CORESET SPECTRAL CLUSTERING WITHOUT KERNEL INTERMEDIARY

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing coreset methods for spectral clustering commonly proceed through the equivalence between normalized cut and weighted kernel k-means. We study whether the compressed graph can instead be constructed and solved directly in the graph domain. GCSC combines degree-stratified vertex sampling, retention of high-weight input edges, spectral clustering on the endpoints of those edges, and average-affinity lifting. Every retained edge keeps its original weight, and no off-support edge is introduced. For a fixed sample and edge count, we establish edge-tail optimality. We also derive an explicit stratified-sampling bound for endpoint coverage. Under stated perturbation/gap, tolerance-separation, balance, and margin conditions, same-vertex spectral stability and robust lifting yield two-sided full-graph NCut fidelity to a fixed reference partition, such as standard spectral clustering. On six benchmark datasets, GCSC attains the highest mean ARI among the compared coreset methods on five datasets at the largest shared coreset size. Mechanism studies further characterize sampling fidelity, edge-budget sensitivity, and affinity lifting.

1 INTRODUCTION

Spectral clustering (SC) is a widely used method for graph partitioning and non-convex clustering (Von Luxburg, 2007; Ding et al., 2024). It is naturally connected to balanced graph objectives such as Normalized Cut (NCut) (Shi & Malik, 2000). Its principal computational bottleneck is the eigendecomposition of a matrix whose size grows with the number of vertices. This has motivated Nyström approximations, transfer-cut methods, iterative eigensolvers, and coreset constructions (Choromanska et al., 2013; Pourkamali-Anaraki, 2020; Li et al., 2012; Huang et al., 2019; Macgregor, 2023).

Coreset-based SC compresses the graph, solves a smaller clustering problem, and lifts the solution to the original vertices. Existing methods commonly obtain the coreset through the equivalence between NCut and weighted kernel k-means (Dhillon et al., 2004; Jiang et al., 2024; Jourdan et al., 2025). This is a valid reduction, but it makes coreset construction depend on an intermediary objective and transformed coreset weights (Figure 1, left).

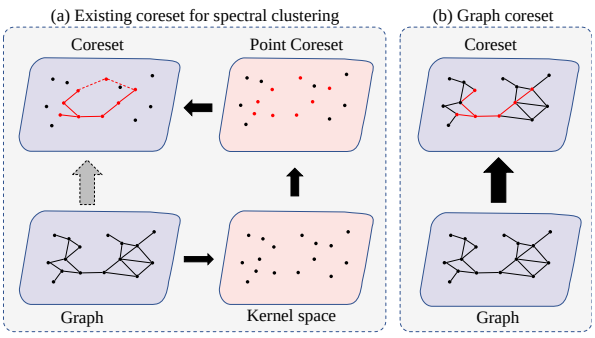


Figure 1: Intermediary-based coreset construction (left) versus graph-native GCSC (right).

We ask a graph-native question: *can the vertices and edges used by the reduced spectral problem be selected directly from the supplied graph?* GCSC answers this question by combining degree-stratified sampling with a weight-preserving edge subgraph (Figure 1, right). The distinction is independent of how the input graph is obtained. Constructing a k-nearest-neighbor similarity graph from vectors is standard practice in spectral clustering and is also used by

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.

CSC (Jourdan et al., 2025); GCSC removes the subsequent kernel k-means intermediary, not the construction of the input graph.

The structural claim is deliberately precise. CSC does not algebraically create new off-diagonal support under its normalized-cut kernel construction. GCSC differs by preserving the exact input weight of every retained edge. If $H = (V_H, E_H, W_H)$ is its reduced graph, then $V_H \subseteq \mathcal{V}$, $E_H \subseteq \mathcal{E}$, and $(W_H)_{uv} = W_{uv}$ for every retained edge. Edge deletion can remove paths, so we claim deletion-only support fidelity and exact retained-weight fidelity rather than universal topology preservation.

Our contributions are:

- We introduce a graph-native compressed-solve-lift pipeline whose spectral core consists of selected input edges and their endpoints.
- We analyze the actual endpoint graph. For fixed S and q, top-edge retention is optimal among q-edge subsets of $\mathcal{E}_S^+$ for edge-mass and entrywise adjacency-tail objectives. An explicit coverage bound and, under stated spectral, tolerance, balance, and margin conditions, robust lifting give two-sided full-graph NCut fidelity.
- We provide a coreset-focused comparison with CSC, CoSC, and Cokkm, together with targeted studies of sampling, edge retention, and propagation.

