# HAPPIER CONSTRAINED CLUSTERING: EXPLOITING PAIRWISE HINTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Constrained clustering incorporates domain knowledge through pairwise must-link and cannot-link constraints and can substantially improve unsupervised clustering. However, these constraints turn the assignment step of $k$-means into a combinatorial problem and, when treated only as feasibility conditions, leave much of their geometric information unexploited. We introduce HAPPIER, a weighted hard-constrained $k$-means framework that quantifies the geometric agreement between feasible assignments and their centers and uses this information to guide point weighting and solution selection. We establish that the fixed-center assignment relaxation is integral for two clusters under mere feasibility of the constraints, and for more than two clusters under acyclic constraint graphs. Consequently, each assignment step admits a globally optimal binary solution obtainable through linear programming, yielding a direct and tractable algorithm. Extensive experiments across diverse datasets, constraint budgets, and noise regimes demonstrate that HAPPIER robustly and consistently outperforms competitive constrained clustering methods in recovering the ground-truth cluster structure. These results confirm the value of exploiting the geometric information encoded by pairwise constraints to improve clustering quality.

## 1 INTRODUCTION

Clustering is a cornerstone of unsupervised learning, providing a fundamental mechanism for partitioning data based on internal similarity. By uncovering latent structures, clustering supports applications across a wide range of domains, including signal processing, computer vision, bioinformatics, natural language processing, and data mining (Ezugwu et al., 2022). The foundational $k$-means algorithm (Steinhaus, 1956; Forgy, 1965; MacQueen et al., 1967; Lloyd, 1982) remains one of the most widely used clustering methods due to its conceptual simplicity, computational efficiency, and broad applicability (Ikotun et al., 2023).

In many real-world applications, partial prior knowledge is available and can be exploited to guide the clustering process. This has led to the development of semi-supervised clustering, which incorporates limited supervision, most commonly through must-link and cannot-link constraints (Wagstaff and Cardie, 2000; Wagstaff et al., 2001). While these constraints guide partitions toward semantic relevance, they couple point assignments, transforming local decisions into a globally complex, intractable combinatorial problem. Consequently, current approaches trade off performance: efficient heuristics lack theoretical guarantees, while global optimization methods fail to scale.

This research addresses a critical gap in constrained clustering: the absence of a principled block coordinate descent method that directly extends the classical $k$-means framework to incorporate hard pairwise constraints. We introduce HAPPIER, which alternates between optimally solving the assignment and centroid update steps, thereby avoiding heuristic assignment procedures and the need for globally optimizing the joint problem. Beyond theoretical guarantees, this framework originates from a key empirical finding regarding centroid-based paradigms that has gone largely unaddressed in the constrained clustering literature: standard algorithms frequently satisfy constraints without *adapting* the underlying cluster geometry, forcing data points into non-nearest clusters. We formalize this diagnostic gap by defining explicit terms to capture geometric misalignments across both model outputs and ground-truth partitions, and design a mechanism to actively exploit it—resulting in remarkably improved clustering quality.



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.

Formally, we position our approach within the domain of *hard center-based clustering under hard pairwise constraints*, framed around the classical weighted $k$-means objective using the squared Euclidean distance metric. This problem is also known as semi-supervised weighted Minimum Sum-of-Squares Clustering (MSSC). Comprehensive overviews of the broader constrained clustering landscape are available in three recent reviews (González-Almagro et al., 2025; Yin et al., 2024; Bagirov et al., 2020). While our formulation focuses on hard constraints, we also compare against soft-constraint approaches in the numerical evaluation (Section 4) to assess the trade-off between constraint satisfaction and clustering quality. Under our setting, the number of clusters $K$ is fixed, and all pairwise constraints (in the form of must-link and cannot-link) must be strictly satisfied. The following section formally defines the mathematical framework for this optimization problem.

**Problem Formulation.** Let $\mathcal{A} = \{a_1, \ldots, a_N\} \subset \mathbb{R}^d$ denote a dataset of $N$ points to be partitioned into $K$ disjoint clusters, represented by a centroid matrix $C = [c_1, \ldots, c_K] \in \mathbb{R}^{d\times K}$. The cluster assignments are captured by the binary matrix $X \in \{0,1\}^{N\times K}$, where $x_{i,k} = 1$ if data point $a_i$ is assigned to cluster $k$, and 0 otherwise. Let $\mathcal{ML}, \mathcal{CL} \subseteq \{1, \ldots, N\} \times \{1, \ldots, N\}$ denote the sets of must-link and cannot-link pairs, respectively: $(i, j) \in \mathcal{ML}$ indicates that $a_i$ and $a_j$ are required to share a cluster, while $(i, j) \in \mathcal{CL}$ indicates that they are required to lie in different clusters. Finally, let $w_i > 0$, $i = 1, \ldots, N$, denote a fixed per-point weight, capturing the relative importance or reliability of each point. Formally, the weighted pairwise constrained $k$-means problem is defined by the following mixed-integer nonlinear program (MINLP):

$$(\mathcal{P}_0) \qquad \min_{\substack{X\in\mathcal{X}_{\text{bin}}\\ C\in\mathbb{R}^{d\times K}}} \mathcal{J}(X, C) = \sum_{i=1}^{N} w_i \sum_{k=1}^{K} x_{i,k}\|a_i - c_k\|^2,$$

where $\mathcal{X}_{\text{bin}} \subseteq \{0,1\}^{N\times K}$ is the binary feasible region of assignment matrices defined by:

$$\sum_{k=1}^{K} x_{i,k} = 1, \qquad \forall i \in \{1, \ldots, N\}, \qquad \text{(unique assignment)} \qquad (1)$$

$$x_{i,k} = x_{j,k}, \qquad \forall (i, j) \in \mathcal{ML},\ \forall k, \qquad \text{(must-link pairs)} \qquad (2)$$

$$x_{i,k} + x_{j,k} \le 1, \qquad \forall (i, j) \in \mathcal{CL},\ \forall k, \qquad \text{(cannot-link pairs)} \qquad (3)$$

$$x_{i,k} \in \{0, 1\}, \qquad \forall i \in \{1, \ldots, N\},\ k \in \{1, \ldots, K\}. \qquad \text{(integrality)} \qquad (4)$$

We address the continuous relaxation formulation obtained by relaxing the binary assignment constraints to the $K$-dimensional unit simplex; in Section 3, we establish that this solves $(\mathcal{P}_0)$ under geometric properties of the cannot-link hint structure. The resulting continuous feasible region $\mathcal{X}_{\text{con}}$ is obtained by replacing the integrality condition (4) in $\mathcal{X}_{\text{bin}}$ with non-negativity:

$$x_{i,k} \ge 0, \quad \forall i \in \{1, \ldots, N\},\ k \in \{1, \ldots, K\}. \qquad (5)$$

Substituting $\mathcal{X}_{\text{con}}$ for $\mathcal{X}_{\text{bin}}$ yields the continuous relaxed problem $(\mathcal{P}_1)$, which we address and analyze in this work:

$$(\mathcal{P}_1) \qquad \min_{\substack{X\in\mathcal{X}_{\text{con}}\\ C\in\mathbb{R}^{d\times K}}} \mathcal{J}(X, C).$$

We deliberately omit cluster non-emptiness constraints from these formulations. In center-based clustering, splitting a cluster into two smaller feasible subclusters cannot increase the objective value $\mathcal{J}(X, C)$; we formally detail how this extends to the constrained setting in Section 3.

**Related Work.** Within the setting considered here, optimization methods target either the full clustering problem or the fixed-center assignment step within an alternating scheme.

Global optimization methods jointly optimize assignments and centers, using approaches such as concave cutting planes (Xia, 2009) and semidefinite relaxations within branch-and-cut (Piccialli et al., 2022), though their applicability is restricted to small-scale settings.

Alternating methods improve scalability by decoupling assignments and centers, but rely on heuristics to solve the challenging constrained assignment subproblem. Among early approaches, the seminal COP-KMeans (Wagstaff et al., 2001) uses a greedy assignment heuristic that may fail even

on feasible instances. To enable a broader search, BCK-Means (Le et al., 2018) combines an LP relaxation with a feasibility pump to recover binary assignments, yet still lacks optimal assignment guarantees. Alternatively, PCCC (Baumann and Hochbaum, 2025) formulates the assignment step as an integer program, gaining scalability by restricting candidate clusters, which comes at the expense of assignment optimality. Beyond exactness, these methods often treat constraints as mere feasibility boundaries rather than informative geometric signals.

The HAPPIER framework addresses both limitations within the scalable alternating paradigm: we establish structural conditions under which the continuous assignment LP itself admits exact binary solutions, and devise a label-free geometric measure to exploit user hints guiding the optimization towards higher-quality partitions characterized by stronger spatial alignment.

**Contribution.** We briefly summarize our contributions as follows:

**1. Selecting Solutions Beyond Constraint Satisfaction.** We introduce unhappy-count as a label-free criterion for selecting among clustering solutions that satisfy the same hard pairwise constraints. HAPPIER uses this geometric signal to guide point weighting and solution selection, exploiting supervision beyond feasibility alone.

**2. Exact Assignment Through Linear Programming.** We establish structural conditions under which the fixed-center assignment relaxation is exact, allowing globally optimal binary assignments through linear programming. This provides a tractable optimization foundation for HAPPIER without rounding or greedy assignments.

**3. Improved Ground-Truth Recovery.** Extensive experiments demonstrate best or tied-best average recovery performance in most benchmark settings, with particularly large gains where the standard clustering objective is a poor recovery proxy. These results support geometric agreement as a useful criterion for selecting among constraint-satisfying solutions.

## 2 Constrained Clustering

Enforcing pairwise constraints can leave points assigned to non-nearest centers, revealing a tension between constraint satisfaction and geometric adaptation. We use this tension as a diagnostic signal to assess how a solution incorporates supervision and how the ground-truth partition aligns with nearest-center geometry. We formalize this geometric tension through the notion of point *happiness*—a point is happy if its closest center and its cluster's center coincide.

