification. We compare rebuilding a flow at every radius, continuing the current flow as radii increase, and enumerating subsets of selected centers. The latter uses the same subset enumeration as the exact verifier of He et al. (2026); our implementation evaluates $D_1(X)$ at all radii

Table 10: Flow size, median time components, and peak process memory over three trials. Time components are preprocessing, network solution with flow checks, and selection recovery with exact optimality checks. LP memory includes incomplete runs.

n	Flow arcs	Preprocess	Solve/check	Recover/check	Flow MiB	LP MiB
80	1,408	0.005	0.013	0.002	131	132
160	5,295	0.020	0.045	0.007	139	142
320	19,154	0.071	0.175	0.022	164	178
640	76,313	0.376	0.867	0.215	270	324
1280	289,608	1.64	4.98	0.474	671	809

Table 11: Varying k at $(n, m) = (320, 240)$, median seconds over three paired trials.

k	Flow	Repair	Verification
3	0.245	0.011	0.0013
6	0.279	0.015	0.0028
12	0.312	0.025	0.0055
24	0.339	0.039	0.0109

by summing agent weights over subsets with the standard zeta transform. All three methods compute $D_1(X)$ and the largest number of missing centers, without stopping at the first violation. Rebuilt and continued flow use the same integer implementation. Thus their comparison measures the effect of reusing the flow rather than a difference between solvers.

There are three paired trials with $n = m = 40$ and $k \in \{4, 8, 12, 16\}$. Agents are sampled from integer points in $[0, 1000]^2$; distinct candidates are sampled from the grid with spacing 20. Seeds are $711820, 711821, 711822$. Each trial reuses coordinates and takes nested selected-center sets as k increases. Euclidean comparisons use squared integer distances. Every method returns the same value of $D_1(X)$ and the same number of missing centers. Table 12 reports the times. The subset method can be competitive for small k, but its running time grows with 2^k, unlike that of the flow methods; these measurements are not a comparison with every available verifier.

Table 12: General-metric verification, including the largest number of missing centers; median seconds over three paired trials with $n = m = 40$. All methods return the same result and process all relevant radii.

k	Rebuilt flow	Continued flow	Subsets
4	0.236	0.060	0.038
8	0.370	0.102	0.051
12	0.474	0.137	0.108
16	0.521	0.161	0.776