2 Related Work

The coreset paradigm has been extensively developed for k-means, including sensitivity sampling and constructions whose size can be independent of the number of input points (Har-Peled & Mazumdar, 2004; Chen, 2009; Feldman & Langberg, 2011; Huang & Vishnoi, 2020). Extending this framework to kernel objectives is more challenging because the feature space may be high-dimensional or implicit. Cokkm constructs coresets for kernel clustering, and CoSC applies this machinery to SC through the weighted kernel k-means equivalence (Jiang et al., 2024). CSC further develops a coreset construction for SC through this intermediary (Jourdan et al., 2025). GCSC addresses the same coreset setting but selects and solves the reduced object directly in the graph domain. Nyström, bipartite, and direct graph solvers approximate different stages of the SC pipeline and are complementary rather than component controls for this coreset-specific question.

3 Problem Formulation

Let $G = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ be an undirected loopless graph with nonnegative symmetric weights, $W_{uu} = 0$, and positive weighted degree at every vertex. Let $2 \le k < |\mathcal{V}|$, and let $\mathbf{D}$ be the diagonal degree matrix. For a k-way partition $\mathcal{C} = (C_1, \dots, C_k)$, the NCut objective is

$$J_G(\mathcal{C}) = \sum_{c=1}^{k} \frac{\mathrm{cut}_G(C_c, \mathcal{V} \setminus C_c)}{\mathrm{vol}_G(C_c)} = \mathrm{Tr}\big(\mathbf{Z}_{\mathcal{C}}^{\top} \mathcal{L}_G \mathbf{Z}_{\mathcal{C}}\big), \tag{1}$$

where $\mathcal{L}_G = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ and $(Z_{\mathcal{C}})_{uc} = \sqrt{d_u / \mathrm{vol}_G(C_c)}$ for $u \in C_c$, and zero otherwise. We write $\mathcal{P}_k(U)$ for the ordered partitions of U into k nonempty cells. For any graph F on vertex set U, define $\mathrm{OPT}_k(F) = \min_{\mathcal{C} \in \mathcal{P}_k(U)} J_F(\mathcal{C})$.

For context, let $\mathbf{W}_C$ denote CSC's diagonal coreset-weight matrix. Under the normalized-cut kernel construction, $\mathbf{A}_C = \mathbf{W}_C[\mathbf{K}]_C\mathbf{W}_C$, where $\mathbf{K} = \mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1} + \sigma\mathbf{D}^{-1}$. For distinct selected vertices $u \neq v$,

$$(A_C)_{uv} = (W_C)_{uu}(W_C)_{vv}\frac{W_{uv}}{d_u d_v}. \tag{2}$$

Thus the diagonal shift does not add off-diagonal support. GCSC instead constructs the reduced adjacency directly from $\mathbf{W}$ and leaves every selected weight unchanged.

Fix a total order $\prec$ on $\mathcal{V}$. For $|V_H| \ge k$ and $\mathcal{C} = (C_1, \dots, C_k) \in \mathcal{P}_k(\mathcal{V})$, begin with the nonempty intersections $C_c \cap V_H$, retained in increasing original cell index. If fewer than k cells remain, repeatedly select the first cell of size at least two under the lexicographic order induced by $\prec$ and split off its $\prec$-largest vertex, appending the resulting singleton to the end of the ordered list. Denote

the resulting ordered k-partition of V_H by $\Gamma_H(\mathcal{C})$. This fixed restriction–repair map uses neither edge weights nor objective values; when every C_c meets V_H, $\Gamma_H(\mathcal{C}) = (C_1 \cap V_H, \ldots, C_k \cap V_H)$.

Definition 3.1 (Strong graph-coreset for spectral clustering)**.** Let $0 < \epsilon < 1$, let $H = (V_H, E_H, W_H)$ satisfy $|V_H| \geq k$ and $d_H(u) > 0$ for every $u \in V_H$, and let Γ_H be the fixed map above. The pair (H, Γ_H) is a strong $(1 \pm \epsilon)$ graph-coreset if

$$(1-\epsilon)J_G(\mathcal{C}) \leq J_H(\Gamma_H(\mathcal{C})) \leq (1+\epsilon)J_G(\mathcal{C}) \quad \text{for every } \mathcal{C} \in \mathcal{P}_k(\mathcal{V}),$$

and

$$E_H \subseteq \big\{\{u,v\} \in \mathcal{E} : u, v \in V_H\big\}, \qquad (W_H)_{uv} = W_{uv} \quad \text{for every } \{u,v\} \in E_H.$$

The fixed map removes a per-query existential choice. Definition 3.1 is an unconditional all-partition benchmark and does not assume or privilege a particular cut. Algorithm 1 is not claimed to satisfy this strong definition; its guarantee is the separately stated conditional, reference-relative solution-coreset result. The weight-preserving subgraph property does hold deterministically.

Its universal quantifier has a concrete solve-and-lift consequence rather than assuming a particular cut: Proposition B.1 proves that an (a, ξ)-approximate solution on a strong coreset lifts to $a(1+\epsilon)/(1-\epsilon)\,\mathrm{OPT}_k(G) + \xi/(1-\epsilon)$ on G.