**Definition 1** (Unhappy Point)**.** *Given a configuration $(X, C)$, a point $a_i$ assigned to cluster $k$ is* unhappy *if its assigned center $c_k$ is not a nearest center to $a_i$: $k \notin \arg\min_{k'} \|a_i - c_{k'}\|^2$. An assignment $X$ is* happy *with respect to $C$ if no point $a_i$ is unhappy under configuration $(X, C)$.*

By construction, unconstrained $k$-means converges to a happy assignment, regardless of whether it aligns with the ground truth.

**Proposition 1** (Unconstrained $k$-means Converges to a Happy Assignment)**.** *If $(X^*, C^*)$ is a fixed point of the $k$-means iterative algorithm, then $X^*$ is a happy assignment with respect to $C^*$.*

While Definition 1 evaluates happiness with respect to an assignment-center configuration $(X, C)$, the concept also provides a fundamental lens for analyzing dataset geometry itself. Measuring ground-truth happiness explicitly captures how well a dataset's class structure fits the geometric assumptions of centroid-based models.

**Definition 2** (Ground-Truth Centers and Happy Points)**.** *Given a ground-truth assignment matrix $X^{\mathrm{gt}}$, the ground-truth center $c_k^{\mathrm{gt}}$ of cluster $k$ is defined as: $c_k^{\mathrm{gt}} = \frac{\sum_{i=1}^{N} x_{i,k}^{\mathrm{gt}} a_i}{\sum_{i=1}^{N} x_{i,k}^{\mathrm{gt}}}$. A point $a_i$ is* happy *with respect to ground-truth centers $C^{\mathrm{gt}}$ if its ground-truth center is a nearest center among all centers in $C^{\mathrm{gt}}$: $k \in \arg\min_{k'} \|a_i - c_{k'}^{\mathrm{gt}}\|^2$. Otherwise, $a_i$ is a* ground-truth unhappy point.

**Definition 3** (Dataset Unhappy-Count / GT-Floor)**.** *The dataset's* ground-truth unhappy-count *(or* GT-floor*), denoted by $\mathcal{U}^{\mathrm{gt}}$, is the number of data points that are unhappy with respect to the ground-truth centers $C^{\mathrm{gt}}$.*

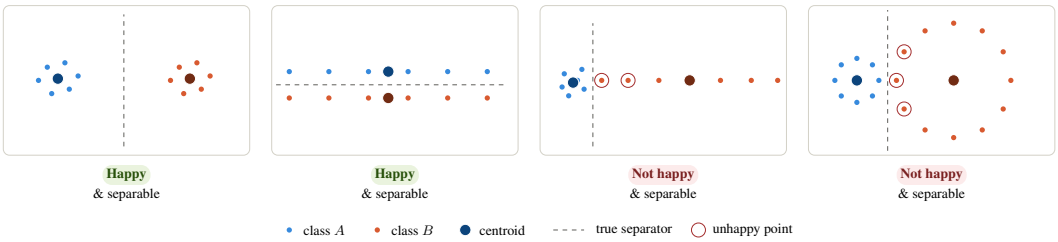



Figure 1: Distinguishing happy vs. separable datasets. A dataset is considered happy if, in the ground-truth partition, every data point is closest to its own class centroid.

**Definition 4** (Dataset Happiness Rate)**.** *The dataset's* ground-truth happiness rate*, denoted by* $\mathcal{H}^{\mathrm{gt}} \in [0, 1]$*, is the proportion of points that are happy with respect to their ground-truth partition:* $\mathcal{H}^{\mathrm{gt}} = 1 - \frac{\mathcal{U}^{\mathrm{gt}}}{N}$*. A dataset is termed a* happy dataset *if* $\mathcal{H}^{\mathrm{gt}} = 1$ *(or equivalently,* $\mathcal{U}^{\mathrm{gt}} = 0$*).*

Dataset happiness is a distinct—and strictly stronger—geometric property than linear separability. As illustrated in Figure 1, a dataset can be perfectly separable yet unhappy ($\mathcal{H}^{\mathrm{gt}} < 1$) due to asymmetric class densities or variances.

**Ground-Truth Geometry and Unexploited Supervision.** Interpreting unhappiness as unexploited supervision requires accounting for the geometry of the true partition. When $\mathcal{H}^{\mathrm{gt}} < 1$, that partition itself contains unhappy points and therefore cannot be a fixed point of unconstrained $k$-means (Proposition 1). Pairwise constraints can support such assignments, but their resulting unhappiness may reflect the target geometry rather than insufficient adaptation. Thus, unhappiness can indicate unexploited supervision when $\mathcal{H}^{\mathrm{gt}} \approx 1$, whereas forcing happiness when the true partition is unhappy may hinder recovery. Understanding when reducing unhappiness is guaranteed to improve recovery remains an important direction for future research.

## 3 Method

In Algorithm 1, we present HAPPIER, which, for fixed geometry-guided weights, alternates between assignment optimization and centroid updates, with exact binary assignments under the structural conditions established below. We begin by reducing the assignment problem to a smaller representative formulation.

**Handling Must-Link Constraints via Domain Reduction.** Must-link constraints are satisfied *by construction* as follows. The set $\mathcal{ML}$ induces an equivalence relation whose transitive closure partitions the dataset into $R$ disjoint equivalence classes $\mathcal{G}_1, \ldots, \mathcal{G}_R$. For each class $\mathcal{G}_r$, a single representative $r \in \mathcal{R} = \{r_1, \ldots, r_R\}$ is defined as $r = \min\{i : i \in \mathcal{G}_r\}$. Enforcing $x_{i,k} = x_{r,k}$ for all $i \in \mathcal{G}_r$ guarantees must-link satisfaction, collapsing every group to a single decision variable. Any cannot-link constraint involving a member of a must-link group is correspondingly propagated to its representative, yielding the reduced cannot-link set $\hat{\mathcal{CL}}$ over the representatives. Substituting this equality into $(\mathcal{P}_1)$ and grouping terms by representatives gives the reduced problem $(\hat{\mathcal{P}}_1)$:

$$(\hat{\mathcal{P}}_1) \qquad \min_{\substack{X \in \hat{\mathcal{X}}_{\mathrm{con}} \\ C \in \mathbb{R}^{d \times K}}} \hat{\mathcal{J}}(X; C) = \sum_{r \in \mathcal{R}} \sum_{k=1}^{K} x_{r,k} \sum_{i \in \mathcal{G}_r} w_i \|a_i - c_k\|^2.$$

where $\hat{\mathcal{X}}_{\mathrm{con}}$ is the continuous feasible region of assignment matrices defined below.

**The Continuous Assignment Relaxation.** After reducing the must-link groups to their representatives, we consider the continuous relaxation of the representative assignment subproblem. For fixed centroids $C$, its objective is linear in $X$ over the continuous feasible region $\hat{\mathcal{X}}_{\mathrm{con}}$, yielding the

following *LP relaxation* of the assignment subproblem:

$$\min_{X \in \hat{\mathcal{X}}_{\text{con}}} \hat{\mathcal{J}}(X, C) = \sum_{r \in \mathcal{R}} \sum_{k=1}^{K} x_{r,k} \left( \sum_{i \in \mathcal{G}_r} w_i \|a_i - c_k\|^2 \right).$$

$$\hat{\mathcal{X}}_{\text{con}} = \Bigg\{ X \in \mathbb{R}^{R \times K} \ : \sum_{k=1}^{K} x_{r,k} = 1, \quad \forall r \in \mathcal{R}, \\ x_{r,k} \geq 0, \quad \forall r \in \mathcal{R}, \ \forall k, \\ x_{r,k} + x_{j,k} \leq 1, \quad \forall (r,j) \in \hat{\mathcal{CL}}, \ \forall k \Bigg\}.$$

**Theoretical Guarantees: Exact Integrality.** The following two theorems establish the conditions under which the linear relaxation of the assignment subproblem is exact.

**Assumption 1** (Constraint Feasibility)**.** *The provided sets of pairwise hints $\mathcal{ML}$ and $\mathcal{CL}$ are* feasible*: there exists at least one binary assignment $X \in \{0,1\}^{N \times K}$ that partitions the $N$ data points into $K$ clusters while respecting every must-link and cannot-link constraint.*

**Lemma 1** (Bipartite Graph, $K = 2$)**.** *Under Assumption 1, if $K = 2$, then the cannot-link graph $G$ is bipartite.*

*Proof of Lemma 1.* Recall that a graph is bipartite if and only if it contains no odd cycle. If cannot-link constraints induce an odd cycle, then at least 3 clusters are needed, which contradicts Assumption 1 that guarantees that a valid assignment exists using only $K = 2$ clusters. □

**Theorem 1** (Integrality of the Assignment Polyhedron, $K = 2$)**.** *Under Assumption 1, if $K = 2$ then all extreme points of $\hat{\mathcal{X}}_{\text{con}}$ are integral.*

*Proof of Theorem 1.* By Lemma 1, the reduced cannot-link graph $G$ is bipartite. Set $y_r := x_{r,1}$, so that $x_{r,2} = 1 - y_r$ and $0 \leq y_r \leq 1$. For each edge $(i,j) \in \hat{\mathcal{CL}}$, the cannot-link inequalities for the two clusters become $y_i + y_j \leq 1$ and $y_i + y_j \geq 1$, hence $y_i + y_j = 1$. For each connected component $G_m$ of $G$ with bipartition $(L_m, R_m)$, these equalities imply $y_i = t_m$ for all $i \in L_m$ and $y_j = 1 - t_m$ for all $j \in R_m$, where $t_m \in [0,1]$. An isolated vertex likewise contributes one independent parameter in $[0,1]$. Conversely, every choice of these component parameters defines a feasible assignment. At an extreme point, every $t_m$ must belong to $\{0,1\}$: otherwise, perturbing that parameter slightly in both directions yields two distinct feasible assignments whose midpoint is the original assignment. Thus all $y_r$ are binary, and so are $x_{r,1} = y_r$ and $x_{r,2} = 1 - y_r$. □

For $K > 2$, we interpret cluster labels as vertex colors and apply the tree-integrality result of Delle Donne and Marenco (2016) for the standard vertex-coloring formulation.

**Lemma 2** (Delle Donne and Marenco, 2016)**.** *Let $T = (V, E)$ be a tree and let $S$ be a finite set of colors with $|S| \geq 2$. Then $\{x \geq 0 : \sum_{k \in S} x_{v,k} = 1 \ \forall v \in V, \quad x_{v,k} + x_{w,k} \leq 1 \ \forall (v,w) \in E, \ \forall k \in S\} = P_{\text{col}}(T, S)$, where $P_{\text{col}}(T, S)$ is the convex hull of the binary assignments representing proper colorings of $T$ using colors from $S$. In particular, this relaxation is integral.*

**Theorem 2** (Integrality of the Assignment Polyhedron, $K > 2$)**.** *Under Assumption 1, if $K > 2$ and the reduced cannot-link graph $G$ is acyclic, then all extreme points of $\hat{\mathcal{X}}_{\text{con}}$ are integral.*

*Proof of Theorem 2.* Each connected component $T_m = (V_m, E_m)$ of $G$ is a tree. Its assignment polyhedron is

$$Q_m = \left\{ x \geq 0 : \sum_{k=1}^{K} x_{r,k} = 1 \ \forall r \in V_m, \quad x_{i,k} + x_{j,k} \leq 1 \ \forall (i,j) \in E_m, \ \forall k \in \{1, \ldots, K\} \right\}.$$

By Lemma 2, $Q_m$ is integral; for an isolated vertex, it is simply a simplex. Since distinct components involve disjoint variables and no constraints couple them, $\hat{\mathcal{X}}_{\text{con}}$ is the Cartesian product of the component polyhedra. Its extreme points are therefore products of integral component extreme points and are themselves integral. □

**Corollary 1** (Exact LP Relaxation)**.** *Under the conditions of Theorem 1 or Theorem 2, the fixed-center assignment LP over $\hat{\mathcal{X}}_{\text{con}}$ admits an optimal extreme point, and every such point is binary. Consequently, a globally optimal binary assignment can be obtained in polynomial time through linear programming.*

**Comment on Acyclicity.** For $K > 2$, Theorem 2 requires the reduced cannot-link graph $G$ to be acyclic. Typical constraint sampling regimes (González-Almagro et al., 2025), including those considered in Section 4, provide sparse supervision, making the formation of cycles less likely, although sparsity alone does not guarantee acyclicity. When hints are sampled from labeled points, cycles can be explicitly avoided in the reduced graph, allowing the forest condition to be enforced during constraint generation. Consequently, acyclicity is not expected to pose a practical limitation in common constrained clustering settings with sparse supervision.

**The BCD Optimization Framework.** Algorithm 1 integrates these components into a block coordinate descent scheme: an exact, LP-based assignment step, followed by a centroid update. The instance weights $w_i$ can either be set uniformly to 1 for a parameter-free algorithm or tuned as detailed below. In the rare event that an empty cluster arises at the end of an iteration, reassigning any singleton representative or a full must-link group to the empty cluster preserves feasibility without increasing $\hat{\mathcal{J}}(X, C)$. Since the destination cluster is empty, no cannot-link constraint can be violated, and setting its centroid to the weighted mean of the moved points does not increase the overall objective.

**Algorithm 1** The *HAPPIER* algorithm for hard-constrained clustering via exact LP relaxation

**Input:** $\{a_i\}_{i=1}^N$, $\{w_i\}_{i=1}^N$, $K$, $\mathcal{ML}$, $\mathcal{CL}$.
**Preprocessing:**
1. Collapse $\mathcal{ML}$ into $R$ equivalence classes $\{\mathcal{G}_r\}_{r=1}^R$ via union-find.
2. Assign representative set $\mathcal{R}$ and define the reduced cannot-link set $\hat{\mathcal{CL}}$.

**Initialize** cluster centers $\{c_k^{(0)}\}_{k=1}^K$; set iteration $t \leftarrow 0$.
**Repeat:**
1. *Optimal Assignment Step (X-Step):* Fix centroids $C^{(t)}$, solve the LP relaxation with deterministic tie-breaks:

$$X^{(t+1)} = \arg\min_{X \in \hat{\mathcal{X}}_{\text{con}}} \hat{\mathcal{J}}(X; C^{(t)})$$

2. *Optimal Centroid Step (C-Step):* Update centroids for each cluster $k$:

$$c_k^{(t+1)} = \frac{\sum_{r=1}^R x_{r,k}^{(t+1)} \sum_{i \in \mathcal{G}_r} w_i a_i}{\sum_{r=1}^R x_{r,k}^{(t+1)} \sum_{i \in \mathcal{G}_r} w_i}$$

3. $t \leftarrow t + 1$.

**Until** convergence ($X^{(t)} = X^{(t-1)}$).
**Output:** Binary assignment matrix $X$ and final cluster centroids $C$.

**Theorem 3** (Convergence of HAPPIER)**.** *Suppose that $(\mathcal{P}_0)$ is feasible, the weights are fixed and positive, and the conditions of Theorem 1 or Theorem 2 hold. With exact subproblem solutions and the stopping rule specified in Algorithm 1, HAPPIER generates feasible iterates with non-increasing objective values and terminates after finitely many iterations. Its output $(X^\star, C^\star)$ satisfies*

$$X^\star \in \arg\min_{X \in \hat{\mathcal{X}}_{\text{con}}} \hat{\mathcal{J}}(X, C^\star), \qquad C^\star \in \arg\min_{C \in \mathbb{R}^{d \times K}} \hat{\mathcal{J}}(X^\star, C).$$

*In particular, the output is a block-optimal feasible solution of $(\mathcal{P}_0)$.*

*Proof.* The proof follows the standard finite-convergence argument for Lloyd's $k$-means algorithm. The X-Step and C-Step are exact minimization steps, with deterministic tie-breaking in the X-Step. Thus, unless the stopping criterion is met, the assignment changes and the objective decreases strictly. Since there are at most $K^R$ possible assignments of the $R$ must-link representatives, only finitely many such decreases can occur. Upon termination, both subproblem minimizations hold, yielding a feasible solution that is block-optimal. □

**Beyond Satisfaction: Exploiting Pairwise Hints.** To better exploit the provided supervision, we incorporate the unhappy-count as a geometric guiding metric for parameter tuning. While Algorithm 1 ensures feasibility in its parameter-free setting $\{w_i = 1\}_{i=1}^N$, it may still converge to a partition with a high unhappy-count, satisfying constraints without adapting the underlying geometry. HAPPIER($\alpha, \beta$) addresses this by tuning two weights to explicitly target a lower unhappy-count at convergence; a simple grid search over $(\alpha, \beta)$ identifies the combination that reduces it most. After propagating the cannot-link constraints to the must-link groups, points receive additional weight $\alpha$ if involved in any must-link constraint and additional weight $\beta$ if their must-link group participates in a cannot-link constraint:

$$w_i = 1 + \alpha \cdot \mathbb{I}[i \text{ appears in } \mathcal{ML}] + \beta \cdot \mathbb{I}[\text{the group } r \text{ representing } i \text{ appears in } \hat{\mathcal{CL}}], \qquad w_i \geq 1.$$

The weights are fixed after preprocessing and remain unchanged throughout the run; consequently, all instances within the same must-link group receive the same weight. These weights function as an attention mechanism, forcing the algorithm to "listen" more closely to the hints. Tuning $(\alpha, \beta)$ reshapes the optimization landscape to target fixed points with lower unhappy-counts, ultimately yielding superior clustering quality as demonstrated in Section 4.

# 4 Numerical Experiments

## 4.1 Experimental Protocol

We evaluate HAPPIER across two tiers: (i) a 20-dataset benchmark suite establishing robustness across diverse data regimes, and (ii) a scarcity imbalanced showcase demonstrating that a happy dataset is not necessarily an easy one. Both share the following common protocol.

**Constraint Sampling and Random Initialization.** Following common practice (González-Almagro et al., 2025), for a dataset with $N$ data points, we sample a fixed percentage of $N$ instance pairs uniformly at random from the set of all possible pairs, labeling each pair according to ground-truth agreement. Thus, for example, a dataset with 100 data points and a 10% constraint rate yields 10 constraints in total. For $K > 2$, the random samples resulted in acyclic cannot-link graphs in all experiments, satisfying the acyclicity assumption of our method, while the random samples used for noisy constraints with $K = 2$ were confirmed feasible. To isolate performance from variability due to constraint sampling and random initialization, we report the mean $\pm$ std across 100 total runs per setup, combining 10 independent constraint draws with 10 random initialization seeds (each using $K$ uniformly sampled data points as initial centers).

**Competing Methods.** We compare HAPPIER against unconstrained $k$-means, three soft-constraint baselines (PCKMeans, LCVQE, PCSK-Means), and two hard-constraint baselines (COP-KMeans, BCK-Means); Table 3 in Appendix A summarizes each baseline's type and mechanism. Although HAPPIER enforces constraints as hard requirements, we include soft-constraint methods since they provide a complementary approach that trades off constraint satisfaction against the clustering objective, which can be useful when constraints are noisy or strict feasibility is not demanded. We test two variants of our approach: the parameter-free HAPPIER$(0, 0)$ and HAPPIER$(\alpha, \beta)$—where the weights are selected separately for each setup by evaluating a prescribed grid of $(\alpha, \beta)$ values and retaining the pair with the lowest final unhappy-count. HAPPIER requires no specialized implementation or problem-specific heuristics; its results follow directly from the simple block coordinate procedure in Algorithm 1.

**Quality Metrics.** Every experiment tracks three quantities. **ARI** (Adjusted Rand Index) measures agreement between the predicted partition and the ground-truth labels, chance-corrected, with 1 indicating perfect recovery. **MSSC** is the $k$-means objective $\mathcal{J}(X, C)$ from $(\mathcal{P}_0)$, evaluated on the final assignment with uniform weights $\{w_i = 1\}_{i=1}^N$. **Unhappy-count** is the number of points in the output partition that are not nearest to their assigned unweighted center.