4 METHOD

Degree-stratified sample. GCSC sorts vertices by weighted degree, partitions the ordering into m strata, and samples $s = \lfloor \frac{p}{m} \rfloor$ vertices uniformly without replacement from each stratum. Let S be the resulting sample. The draws in distinct strata are independent.

This controls representation across the degree range without treating degree as a class label. In a stratum T_j of size N_j with quota s_j, every $u \in T_j$ has inclusion probability s_j/N_j; hence, for any fixed vertex set B,

$$\mathbb{E}|S \cap B| = \sum_j \frac{s_j}{N_j}|T_j \cap B|.$$

When the stratum rates are equal, this is proportional to $|B|$ regardless of the correlation between degree and a semantic label.

Top-edge endpoint graph. Let

$$\mathcal{E}_S^+ = \big\{\{u,v\} \in \mathcal{E} : u, v \in S,\ u \prec v,\ W_{uv} > 0\big\}.$$

Before sampling, fix a total order on all positive full-graph edges by non-increasing weight, breaking ties lexicographically by their ordered endpoints. The sampled edges inherit the restriction of this order. GCSC's default construction uses

$$k_E = \left\lfloor \frac{s^2}{4} \right\rfloor, \qquad e_H = \min\{k_E, |\mathcal{E}_S^+|\}, \qquad E_H = \mathrm{Top}_{e_H}(\mathcal{E}_S^+). \tag{3}$$

The reduced graph is formed by the retained edges and their endpoints:

$$V_H = \bigcup_{\{u,v\} \in E_H} \{u,v\}, \qquad H = (V_H, E_H, W|_{E_H}). \tag{4}$$

Thus the spectral vertex set is V_H, not all of S. Every vertex in V_H has positive degree in H, while $S \setminus V_H$ is handled by the same lifting step as $\mathcal{V} \setminus S$.

Spectral solution and lift. GCSC applies normalized spectral clustering to H and obtains $C_H = (C_{H,1}, \ldots, C_{H,k})$. Every $u \in \mathcal{V} \setminus V_H$ is assigned by average affinity,

$$\widehat{y}_u = \arg\max_{c \in [k]} \frac{1}{|C_{H,c}|} \sum_{v \in C_{H,c}} W_{uv}, \tag{5}$$

with deterministic tie-breaking. Dividing by $|C_{H,c}|$ removes the systematic advantage that an unnormalized sum gives to a cluster with more active representatives.

Complete pseudocode, including the insufficient-active-core check, appears in Appendix Algorithm 1.

Complexity. The dominant spectral problem contracts from $n = |\mathcal{V}|$ vertices to $h = |V_H| \le p$. Edge selection uses a size-k_E heap, and the lift can be implemented by scanning active-to-inactive edges. Full pseudocode and the sparse- and dense-input cost expressions are given in Appendix A.

5 Theoretical Analysis

5.1 Structural properties of top-edge endpoints

The first result is unconditional: it holds for every realized sample and does not invoke a cut-preservation assumption. For a matrix $\mathbf{B}$, write $\|\mathbf{B}\|_{\mathrm{ent},r}^r = \sum_{u,v} |B_{uv}|^r$. For $F \subseteq \mathcal{E}_S^+$, define its common-index, zero-padded adjacency on S by $(\mathbf{A}_F^S)_{uv} = W_{uv}\mathbf{1}\{\{u,v\} \in F\}$, and let $\mathbf{A}_S^S = \mathbf{W}[S,S]$.

Proposition 5.1 (Top-edge optimality and endpoint geometry)**.** *Fix S, let $M = |\mathcal{E}_S^+|$, $q = e_H$, and let $w_{(1)} \ge \cdots \ge w_{(M)} > 0$ follow the fixed edge order. Among all q-edge subsets $F \subseteq \mathcal{E}_S^+$, E_H maximizes $\sum_{e\in F} \varphi(W_e)$ for every nondecreasing φ. Consequently, for every $1 \le r < \infty$,*

$$E_H \in \arg\min_{|F|=q} \|\mathbf{A}_S^S - \mathbf{A}_F^S\|_{\mathrm{ent},r}, \qquad \|\mathbf{A}_S^S - \mathbf{A}_{E_H}^S\|_{\mathrm{ent},r}^r = 2\sum_{i=q+1}^{M} w_{(i)}^r. \tag{6}$$

If $q \ge 1$ and $h = |V_H|$, then

$$\left\lceil \frac{1+\sqrt{1+8q}}{2} \right\rceil \le h \le \min\{|S|, 2q\}, \qquad d_H(u) \ge \min_{e\in E_H} W_e > 0. \tag{7}$$

In particular, $q > \binom{k-1}{2}$ guarantees $h \ge k$, and increasing the edge budget for a fixed S gives nested edge and endpoint sets.

Equation (6) proves optimal retained edge mass and entrywise adjacency approximation among q-edge subsets of $\mathcal{E}_S^+$, not universal cut or spectral preservation. The latter is measured separately below. Proofs, including the exchange argument and the sharp endpoint-count bounds, are in Appendix B.2.