## 4.2 Cross-Dataset Comparison Suite

To test HAPPIER beyond any single hand-picked domain, we assembled a suite of 20 widely used benchmark datasets from the constrained clustering literature. The suite spans categorical, binary, continuous, sparse, high-dimensional, data-scarce ($D \gg N$), and class-imbalanced regimes, with $K$

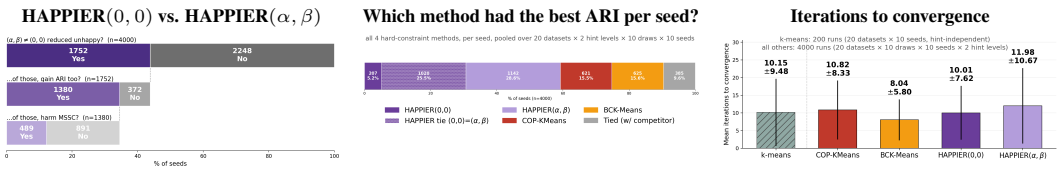



Figure 2: Per-seed diagnostics across the 20-dataset comparison suite.

ranging from 2 to 10. The characteristics of all datasets are summarized in Table 4 in Appendix B. For each dataset, we report the GT-floor (Definition 3) and the corresponding Happy% $= 100\,(1 -$ GT-floor$/N)$. The datasets are ordered by decreasing Happy%, from Soybean (100%) to Balance Scale (75%).

**Results.** Table 5 in Appendix B reports ARI at two hint budgets (10% and 20% of $N$) for all 8 methods. These results indicate that reducing unhappy-count can improve ARI even when the dataset is not happy. HAPPIER$(\alpha, \beta)$ attains the best or tied-best ARI in 33 of 40 rows; the margin over the strongest hard-constraint competitor exceeds 0.1 ARI on 9 of those rows (highlighted green). The method loses to a competitor on only 3 datasets. Figure 2 presents this analysis at the per-seed level: in the left panel, reducing unhappy-count translates to an ARI gain in 78.8% of seeds where it occurs (1380/1752). HAPPIER$(\alpha, \beta)$ achieves the best ARI on more individual seeds than any competitor (middle panel). The right panel shows that at a fixed $(\alpha, \beta)$, HAPPIER converges in roughly the same number of alternating iterations as unconstrained $k$-means or any hard-constraint competitor. However, selecting $(\alpha, \beta)$ requires evaluating multiple such runs across the grid, so the real cost of tuning shows up in the number of runs performed (allowing users to scale search granularity to match their compute budget). Table 6 in Appendix B gives a zoom-in per-dataset breakdown behind these ARI numbers—showing unhappy-count and MSSC for all 8 methods at the 20% hint budget.

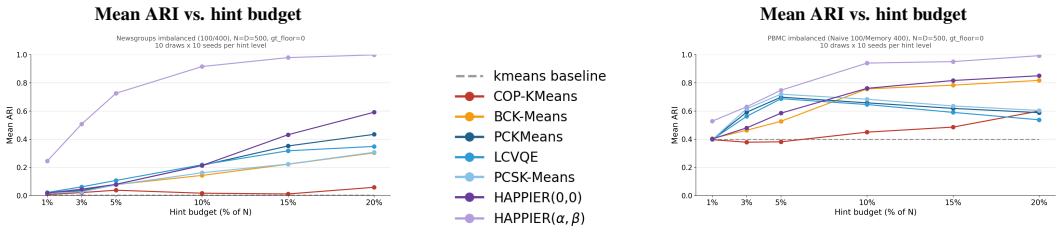
### 4.3 Happy Datasets Are Not Necessarily Easy

Dataset happiness characterizes the ground-truth partition as a fixed point of $k$-means, but says nothing about its basin of attraction or the presence of competing, lower-quality fixed points. Three structural features—imbalance, scarcity, and sparsity—can therefore make a happy dataset difficult to recover. In such settings, MSSC can be poorly correlated with ARI, so minimizing the objective may improve clustering loss while degrading recovery. Consequently, the standard strategy of selecting the lowest-MSSC solution across random $k$-means restarts can fail. We show that HAPPIER$(\alpha, \beta)$ can overcome this difficulty by exploiting the available hints, often enabling near-perfect recovery across random initializations when sufficient supervision is available.

In this section, we focus on two confirmed-happy datasets with $K = 2$ (one text and one genomic) that exemplify the $N \approx D$ scarcity regime. From the 20 Newsgroups dataset (Lang, 1995), we use the `talk.politics.guns` and `sci.med` topics, represented as sparse binary bag-of-words vectors, while the PBMC dataset (10x Genomics FACS-sorted T cells) (Zheng et al., 2017) consists of dense, continuous single-cell gene expression profiles after z-scoring. These datasets illustrate that even when the ground-truth partition aligns with nearest-center geometry, it can still challenge the $k$-means framework. All experiments follow the protocol described in Section 4.1; Table 7 in Appendix C details the class composition of both datasets and their preprocessing.

**ARI Across Hint Budgets.** Figure 3 compares mean ARI across hint budgets (1% to 20% of $N$) on both datasets. HAPPIER$(\alpha, \beta)$ consistently outperforms all competitors, reaching near-perfect recovery at 20% hints. Figure 5 in Appendix C aggregates seed-level results across both datasets, showing that reducing the unhappy-count improves ARI in 92.2% of seeds (559/606). Moreover, in 73.2% of these cases (409/559), the ARI improvement actually increases MSSC.

**Zoom-In: Diagnostics Beyond ARI.** Tables 1 and 2 detail all metrics for the 10%-hint benchmark, including the number of violated must-link and cannot-link hints. While all hard-constraint baselines satisfy the constraints, HAPPIER$(\alpha, \beta)$ additionally reduces the unhappy-count, yielding better alignment with the ground-truth clustering.




**(a) Text (Newsgroups), imbalanced 100/400, sparse binary.**

**(b) Genomic (PBMC), imbalanced 100/400, dense continuous.**

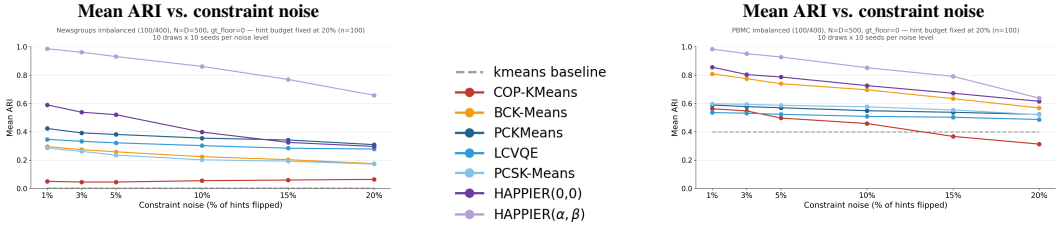
Figure 3: Scarce & imbalanced regime: ARI across hint budgets.

| | Method | ARI | Unhappy | MSSC | Iters | ML-viol | CL-viol |
|---|---|---|---|---|---|---|---|
| | $k$-means | 0.003±0.040 | — | 15269.8±248.9 | 5.8±3.5 | 3.8±2.9 | 13.6±2.2 |
| Soft | PCKMeans | 0.214±0.207 | 21.5±10.5 | 15400.4±118.5 | 99.0±9.8 | 3.2±4.4 | 9.9±3.5 |
| Soft | LCVQE | 0.219±0.235 | 23.0±10.2 | 15419.5±122.9 | 100.0±0.0 | 6.3±4.5 | 13.7±2.7 |
| Soft | PCSK-Means | 0.162±0.187 | 17.5±6.2 | 15358.8±126.8 | 99.0±9.8 | 0.0±0.0 | 11.8±3.6 |
| Hard | COP-KMeans | 0.017±0.071 | 17.0±3.0 | 15411.4±132.6 | 7.1±2.9 | 0.0±0.0 | 0.0±0.0 |
| Hard | BCK-Means | 0.144±0.206 | 19.5±6.2 | 15355.3±135.5 | 7.5±2.4 | 0.0±0.0 | 0.0±0.0 |
| Hard | HAPPIER(0,0) | 0.214±0.241 | 14.8±4.8 | 15338.3±134.0 | 7.9±4.2 | 0.0±0.0 | 0.0±0.0 |
| Hard | HAPPIER($\alpha, \beta$) | **0.916±0.229** | **2.1±4.9** | 15398.9±35.1 | 7.8±3.3 | 0.0±0.0 | 0.0±0.0 |

Table 1: Text (Newsgroups), imbalanced 100/400, sparse binary, 10% hints.

| | Method | ARI | Unhappy | MSSC | Iters | ML-viol | CL-viol |
|---|---|---|---|---|---|---|---|
| | $k$-means | 0.399±0.489 | — | 246254.7±1940.8 | 5.4±3.8 | 4.8±7.2 | 7.0±6.5 |
| Soft | PCKMeans | 0.658±0.348 | 17.4±6.5 | 245638.8±371.8 | 100.0±0.0 | 4.3±5.2 | 10.1±3.2 |
| Soft | LCVQE | 0.646±0.339 | 19.2±6.3 | 245712.7±337.5 | 100.0±0.0 | 6.1±6.0 | 11.4±3.6 |
| Soft | PCSK-Means | 0.684±0.357 | 14.4±4.8 | 245476.6±342.0 | 100.0±0.0 | 0.0±0.0 | 12.0±3.7 |
| Hard | COP-KMeans | 0.450±0.490 | 9.8±9.7 | 245600.3±1143.9 | 12.5±8.0 | 0.0±0.0 | 0.0±0.0 |
| Hard | BCK-Means | 0.757±0.428 | 5.0±9.1 | 244969.1±758.1 | 7.5±3.8 | 0.0±0.0 | 0.0±0.0 |
| Hard | HAPPIER(0,0) | 0.761±0.423 | 4.2±8.0 | 244956.8±742.0 | 5.6±3.0 | 0.0±0.0 | 0.0±0.0 |
| Hard | HAPPIER($\alpha, \beta$) | **0.940±0.225** | **0.7±3.5** | 244702.6±553.7 | 8.0±7.3 | 0.0±0.0 | 0.0±0.0 |

Table 2: Genomic (PBMC), imbalanced 100/400, dense continuous, 10% hints.




**(a) Text (Newsgroups), imbalanced 100/400, sparse binary, 20% hints.**

**(b) Genomic (PBMC), imbalanced 100/400, dense continuous, 20% hints.**

Figure 4: Robustness to constraint noise.

**Robustness to Noisy Constraints.** All experiments above assume that the pairwise oracle is perfectly reliable, meaning every must-link and cannot-link hint agrees with the ground-truth partition. However, real-world annotators are rarely so dependable. This motivates us to test how HAPPIER and its competitors degrade when a fraction of the must-link/cannot-link hints are flipped. We examine whether the exact satisfaction of hard constraints, an asset when the oracle is trustworthy, becomes a liability under noisy supervision. Remarkably, as shown in Figure 4, HAPPIER$(\alpha, \beta)$ outperforms all competitors across all noise levels on both datasets, exhibiting graceful degradation under substantial constraint noise.

## 5 Conclusion and Future Work

In this work, we introduced HAPPIER, a theoretically grounded and scalable framework for constrained clustering that leverages information from pairwise must-link and cannot-link supervision. The framework employs a novel, label-free "unhappy" diagnostic as a guiding metric for parameter tuning, using the constraints to direct the search toward feasible solutions that make greater geometric sense. Our numerical results demonstrate consistent performance improvements, including in settings with noisy constraints and unhappy ground-truth structures. Looking ahead, future research should focus on three main directions: developing a principled theoretical framework for weight selection, extending the methodology to broader objective formulations, and incorporating richer or alternative constraint structures.

## THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this work, we used generative AI tools (specifically ChatGPT, Gemini, and Claude) for formulating and refining mathematical claims (notation and conditions), designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting datasets, and interpreting results. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, assisting with translation, or supporting qualitative and thematic data analysis.

Additionally, we used generative AI tools for editing the paper to improve readability, formatting references, brainstorming, identifying and searching for relevant literature, creating or editing software code, and creating or modifying scientific figures.

All LLM-generated code was executed, debugged, and tested by the authors. All literature suggestions, mathematical notation edits, and analytical observations were manually reviewed to identify potential hallucinations, verify citation accuracy, and verify logical correctness. The authors take full responsibility for all final content, claims, and artifacts produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

All experiments in this paper are fully reproducible: every $k$-means initialization and constraint draw uses a fixed, documented seed (Section 4.1), and re-running the pipeline reproduces every reported ARI, unhappy-count, and MSSC value exactly across all 20 comparison-suite datasets and both scarcity-regime case studies. The proof of exact integrality of the LP relaxation is given in Section 3. All datasets are public, with exact access routes and DOIs cited per dataset in Table 4.

## REFERENCES

Stefan Aeberhard, Danny Coomans, and Olivier De Vel. Comparison of classifiers in high dimensional settings. In *Dept. Math. Statist., James Cook Univ., North Queensland, Australia, Tech. Rep*, number 92-02, 1992.

Ethem Alpaydin and Cenk Kaynak. Cascading classifiers. *Kybernetika*, 34(4):369–374, 1998. Source of the Optical Recognition of Handwritten Digits data used by `sklearn.load_digits`.

Adil M Bagirov, Napsu Karmitsa, and Sona Taheri. Partitional clustering via nonsmooth optimization. *Cham, Switzerland: Springer Nature*, 2020.

Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM)*, pages 333–344. SIAM, 2004.

Philipp Baumann and Dorit S. Hochbaum. An algorithm for clustering with confidence-based must-link and cannot-link constraints. *INFORMS Journal on Computing*, 37(4):1044–1068, 2025. doi: 10.1287/ijoc.2023.0419.

Arindam Bhattacharjee, William G Richards, Jane Staunton, Cheng Li, Stefano Monti, Priya Vasa, Christine Ladd, Javad Beheshti, Raphael Bueno, Michael Gillette, et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98(24):13790–13795, 2001.

Diego Delle Donne and Javier Marenco. Polyhedral studies of vertex coloring problems: The standard formulation. *Discrete Optimization*, 21:1–13, 2016. doi: 10.1016/j.disopt.2016.05.001.

Absalom E. Ezugwu, Abiodun M. Ikotun, Olaide O. Oyelade, Laith Abualigah, Jeffrey O. Agushaka, Christopher I. Eke, and Andronicus A. Akinyelu. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743, 2022.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936.

Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

R. Forsyth. Zoo [dataset]. UCI Machine Learning Repository, 1990.

Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

Germán González-Almagro, Daniel Peralta, Eli De Poorter, José-Ramón Cano, and Salvador García. Semi-supervised constrained clustering: An in-depth overview, ranked taxonomy and future research directions. *Artificial Intelligence Review*, 58(5):157, 2025.

hepatitis. Hepatitis [dataset]. UCI Machine Learning Repository, 1983.

Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023.

N. Ilter and H. Guvenir. Dermatology [dataset]. UCI Machine Learning Repository, 1998.

imageseg. Statlog (image segmentation) [dataset]. UCI Machine Learning Repository, 1990.

A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. Heart disease [dataset]. UCI Machine Learning Repository, 1989.

Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*, pages 331–339, 1995.

Huu M. Le, Anders Eriksson, Thanh-Toan Do, and Michael Milford. A binary optimization approach for constrained k-means clustering. In *Asian Conference on Computer Vision (ACCV)*, 2018. arXiv:1810.10134.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Computing Surveys*, 50(6):1–45, 2018.

Stuart P Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. University of California Press, 1967.

R. Michalski. Soybean (small) [dataset]. UCI Machine Learning Repository, 1980.

mushroom. Mushroom [dataset]. UCI Machine Learning Repository, 1981.

K. Nakai. Ecoli [dataset]. UCI Machine Learning Repository, 1996.

Kenta Nakai and Minoru Kanehisa. Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins: Structure, Function, and Bioinformatics*, 11(2):95–110, 1991.

Michiel O Noordewier, Geoffrey G Towell, and Jude W Shavlik. Training knowledge-based neural networks to recognize genes in DNA sequences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 3, 1991.

Mojca Olave, Vladislav Rajkovič, and Marko Bohanec. An application for admission in public school systems. In *Expert Systems in Public Administration*, pages 145–160. Elsevier, 1989.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Dan Pelleg and Dorit Baras. K-means with large and noisy constraint sets. In *European Conference on Machine Learning*, pages 674–682. Springer, 2007.

Veronica Piccialli, Anna Russo Russo, and Antonio M. Sudoso. An exact algorithm for semi-supervised minimum sum-of-squares clustering. *Computers & Operations Research*, 147:105958, 2022. doi: 10.1016/j.cor.2022.105958.

V. Rajkovic. Nursery [dataset]. UCI Machine Learning Repository, 1989.

R. Siegler. Balance scale [dataset]. UCI Machine Learning Repository, 1976.

V. Sigillito, S. Wing, L. Hutton, and K. Baker. Ionosphere [dataset]. UCI Machine Learning Repository, 1989.

splice. Molecular biology (splice-junction gene sequences) [dataset]. UCI Machine Learning Repository, 1991.

Hugo Steinhaus. Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences, Classe III*, 4(12):801–804, 1956.

Wei Tang, Hui Xiong, Shi Zhong, and J. Wu. Enhancing semi-supervised clustering: a feature projection perspective. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 707–716, 2007.

voting. Congressional voting records [dataset]. UCI Machine Learning Repository, 1987.

Kiri Wagstaff and Claire Cardie. Clustering with instance-level constraints. *AAAI/IAAI*, 1097(577-584):197, 2000.

Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584, 2001.

William H Wolberg, W Nick Street, and Olvi L Mangasarian. Breast cancer Wisconsin (diagnostic) data set. *UCI Machine Learning Repository*, 1995.

Yu Xia. A global optimization method for semi-supervised clustering. *Data mining and knowledge discovery*, 18(2):214–256, 2009.

Hui Yin, Amir Aryani, Stephen Petrie, Aishwarya Nambissan, Aland Astudillo, and Shengyuan Cao. A rapid review of clustering algorithms. *arXiv preprint arXiv:2401.07389*, 2024.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, 2017.

M. Zwitter and M. Soklic. Lymphography [dataset]. UCI Machine Learning Repository, 1988.

# A COMPETING METHODS

Table 3 summarizes the type and mechanism of each baseline used throughout Section 4, first introduced in Section 4.1.

Table 3: Overview of competing methods.

| Method | Type | Mechanism |
|---|---|---|
| **PCKMeans** (Basu et al., 2004) | Soft | Constant penalty per violated constraint. |
| **LCVQE** (Pelleg and Baras, 2007) | Soft | Violation penalty scaled by inter-point distance. |
| **PCSK-Means** (Tang et al., 2007) | Hybrid | Hard ML via union-find chunklets; soft constant CL penalties. |
| **COP-KMeans** (Wagstaff et al., 2001) | Hard | Greedy assignment; falls back to nearest centroid on infeasibility. |
| **BCK-Means** (Le et al., 2018) | Hard | LP relaxation with feasibility-pump binary projection. |

# B CROSS-DATASET COMPARISON SUITE: DATASETS AND FULL RESULTS

Section 4.2's experimental suite is detailed across three tables: Table 4 summarizes the metadata for all 20 benchmark datasets, Table 5 presents the complete per-dataset ARI performance across the two hint budgets (10% and 20%), while Table 6 provides a detailed diagnostic breakdown (unhappy-count and MSSC) at 20% hint budget.