Fix before sampling a reference partition $C^{\mathrm{ref}} = (C_1^{\mathrm{ref}}, \ldots, C_k^{\mathrm{ref}}) \in \mathcal{P}_k(\mathcal{V})$. For each $u \in \mathcal{V}$, let y_u^{ref} be the unique index such that $u \in C_{y_u^{\mathrm{ref}}}^{\mathrm{ref}}$. For two k-partitions C and C' of the same vertex set, write $C \sim C'$ if their cells agree under a permutation of the cluster indices. It may be the deterministic output of standard SC on G and need not be ground truth. The construction, reduced solver, and lift do not read its labels. All probabilities below are over the degree-stratified sample and any randomization of the reduced solver. The reference partition, vertex order, and tie-breaking rules are fixed before these random choices; termination due to $|V_H| < k$ counts as a failed realization.

For $C_H = (C_{H,1}, \ldots, C_{H,k}) \in \mathcal{P}_k(V_H)$, define the full label map

$$\big[\Pi_{G,V_H}^{\mathrm{aff}}(C_H)\big](u) = \begin{cases} c, & u \in C_{H,c}, \\ \arg\max\limits_{c\in[k]} \dfrac{1}{|C_{H,c}|} \displaystyle\sum_{v\in C_{H,c}} W_{uv}, & u \in \mathcal{V}\setminus V_H, \end{cases} \tag{8}$$

using the fixed tie-breaking rule. We identify this label map with its induced ordered partition in $\mathcal{P}_k(\mathcal{V})$; every cell is nonempty because it contains $C_{H,c}$. For a reduced solver $\mathcal{A}$, the full GCSC output is $\widehat{C} = \Pi_{G,V_H}^{\mathrm{aff}}(\mathcal{A}(H))$.

Definition 5.2 (Reference-relative graph solution coreset)**.** Fix a reduced-graph solver $\mathcal{A}$. A randomized construction returns $(H_\omega, \Pi_{G,V_{H_\omega}}^{\mathrm{aff}})$ under outcome ω, which collects construction and solver randomness. On $|V_{H_\omega}| \ge k$, let $\widehat{C}_\omega = \Pi_{G,V_{H_\omega}}^{\mathrm{aff}}(\mathcal{A}(H_\omega))$; otherwise declare the success event below false. For $\alpha \ge 1$, $\zeta \ge 0$, and $\delta \in [0,1]$, this random pair is an (α,ζ,δ) reference-relative graph solution coreset if

$$\Pr\Big[|V_{H_\omega}| \ge k \quad \text{and} \quad J_G(\widehat{C}_\omega) \le \alpha J_G(C^{\mathrm{ref}}) + \zeta\Big] \ge 1-\delta.$$

5.2 SAME-VERTEX SPECTRAL STABILITY

All results in this subsection condition on a nonfailed realization $|V_H| \geq k$. The deletion analysis must compare graphs on the same vertices. Define

$$G_R := G[V_H],$$

and write $\mathbf{A}_R, \mathbf{D}_R, \mathcal{L}_R$ and $\mathbf{A}_H, \mathbf{D}_H, \mathcal{L}_H$ for G_R and H. Let

$$\mathbf{R}_H = \mathbf{A}_R - \mathbf{A}_H, \qquad \mathbf{r} = \mathbf{R}_H \mathbf{1}, \qquad \rho_H = \max_{u \in V_H} \frac{r_u}{d_R(u)}. \tag{9}$$

Lemma 5.3 (Normalized perturbation of the retained graph)**.** *Every $u \in V_H$ is incident to a retained positive edge, so $0 \leq \rho_H < 1$. Moreover,*

$$\eta_H := \|\mathcal{L}_H - \mathcal{L}_R\|_2 \leq \min\left\{2, \frac{2\rho_H}{1-\rho_H}\right\}. \tag{10}$$

Lemma 5.3 is deterministic conditional on the realized S and H; it does not assert that the fixed edge budget has small error for every graph. Two further deterministic properties are given in Appendix B.5: Corollary B.3 turns the total discarded edge weight into a computable certificate for η_H, while Proposition B.4 bounds the NCut distortion simultaneously for every k-partition of the realized active vertex set and transfers an approximate solve on H back to G_R.

Let $h = |V_H| \geq k+1$ and let $\mathbf{U}_R, \mathbf{U}_H \in \mathbb{R}^{h \times k}$ contain orthonormal bottom-k eigenvectors of $\mathcal{L}_R$ and $\mathcal{L}_H$. Define

$$g_R = \lambda_{k+1}(\mathcal{L}_R) - \lambda_k(\mathcal{L}_R).$$

For $C \in \mathcal{P}_k(V_H)$, let $(\mathbf{Z}_C^{\mathrm{km}})_{uc} = |C_c|^{-1/2}\mathbf{1}\{u \in C_c\}$, set $\mathbf{P}_C = \mathbf{Z}_C^{\mathrm{km}}(\mathbf{Z}_C^{\mathrm{km}})^\top$, and define $\Phi_{\mathbf{U}}(C) = \|(\mathbf{I} - \mathbf{P}_C)\mathbf{U}\|_F^2$.