Table 4: Datasets metadata.

| # | Dataset | $K$ | $N$ | $D$ | GT floor | Happy% | Class sizes | Characteristics | Source / preprocessing |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Soybean | 4 | 47 | 35 | 0 (0.0%) | 100.0% | 17 / 10 / 10 / 10 | Categorical (one-hot → binary), imbalanced | UCI #91, one-hot (Michalski, 1980) |
| 2 | Dermatology | 6 | 366 | 129 | 6 (1.6%) | 98.4% | 112 / 72 / 61 / 52 / 49 / 20 | Categorical (one-hot → binary), imbalanced | UCI #33, one-hot (Ilter and Guvenir, 1998) |
| 3 | Wine | 3 | 178 | 13 | 4 (2.2%) | 97.8% | 71 / 59 / 48 | — | `sklearn` + StandardScaler (Aeberhard et al., 1992; Pedregosa et al., 2011) |
| 4 | Lung | 5 | 203 | 3312 | 9 (4.4%) | 95.6% | 139 / 21 / 20 / 17 / 6 | Scarce ($D \gg N$), imbalanced | Bhattacharjee (2001) microarray + StandardScaler (Bhattacharjee et al., 2001; Li et al., 2018) |
| 5 | Splice | 3 | 3190 | 287 | 210 (6.6%) | 93.4% | 767 / 768 / 1655 | Categorical (one-hot → binary), sparse (∼79.1%) | UCI #69, one-hot (splice; Noordewier et al., 1991) |
| 6 | Breast Cancer | 2 | 569 | 30 | 39 (6.9%) | 93.1% | 357 / 212 | Imbalanced | `sklearn` + StandardScaler (Wolberg et al., 1995; Pedregosa et al., 2011) |
| 7 | Leukemia | 2 | 72 | 7129 | 5 (6.9%) | 93.1% | 47 / 25 | Scarce ($D \gg N$), imbalanced | Golub (1999) microarray + StandardScaler (Golub et al., 1999; Li et al., 2018) |
| 8 | Iris | 3 | 150 | 4 | 11 (7.3%) | 92.7% | 50 / 50 / 50 | — | `sklearn` (Fisher, 1936; Pedregosa et al., 2011) |
| 9 | Voting | 2 | 435 | 16 | 44 (10.1%) | 89.9% | 267 / 168 | Binary, imbalanced | UCI #105, mode imputation (voting) |
| 10 | Digits | 10 | 1797 | 64 | 183 (10.2%) | 89.8% | ≈180 per class | — | `sklearn` + StandardScaler (Alpaydin and Kaynak, 1998; Pedregosa et al., 2011) |
| 11 | Mushroom | 2 | 8124 | 116 | 900 (11.1%) | 88.9% | 4208 / 3916 | Categorical (one-hot → binary), sparse (∼81%) | UCI #73, one-hot (mushroom) |
| 12 | Lymphography | 4 | 148 | 55 | 20 (13.5%) | 86.5% | 81 / 61 / 4 / 2 | Categorical (one-hot → binary), imbalanced, sparse (∼67.3%) | UCI #63, one-hot (Zwitter and Soklic, 1988) |
| 13 | Zoo | 7 | 101 | 16 | 14 (13.9%) | 86.1% | 41 / 20 / 13 / 10 / 8 / 5 / 4 | Binary, imbalanced | UCI #111 (Forsyth, 1990) |
| 14 | Image Segmentation | 7 | 2310 | 18 | 347 (15.0%) | 85.0% | 330 per class | — | UCI #147 + StandardScaler (imageseg) |
| 15 | Heart | 2 | 303 | 13 | 47 (15.5%) | 84.5% | 164 / 139 | — | UCI #45, target binarized, mean imputation + StandardScaler (Janosi et al., 1989) |
| 16 | Ecoli | 8 | 336 | 7 | 53 (15.8%) | 84.2% | 143 / 77 / 52 / 35 / 20 / 5 / 2 / 2 | Imbalanced | UCI #39 (Nakai, 1996; Nakai and Kanehisa, 1991) |
| 17 | Nursery | 4 | 12958 | 27 | 2078 (16.0%) | 84.0% | 4320 / 4266 / 4044 / 328 | Categorical (one-hot → binary), imbalanced, sparse (∼70.4%) | UCI #76, 4-class subset (Rajkovic, 1989; Olave et al., 1989) |
| 18 | Hepatitis | 2 | 155 | 19 | 25 (16.1%) | 83.9% | 123 / 32 | Imbalanced, sparse (∼64.2%) | UCI #46, mode/mean imputation + StandardScaler (hepatitis) |
| 19 | Ionosphere | 2 | 351 | 34 | 61 (17.4%) | 82.6% | 225 / 126 | Imbalanced | UCI #52 + StandardScaler (Sigillito et al., 1989) |
| 20 | Balance Scale | 3 | 625 | 4 | 156 (25.0%) | 75.0% | 288 / 288 / 49 | Imbalanced | UCI #12 (Siegler, 1976) |

Table 5: ARI summary at hint budgets 10% and 20% of $N$, mean±std pooled over 10 draws×10 seeds (100 runs/cell).

| # | Dataset | Hints | Unconstrained | Soft-CC | | | Hard-CC | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $k$-means | PCKMeans | LCVQE | PCSK-Means | COP-KMeans | BCK-Means | HAPPIER(0,0) | HAPPIER($\alpha, \beta$) |
| 1 | Soybean | 10% | 0.543±0.110 | 0.532±0.114 | 0.522±0.112 | 0.569±0.119 | 0.555±0.124 | 0.535±0.136 | 0.564±0.114 | **0.608±0.156** |
| | | 20% | | 0.559±0.160 | 0.535±0.148 | 0.612±0.147 | 0.557±0.116 | 0.550±0.137 | 0.580±0.136 | **0.694±0.185** |
| 2 | Dermatology | 10% | 0.693±0.136 | 0.664±0.125 | 0.669±0.122 | 0.697±0.119 | 0.701±0.137 | 0.620±0.154 | 0.698±0.124 | **0.719±0.115** |
| | | 20% | | 0.653±0.107 | 0.655±0.104 | 0.705±0.110 | 0.705±0.140 | 0.637±0.138 | 0.699±0.117 | **0.731±0.114** |
| 3 | Wine | 10% | 0.840±0.174 | 0.821±0.070 | 0.811±0.073 | 0.888±0.028 | 0.892±0.025 | 0.840±0.137 | **0.903±0.019** | **0.903±0.019** |
| | | 20% | | 0.764±0.113 | 0.734±0.113 | 0.878±0.045 | 0.894±0.029 | 0.880±0.088 | **0.915±0.026** | **0.915±0.026** |
| 4 | Lung | 10% | 0.412±0.107 | 0.344±0.092 | 0.342±0.088 | 0.447±0.106 | 0.416±0.097 | 0.445±0.097 | 0.448±0.112 | **0.487±0.139** |
| | | 20% | | 0.336±0.122 | 0.319±0.105 | 0.456±0.109 | 0.389±0.097 | 0.478±0.104 | 0.462±0.123 | **0.546±0.144** |
| 5 | Splice | 10% | 0.488±0.090 | 0.400±0.099 | 0.401±0.095 | 0.580±0.112 | 0.547±0.122 | 0.324±0.155 | 0.620±0.127 | **0.678±0.074** |
| | | 20% | | 0.330±0.091 | 0.316±0.087 | 0.612±0.116 | 0.624±0.091 | 0.369±0.165 | 0.687±0.143 | **0.761±0.084** |
| 6 | Breast Cancer | 10% | 0.664±0.010 | 0.577±0.087 | 0.545±0.081 | 0.662±0.036 | 0.648±0.014 | **0.714±0.021** | **0.714±0.019** | **0.714±0.020** |
| | | 20% | | 0.502±0.145 | 0.455±0.131 | 0.645±0.061 | 0.650±0.025 | **0.765±0.020** | **0.765±0.021** | **0.765±0.021** |
| 7 | Leukemia | 10% | 0.097±0.088 | 0.117±0.109 | 0.144±0.108 | 0.120±0.101 | **0.115±0.114** | 0.102±0.096 | 0.083±0.096 | 0.099±0.108 |
| | | 20% | | 0.090±0.111 | 0.097±0.101 | 0.107±0.107 | 0.126±0.122 | 0.120±0.122 | 0.124±0.131 | **0.139±0.134** |
| 8 | Iris | 10% | 0.663±0.117 | 0.690±0.084 | 0.705±0.038 | 0.705±0.082 | 0.727±0.052 | 0.612±0.137 | 0.737±0.064 | **0.750±0.022** |
| | | 20% | | 0.688±0.074 | 0.682±0.057 | 0.732±0.047 | 0.737±0.027 | 0.669±0.112 | 0.767±0.043 | **0.771±0.022** |
| 9 | Voting | 10% | 0.576±0.003 | 0.529±0.079 | 0.482±0.065 | 0.595±0.032 | 0.580±0.021 | 0.594±0.043 | **0.620±0.025** | **0.620±0.027** |
| | | 20% | | 0.480±0.131 | 0.418±0.108 | 0.606±0.066 | 0.590±0.025 | 0.643±0.047 | **0.670±0.029** | 0.666±0.033 |
| 10 | Digits | 10% | 0.513±0.048 | 0.508±0.045 | 0.508±0.045 | 0.525±0.046 | 0.517±0.043 | 0.516±0.047 | 0.521±0.047 | **0.536±0.044** |
| | | 20% | | 0.507±0.043 | 0.506±0.042 | 0.533±0.044 | 0.517±0.046 | 0.523±0.049 | 0.535±0.047 | **0.553±0.044** |
| 11 | Mushroom | 10% | 0.195±0.226 | 0.227±0.242 | 0.219±0.233 | 0.261±0.255 | 0.324±0.276 | 0.240±0.272 | 0.419±0.264 | **0.649±0.004** |
| | | 20% | | 0.235±0.239 | 0.206±0.212 | 0.301±0.255 | 0.321±0.285 | 0.275±0.279 | **0.689±0.006** | **0.689±0.006** |
| 12 | Lymphography | 10% | 0.179±0.053 | 0.182±0.037 | 0.182±0.036 | 0.206±0.042 | 0.204±0.045 | 0.181±0.052 | 0.208±0.042 | **0.213±0.041** |
| | | 20% | | 0.187±0.037 | 0.179±0.041 | 0.234±0.048 | 0.216±0.052 | 0.208±0.060 | 0.238±0.053 | **0.240±0.050** |
| 13 | Zoo | 10% | 0.590±0.111 | 0.597±0.108 | 0.594±0.109 | 0.624±0.101 | **0.640±0.090** | 0.584±0.138 | 0.582±0.096 | 0.613±0.102 |
| | | 20% | | 0.603±0.104 | 0.598±0.100 | 0.643±0.099 | **0.651±0.093** | 0.598±0.136 | 0.591±0.100 | 0.625±0.097 |
| 14 | Image-Seg | 10% | 0.466±0.040 | 0.459±0.044 | 0.459±0.045 | 0.469±0.045 | 0.468±0.041 | 0.467±0.046 | 0.471±0.042 | **0.487±0.043** |
| | | 20% | | 0.450±0.045 | 0.448±0.043 | 0.469±0.045 | 0.467±0.040 | 0.466±0.048 | 0.473±0.042 | **0.503±0.052** |
| 15 | Heart | 10% | 0.433±0.017 | 0.381±0.048 | 0.370±0.045 | 0.438±0.038 | 0.442±0.034 | 0.445±0.046 | 0.487±0.035 | **0.488±0.035** |
| | | 20% | | 0.354±0.058 | 0.322±0.058 | 0.425±0.056 | 0.434±0.056 | 0.493±0.047 | **0.536±0.043** | 0.534±0.039 |
| 16 | Ecoli | 10% | 0.412±0.094 | 0.397±0.068 | 0.418±0.072 | 0.416±0.061 | 0.404±0.058 | 0.338±0.083 | 0.418±0.069 | **0.421±0.069** |
| | | 20% | | 0.385±0.058 | 0.417±0.056 | 0.424±0.052 | 0.397±0.058 | 0.346±0.086 | 0.427±0.061 | **0.438±0.074** |
| 17 | Nursery | 10% | 0.035±0.031 | 0.047±0.049 | 0.045±0.047 | 0.062±0.064 | 0.037±0.038 | 0.037±0.016 | 0.071±0.070 | **0.433±0.103** |
| | | 20% | | 0.059±0.056 | 0.055±0.054 | 0.085±0.081 | 0.065±0.064 | 0.039±0.017 | 0.101±0.090 | **0.467±0.053** |
| 18 | Hepatitis | 10% | 0.230±0.109 | 0.223±0.080 | 0.207±0.070 | 0.247±0.084 | 0.247±0.118 | 0.246±0.138 | 0.316±0.086 | **0.329±0.101** |
| | | 20% | | 0.206±0.102 | 0.172±0.075 | 0.259±0.094 | 0.307±0.088 | 0.336±0.137 | 0.392±0.091 | **0.406±0.079** |
| 19 | Ionosphere | 10% | 0.168±0.001 | 0.143±0.038 | 0.135±0.028 | 0.145±0.023 | **0.188±0.015** | 0.149±0.019 | 0.155±0.017 | 0.156±0.016 |
| | | 20% | | 0.125±0.054 | 0.100±0.040 | 0.123±0.041 | **0.215±0.021** | 0.157±0.029 | 0.155±0.026 | 0.166±0.032 |
| 20 | Balance Scale | 10% | 0.139±0.061 | 0.160±0.068 | 0.148±0.063 | 0.193±0.067 | 0.157±0.054 | 0.171±0.095 | 0.204±0.066 | **0.273±0.080** |
| | | 20% | | 0.188±0.086 | 0.172±0.083 | 0.264±0.080 | 0.195±0.058 | 0.203±0.106 | 0.302±0.077 | **0.379±0.062** |

Table 6: Full per-dataset diagnostics at 20% hint budget (mean±std over 10 draws×10 seeds).

| **Soybean** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .54±.11 | — | 255.1±62.3 |
| PCKMeans | .56±.16 | 2.0±1.8 | 261.9±63.7 |
| LCVQE | .54±.15 | 2.4±2.0 | 271.3±62.1 |
| PCSK-M | .61±.15 | 1.1±1.0 | 240.8±51.4 |
| COP-KM | .56±.12 | 1.2±1.0 | 256.8±65.0 |
| BCK-M | .55±.14 | 2.0±1.4 | 269.3±67.9 |
| H(0,0) | .58±.14 | 1.1±1.0 | 253.7±61.3 |
| H($\alpha$, $\beta$) | **.69±.18** | 0.2±0.5 | 232.4±39.6 |

| **Dermatology** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .69±.14 | — | 3.6±0.1k |
| PCKMeans | .65±.11 | 15.5±6.2 | 3.7±0.1k |
| LCVQE | .66±.10 | 15.7±6.5 | 3.7±0.1k |
| PCSK-M | .70±.11 | 7.0±4.0 | 3.6±0.1k |
| COP-KM | .70±.14 | 5.4±3.1 | 3.6±0.1k |
| BCK-M | .64±.14 | 18.8±8.0 | 3.7±0.1k |
| H(0,0) | .70±.12 | 5.1±2.7 | 3.6±0.1k |
| H($\alpha$, $\beta$) | **.73±.11** | 3.8±2.0 | 3.6±0.1k |

| **Wine** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .84±.17 | — | 1.3±0.1k |
| PCKMeans | .76±.11 | 8.5±6.2 | 1.4±0.1k |
| LCVQE | .73±.11 | 10.2±6.7 | 1.4±0.1k |
| PCSK-M | .88±.04 | 3.2±2.9 | 1.3±0.0k |
| COP-KM | .89±.03 | 1.1±1.1 | 1.3±0.0k |
| BCK-M | .88±.09 | 2.7±2.1 | 1.3±0.1k |
| H(0,0) | **.92±.03** | 0.8±0.7 | 1.3±0.0k |
| H($\alpha$, $\beta$) | **.92±.03** | 0.8±0.7 | 1.3±0.0k |

| **Lung** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .41±.11 | — | 509.7±5.1k |
| PCKMeans | .34±.12 | 15.9±5.2 | 528.7±14.8k |
| LCVQE | .32±.10 | 16.5±5.3 | 529.0±12.7k |
| PCSK-M | .46±.11 | 9.4±2.9 | 513.4±4.5k |
| COP-KM | .39±.10 | 9.8±2.7 | 517.2±5.8k |
| BCK-M | .48±.10 | 8.9±2.8 | 513.6±4.7k |
| H(0,0) | .46±.12 | 8.8±2.8 | 513.7±4.5k |
| H($\alpha$, $\beta$) | **.55±.14** | 6.2±2.4 | 513.3±5.0k |

| **Splice** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .49±.09 | — | 137.7±0.1k |
| PCKMeans | .33±.09 | 349.2±32.1 | 138.7±0.1k |
| LCVQE | .32±.09 | 362.5±31.0 | 138.7±0.1k |
| PCSK-M | .61±.12 | 142.7±27.8 | 138.0±0.2k |
| COP-KM | .62±.09 | 93.8±23.9 | 137.9±0.2k |
| BCK-M | .37±.17 | 430.5±125.4 | 138.4±0.6k |
| H(0,0) | .69±.14 | 89.4±36.0 | 137.8±0.2k |
| H($\alpha$, $\beta$) | **.76±.08** | 72.3±18.9 | 137.8±0.1k |

| **Breast Cancer** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .66±.01 | — | 11.6±0.0k |
| PCKMeans | .50±.14 | 55.6±33.1 | 13.1±1.0k |
| LCVQE | .46±.13 | 63.9±33.2 | 13.5±1.1k |
| PCSK-M | .65±.06 | 33.1±13.6 | 12.4±0.4k |
| COP-KM | .65±.03 | 15.8±2.1 | 12.1±0.1k |
| BCK-M | **.77±.02** | 16.6±2.6 | 11.8±0.0k |
| H(0,0) | **.77±.02** | 14.5±2.6 | 11.8±0.0k |
| H($\alpha$, $\beta$) | **.77±.02** | 14.5±2.6 | 11.8±0.0k |

| **Leukemia** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .10±.09 | — | 469.4±13.2k |
| PCKMeans | .09±.11 | 5.6±2.7 | 475.3±9.3k |
| LCVQE | .10±.10 | 6.2±3.2 | 475.3±9.2k |
| PCSK-M | .11±.11 | 6.0±2.6 | 474.9±7.8k |
| COP-KM | .13±.12 | 4.5±1.5 | 476.5±8.1k |
| BCK-M | .12±.12 | 4.9±1.6 | 474.3±7.4k |
| H(0,0) | .12±.13 | 4.7±1.6 | 474.5±7.6k |
| H($\alpha$, $\beta$) | **.14±.13** | 4.2±1.5 | 474.8±8.2k |

| **Iris** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .66±.12 | — | 91.6±25.6 |
| PCKMeans | .69±.07 | 6.9±3.1 | 98.1±16.1 |
| LCVQE | .68±.06 | 9.6±4.1 | 101.3±15.7 |
| PCSK-M | .73±.05 | 5.7±2.7 | 87.8±9.9 |
| COP-KM | .74±.03 | 3.0±0.8 | 84.2±3.7 |
| BCK-M | .67±.11 | 10.4±4.2 | 99.3±26.3 |
| H(0,0) | .77±.04 | 2.9±0.9 | 82.5±9.2 |
| H($\alpha$, $\beta$) | **.77±.02** | 2.9±0.8 | 81.6±1.1 |

| **Voting** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .58±.00 | — | 990.9±0.0 |
| PCKMeans | .48±.13 | 34.2±17.8 | 1.1±0.1k |
| LCVQE | .42±.11 | 38.2±19.5 | 1.2±0.1k |
| PCSK-M | .61±.07 | 22.4±8.5 | 1.1±0.0k |
| COP-KM | .59±.03 | 15.8±4.4 | 1.1±0.0k |
| BCK-M | .64±.05 | 24.5±3.7 | 1.0±0.0k |
| H(0,0) | **.67±.03** | 15.5±3.1 | 1.0±0.0k |
| H($\alpha$, $\beta$) | .67±.03 | 15.3±3.4 | 1.0±0.0k |