Lemma 5.4 (Eigenspace, eigenvalue, and objective stability)**.** *If $g_R > 0$ and $\eta_H < g_R/2$, then*

$$\min_{\mathbf{O}^\top\mathbf{O}=\mathbf{I}} \|\mathbf{U}_H\mathbf{O} - \mathbf{U}_R\|_F \leq e_{\mathrm{emb}} := \frac{\sqrt{2k}\,\eta_H}{g_R - \eta_H}. \tag{11}$$

Furthermore, for every i,

$$|\lambda_i(\mathcal{L}_H) - \lambda_i(\mathcal{L}_R)| \leq \eta_H, \quad g_H \geq g_R - 2\eta_H, \quad \left|\sum_{i=1}^k \lambda_i(\mathcal{L}_H) - \sum_{i=1}^k \lambda_i(\mathcal{L}_R)\right| \leq k\eta_H, \tag{12}$$

where $g_H = \lambda_{k+1}(\mathcal{L}_H) - \lambda_k(\mathcal{L}_H)$. Moreover, uniformly over $C \in \mathcal{P}_k(V_H)$,

$$|\Phi_{\mathbf{U}_H}(C) - \Phi_{\mathbf{U}_R}(C)| \leq d_{\mathrm{emb}} := 2\sqrt{k}\,e_{\mathrm{emb}} + e_{\mathrm{emb}}^2. \tag{13}$$

The projector here is not the degree-scaled NCut indicator in Eq. (1). Equation (13) controls the discrete rounding objective uniformly, which permits a tolerance result rather than requiring a unique, exactly recovered partition.

5.3 ENDPOINT COVERAGE, ROBUST LIFTING, AND FULL-GRAPH FIDELITY

Let the degree strata be T_j, with $N_j = |T_j|$ and quota s_j (GCSC uses $s_j = s$). Assume $1 \leq s_j \leq N_j$ for every stratum. For reference cell c, define

$$M_{jc} = |T_j \cap C_c^{\mathrm{ref}}|, \quad X_c = |S \cap C_c^{\mathrm{ref}}|, \quad Y_c = |V_H \cap C_c^{\mathrm{ref}}|, \quad \mu_c = \mathbb{E}X_c = \sum_j \frac{s_j}{N_j} M_{jc}.$$

Set $E_{\mathrm{cov}}^H = \{\min_c Y_c \geq 1\}$. Proposition B.2 derives a general lower-tail bound and, in particular, the explicit certificate

$$\Pr[(E_{\mathrm{cov}}^H)^c] \leq \delta_{\mathrm{cov}} := \min\left\{1, \sum_{c=1}^k (e^{-\mu_c} + \chi_c)\right\}, \tag{14}$$

where the computable χ_c in Eq. (21) bounds the expected number of sampled cell-c vertices that fail to become endpoints. The proof uses a global-rank protection property of top edges, so it does not identify sample coverage with endpoint coverage.

On E^H_{cov}, write $A_c = V_H \cap C^{\mathrm{ref}}_c$ and fix $b_H \geq 0$, $\bar{\theta}_{1:k} \in [0,1]^k$, and $\xi \geq 0$ before the random draw. Appendix B.7 defines the degree-weighted label alignment, active disagreement q_H, cellwise uniform-measure discrepancies θ_c, tolerance profile κ_R, and event E^H_{core}. On that event, Proposition B.5 guarantees an aligned solver output with $q_H \leq b_H \operatorname{vol}_G(\mathcal{V})$ and $\theta_c \leq \bar{\theta}_c$ for every cell. Thus the analysis permits active error; exact active recovery is only the special case $b_H = 0$ and $\bar{\theta}_c = 0$ for all c. Write the aligned active partition as $\widehat{C}_H = (\widehat{A}_1, \ldots, \widehat{A}_k)$ and the aligned full output as $\widehat{C}$; relabeling does not change NCut.

For $u \in \mathcal{V} \setminus V_H$, define the reference scores and margin

$$a^H_{uc} = |A_c|^{-1} \sum_{v \in A_c} W_{uv}, \qquad m^H_u = a^H_{u,y^{\mathrm{ref}}_u} - \max_{c \neq y^{\mathrm{ref}}_u} a^H_{uc},$$

and the active affinity range $\Omega^H_u = \max_{v \in V_H} W_{uv} - \min_{v \in V_H} W_{uv}$. Put

$$\bar{r}^H_u = \Omega^H_u \left(\bar{\theta}_{y^{\mathrm{ref}}_u} + \max_{c \neq y^{\mathrm{ref}}_u} \bar{\theta}_c \right), \quad B^H_{\Delta,\mathrm{rob}} = \{ u \notin V_H : m^H_u < \Delta + \bar{r}^H_u \}, \tag{15}$$

where $\Delta > 0$.

Lemma 5.5 (Robust average-affinity lifting). *For the aligned solver cells $\widehat{A}_c = \widehat{A}_c(C^{\mathrm{raw}}_H)$,*

$$\left| |\widehat{A}_c|^{-1} \sum_{v \in \widehat{A}_c} W_{uv} - a^H_{uc} \right| \leq \Omega^H_u \theta_c(C^{\mathrm{raw}}_H). \tag{16}$$

Hence, if $C^{\mathrm{raw}}_H \in \mathcal{G}_R(b_H, \bar{\theta}_{1:k})$, the affinity lift assigns every $u \notin V_H \cup B^H_{\Delta,\mathrm{rob}}$ its reference label. Moreover, Eq. (26) *is exactly the total-variation distance between the uniform measures on A_c and $\widehat{A}_c$.*

This result covers both $\mathcal{V} \setminus S$ and $S \setminus V_H$ and allows the active core itself to contain errors.

Lemma 5.6 (Two-sided NCut stability). *For any fixed alignment of $C = (C_c)_{c=1}^k$ and $\widetilde{C} = (\widetilde{C}_c)_{c=1}^k$, let*

$$t_c = \operatorname{vol}_G(C_c \triangle \widetilde{C}_c), \quad v_c = \operatorname{vol}_G(C_c), \quad r_c = t_c / v_c, \quad r = \max_c r_c, \quad R = \sum_c r_c.$$

If $r < 1$, then

$$\max\left\{0, \frac{J_G(C) - R}{1 + r}\right\} \leq J_G(\widetilde{C}) \leq \frac{J_G(C) + R}{1 - r}, \qquad |J_G(\widetilde{C}) - J_G(C)| \leq \frac{2R}{1 - r}. \tag{17}$$

If M is the disagreement set under this alignment, then $\sum_c t_c = 2 \operatorname{vol}_G(M)$.

Theorem 5.7 (Conditional robust endpoint-graph solution guarantee). *Assume, for some $\beta \in (0, 1/k]$, $\operatorname{vol}_G(C^{\mathrm{ref}}_c) \geq \beta \operatorname{vol}_G(\mathcal{V})$ for every c. Choose $b_H, b_L \geq 0$ and let*

$$E^H_{\mathrm{mar}} = E^H_{\mathrm{cov}} \cap \{ \operatorname{vol}_G(B^H_{\Delta,\mathrm{rob}}) \leq b_L \operatorname{vol}_G(\mathcal{V}) \}.$$

Use δ_{cov} from Eq. (14)*, or a sharper choice from Proposition B.2. For $\delta_{\mathrm{core}}, \delta_{\mathrm{mar}} \in [0,1]$, suppose*

$$\Pr[(E^H_{\mathrm{core}})^c \cap E^H_{\mathrm{cov}}] \leq \delta_{\mathrm{core}}, \qquad \Pr[(E^H_{\mathrm{mar}})^c \cap E^H_{\mathrm{cov}} \cap E^H_{\mathrm{core}}] \leq \delta_{\mathrm{mar}}.$$

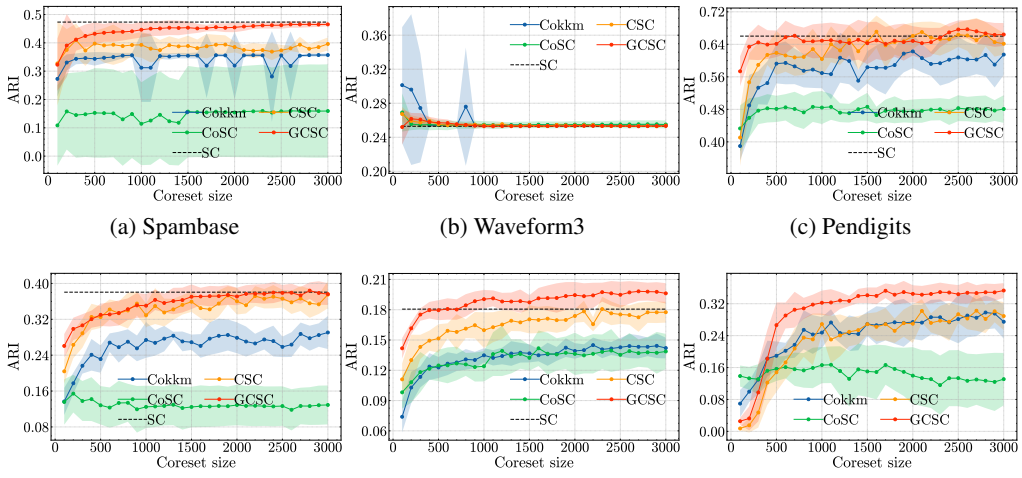
If $b = b_H + b_L < \beta$, define $\rho = b/\beta$, $\qquad \delta_{\mathrm{tot}} = \delta_{\mathrm{cov}} + \delta_{\mathrm{core}} + \delta_{\mathrm{mar}} \leq 1$.