| **Digits** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .51±.05 | — | 71.0±1.2k |
| PCKMeans | .51±.04 | 48.2±11.3 | 72.1±1.0k |
| LCVQE | .51±.04 | 48.5±11.6 | 72.2±1.0k |
| PCSK-M | .53±.04 | 27.8±5.8 | 71.3±0.9k |
| COP-KM | .52±.05 | 31.4±5.5 | 71.7±1.0k |
| BCK-M | .52±.05 | 45.1±7.2 | 71.5±1.0k |
| H(0,0) | .54±.05 | 29.7±5.0 | 71.3±0.8k |
| H($\alpha$, $\beta$) | **.55±.04** | 28.0±4.7 | 71.0±0.8k |

| **Mushroom** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .20±.23 | — | 79.3±1.5k |
| PCKMeans | .23±.24 | 666.7±225.1 | 82.3±2.2k |
| LCVQE | .21±.21 | 666.3±203.0 | 82.4±1.9k |
| PCSK-M | .30±.26 | 618.4±209.0 | 81.7±2.2k |
| COP-KM | .32±.28 | 454.1±197.7 | 81.3±2.7k |
| BCK-M | .28±.28 | 702.6±297.9 | 81.9±2.5k |
| H(0,0) | **.69±.01** | 264.1±11.5 | 78.4±0.1k |
| H($\alpha$, $\beta$) | **.69±.01** | 264.0±11.6 | 78.4±0.1k |

| **Lymphography** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .18±.05 | — | 939.2±14.2 |
| PCKMeans | .19±.04 | 10.9±3.8 | 967.1±15.5 |
| LCVQE | .18±.04 | 11.3±4.2 | 968.0±16.7 |
| PCSK-M | .23±.05 | 8.9±2.9 | 958.8±13.8 |
| COP-KM | .22±.05 | 8.4±2.6 | 965.7±13.7 |
| BCK-M | .21±.06 | 19.4±4.9 | 991.7±29.1 |
| H(0,0) | .24±.05 | 8.8±2.5 | 958.6±12.0 |
| H($\alpha$, $\beta$) | **.24±.05** | 7.6±2.2 | 958.1±12.1 |

| **Zoo** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .59±.11 | — | 139.7±12.6 |
| PCKMeans | .60±.10 | 2.3±1.7 | 145.9±12.2 |
| LCVQE | .60±.10 | 2.8±2.0 | 148.7±15.9 |
| PCSK-M | .64±.10 | 2.1±1.4 | 147.1±11.7 |
| COP-KM | **.65±.09** | 1.9±1.3 | 148.3±13.6 |
| BCK-M | .60±.14 | 5.5±2.8 | 157.9±16.2 |
| H(0,0) | .59±.10 | 2.6±1.6 | 147.4±12.4 |
| H($\alpha$, $\beta$) | .62±.10 | 2.0±1.3 | 146.6±12.4 |

| **Image-Seg** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .47±.04 | — | 14.5±1.2k |
| PCKMeans | .45±.04 | 95.3±22.6 | 15.4±1.4k |
| LCVQE | .45±.04 | 102.1±21.5 | 16.3±1.4k |
| PCSK-M | .47±.04 | 78.2±18.1 | 15.1±1.2k |
| COP-KM | .47±.04 | 65.2±10.4 | 15.3±1.1k |
| BCK-M | .47±.05 | 134.0±24.2 | 15.9±1.5k |
| H(0,0) | .47±.04 | 69.3±10.6 | 15.1±1.1k |
| H($\alpha$, $\beta$) | **.50±.05** | 63.9±9.4 | 15.0±1.1k |

| **Heart** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .43±.02 | — | 3.3±0.0k |
| PCKMeans | .35±.06 | 24.6±6.7 | 3.4±0.1k |
| LCVQE | .32±.06 | 28.1±7.4 | 3.4±0.1k |
| PCSK-M | .42±.06 | 22.0±6.0 | 3.4±0.0k |
| COP-KM | .43±.06 | 14.5±4.2 | 3.4±0.0k |
| BCK-M | .49±.05 | 20.8±5.0 | 3.3±0.0k |
| H(0,0) | **.54±.04** | 15.2±3.2 | 3.3±0.0k |
| H($\alpha$, $\beta$) | .53±.04 | 14.4±3.4 | 3.3±0.0k |

| **Ecoli** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .41±.09 | — | 15.1±0.9 |
| PCKMeans | .39±.06 | 17.9±4.8 | 16.4±1.0 |
| LCVQE | .42±.06 | 9.9±3.2 | 15.4±0.7 |
| PCSK-M | .42±.05 | 14.1±3.1 | 15.7±0.8 |
| COP-KM | .40±.06 | 12.7±2.9 | 16.2±0.9 |
| BCK-M | .35±.09 | 63.8±19.8 | 20.7±2.4 |
| H(0,0) | .43±.06 | 13.6±3.1 | 15.7±0.9 |
| H($\alpha$, $\beta$) | **.44±.07** | 11.6±2.7 | 15.7±0.8 |

| **Nursery** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .04±.03 | — | 60.6±0.4k |
| PCKMeans | .06±.06 | 1117.1±71.3 | 62.1±0.2k |
| LCVQE | .06±.05 | 930.9±63.7 | 61.9±0.2k |
| PCSK-M | .09±.08 | 933.7±59.7 | 61.8±0.3k |
| COP-KM | .07±.06 | 884.2±72.1 | 62.1±0.4k |
| BCK-M | .04±.02 | 2686.8±234.2 | 65.1±0.2k |
| H(0,0) | .10±.09 | 903.0±100.1 | 61.7±0.3k |
| H($\alpha$, $\beta$) | **.47±.05** | 552.2±57.3 | 61.3±0.3k |

| **Hepatitis** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .23±.11 | — | 2.5±0.1k |
| PCKMeans | .21±.10 | 14.3±6.2 | 2.6±0.1k |
| LCVQE | .17±.07 | 15.9±6.2 | 2.6±0.1k |
| PCSK-M | .26±.09 | 13.4±4.8 | 2.6±0.1k |
| COP-KM | .31±.09 | 8.9±2.2 | 2.6±0.0k |
| BCK-M | .34±.14 | 10.6±3.0 | 2.6±0.1k |
| H(0,0) | .39±.09 | 8.7±1.7 | 2.5±0.0k |
| H($\alpha$, $\beta$) | **.41±.08** | 8.0±1.7 | 2.5±0.0k |

| **Ionosphere** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .17±.00 | — | 9.1±0.0k |
| PCKMeans | .12±.05 | 39.2±13.5 | 10.0±0.3k |
| LCVQE | .10±.04 | 44.0±13.9 | 10.0±0.3k |
| PCSK-M | .12±.04 | 38.4±9.7 | 9.9±0.2k |
| COP-KM | **.21±.02** | 23.8±5.0 | 9.8±0.2k |
| BCK-M | .16±.03 | 27.8±4.1 | 9.7±0.1k |
| H(0,0) | .15±.03 | 26.4±4.0 | 9.7±0.1k |
| H($\alpha$, $\beta$) | .17±.03 | 26.1±3.8 | 9.7±0.1k |

| **Balance Scale** | ARI | Unh | MSSC |
|---|---|---|---|
| $k$-means | .14±.06 | — | 3.5±0.0k |
| PCKMeans | .19±.09 | 45.1±10.9 | 3.7±0.1k |
| LCVQE | .17±.08 | 44.3±11.5 | 3.6±0.0k |
| PCSK-M | .26±.08 | 43.3±7.5 | 3.7±0.0k |
| COP-KM | .20±.06 | 41.0±6.3 | 3.7±0.0k |
| BCK-M | .20±.11 | 69.8±8.1 | 3.7±0.0k |
| H(0,0) | .30±.08 | 39.4±5.6 | 3.6±0.0k |
| H($\alpha$, $\beta$) | **.38±.06** | 31.6±4.9 | 3.6±0.0k |

## C SCARCE IMBALANCED DATASETS

Table 7 details the class composition and preprocessing of the two real datasets used in the scarce, imbalanced case studies of Section 4.3 (both $K = 2$, $N = D = 500$, 100/400 split).

Table 7: Scarce Imbalanced Datasets ($K = 2$, $N = D = 500$, 100/400 split).

| **Dataset** | **Classes & Domain** | **Type & Processing** |
|---|---|---|
| **20 Newsgroups** (Lang, 1995) | Text: `talk.politics.guns` vs. `sci.med` | Sparse binary bag-of-words (88–91% zeros) |
| **PBMC** (Zheng et al., 2017) | Genomic: Naive vs. Memory T cells | Dense continuous scRNA-seq (log-norm, HVGs, $z$-scored) |

Figure 5 aggregates seed-level results from the hint-budget sweep of Section 4.3 across both datasets and all hint budgets.

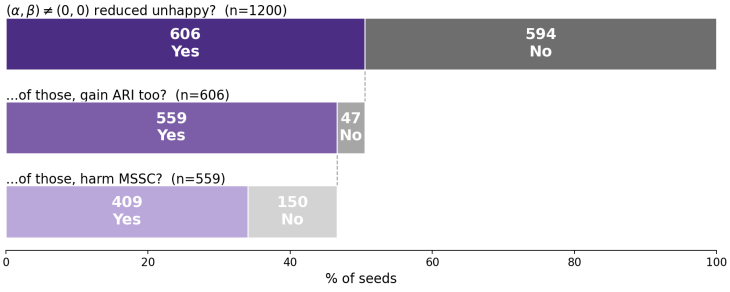



Figure 5: HAPPIER$(0, 0)$ vs. HAPPIER$(\alpha, \beta)$: Newsgroups & PBMC, per seed.