Then, with probability at least $1 - \delta_{\mathrm{tot}}$,

$$\max\left\{0, \frac{J_G(C^{\mathrm{ref}}) - 2\rho}{1 + \rho}\right\} \leq J_G(\widehat{C}) \leq \frac{J_G(C^{\mathrm{ref}}) + 2\rho}{1 - \rho}, \qquad |J_G(\widehat{C}) - J_G(C^{\mathrm{ref}})| \leq \frac{4\rho}{1 - \rho}. \tag{18}$$

Thus Definition 5.2 holds with $\alpha = (1 - \rho)^{-1}$, $\zeta = 2\rho/(1 - \rho)$, and $\delta = \delta_{\mathrm{tot}}$.

The conditional guarantee degrades continuously with active-core and lifting error; it does not require perfect reduced clustering. When C^{ref} is standard Full SC, Eq. (18) states objective-value fidelity to Full SC, not perfect semantic recovery or agreement with ground truth. Only when C^{ref} is an NCut optimum may $J_G(C^{\mathrm{ref}})$ be replaced by $\mathrm{OPT}_k(G)$.



(a) Spambase (b) Waveform3 (c) Pendigits

(d) USPS (e) Letters (f) MNIST

Figure 2: ARI as a function of coreset size. Method-specific affinity settings are selected at $p = 3000$ and held fixed across the curve. Solid curves show ten-trial means with one-sample-standard-deviation bands; the dashed line is the single-run Full-SC reference, where available.

6 Experiments

We evaluate GCSC on Spambase (Hopkins et al., 1999), Waveform3 (Breiman & Stone, 1984), PenDigits (Alpaydin & Alimoglu, 1996), USPS (Hull, 1994), Letters (Frey & Slate, 1991), and MNIST (LeCun et al., 1998).

Following common SC practice and the evaluation protocol of CSC, standardized vectors are converted to exponential-distance-weighted 500-nearest-neighbor graphs. For a directed neighbor edge,

$$W_{ij}^{\mathrm{dir}} = \exp\left(-\frac{\|x_i - x_j\|_2}{2\gamma^2}\right), \qquad \mathbf{W} = \mathbf{W}^{\mathrm{dir}} + (\mathbf{W}^{\mathrm{dir}})^\top.$$

For each γ, the same graph is supplied to every method. We compare only with the aligned coreset baselines Cokkm, CoSC (Jiang et al., 2024), and CSC (Jourdan et al., 2025); Full SC is the uncompressed reference rather than an additional coreset competitor.

We report ARI (Rand, 1971; Gates & Ahn, 2017), NMI (McDaid et al., 2011), full-graph NCut, coreset construction time, and total algorithm runtime after the common feature-standardization and kNN graph-construction stages. We use $m = 10$ and $p \in \{100, 200, \ldots, 3000\}$.

For each coreset method–dataset pair, the ARI, NMI, and total-runtime curves select the setting maximizing mean ARI at $p = 3000$ and hold it fixed across all displayed budgets. Each coreset curve aggregates ten independent trials; solid curves are means and bands show one sample standard deviation. The common search grid, parameter settings, and Full-SC protocol are given in Appendix C.1.

6.1 Comparison within the coreset family

Figure 2 shows the quality–budget tradeoff. At the largest shared budget, GCSC has the highest mean ARI on five of the six manuscript datasets (Table 1). The advantage is not universal: CoSC is slightly better on Waveform3, and the difference on USPS is negligible. The improvements over the strongest coreset baseline are significant on Letters, MNIST, and Spambase.

The NMI trends in Appendix Figure 6 are consistent with the ARI comparison: GCSC is broadly competitive within the coreset family, while gains vary by dataset and budget.

Table 1: Mean ARI at $p = 3000$. Baseline is the strongest coreset comparator. The last row reports Holm-adjusted two-sided p-values.

	Letters	MNIST	PenDigits	Spambase	USPS	Waveform3
GCSC	.196	.353	.664	.465	.376	.253
Best baseline	CSC .178	CSC .289	CSC .641	CSC .396	CSC .375	CoSC .255
ΔARI	+.019	+.064	+.023	+.069	+.001	-.002
Holm p	.003	$< .001$	.318	$< .001$	1.000	1.000

Table 2: Fidelity of the unpruned sampling-and-lifting mechanism to Full SC at $p = 3000$. The last row is the absolute relative full-graph NCut gap.

	Letters	PenDigits	Spambase	USPS	Waveform3	Mean
ARI to Full SC	.664	.936	.944	.727	.964	.847
NMI to Full SC	.793	.944	.894	.776	.938	.869
$\lvert\Delta J\rvert / J_{\mathrm{SC}}$	.027	.050	.004	.007	.003	.018

6.2 Fidelity to standard spectral clustering

The approximation target is Full SC on the same input graph, not perfect semantic recovery under every graph model. To isolate the sampling and lifting mechanism from edge deletion, we solve the unpruned sampled graph at $p = 3000$ and compare the lifted partition directly with Full SC.

Table 2 shows a small objective gap on all five datasets. Partition agreement is highest on PenDigits, Spambase, and Waveform3, and lower on Letters and USPS. MNIST is omitted because a Full-SC reference is unavailable. These results address weak degree–label correlation directly: stratified compression can follow the Full-SC partition even when degree is not a semantic classifier. At $p = 600$, degree-stratified and uniform sampling attain mean ARI-to-Full-SC of .732 and .728, respectively, indicating that stratification behaves close to uniform sampling when degree provides little cluster information.

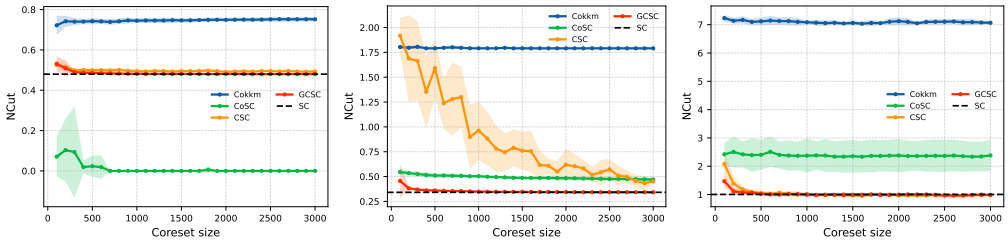
6.3 Objective fidelity and construction efficiency

Figure 3 evaluates the objective optimized by SC. GCSC approaches the Full-SC value on several datasets as the coreset grows, while similar NCut values can still correspond to different discrete partitions when solutions are nearly tied. Figure 4 isolates coreset construction from the subsequent eigensolve and lift. GCSC is the fastest construction method on Letters, USPS, and MNIST and remains competitive on the smaller datasets, supporting the benefit of operating directly on the graph.

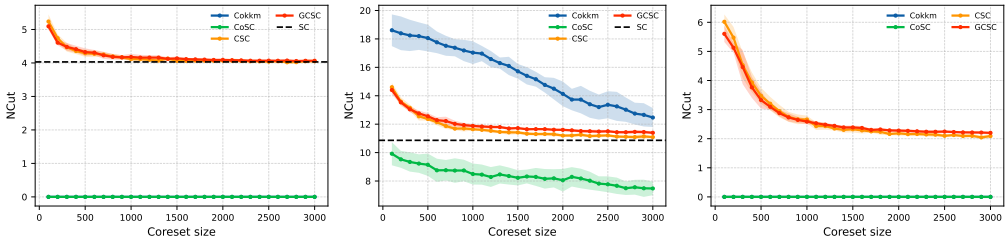
6.4 Propagation and edge retention

Average affinity is designed to remove representative-count bias. At $p = 600$, using the parameter that maximizes GCSC's ten-trial mean ARI at this budget for each dataset, the controlled study in Appendix C.5 shows an ARI gain over unnormalized sum of .141 under minority under-representation, compared with .009 for the natural oracle core and .002 for the unpruned spectral core. This supports the normalization mechanism, but not universal dominance over every lifting rule.

Retained-edge truncation is a separate source of approximation error. The same-vertex comparison required by Section 5 is H versus $G[V_H]$, not H versus $G[S]$. The diagnostic in Appendix C.6 shows that the fixed budget can be aggressive on some graphs, whereas retaining the unpruned sampled graph substantially improves Full-SC fidelity. This does not invalidate kNN graph construction or top-edge selection as a principle; it identifies the edge truncation at the chosen budget as a graph-dependent limitation and motivates the conditional form of Theorem 5.7.

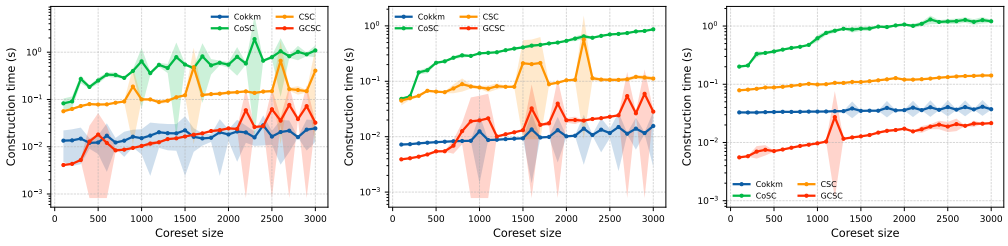


(a) Spambase (b) Waveform3 (c) Pendigits

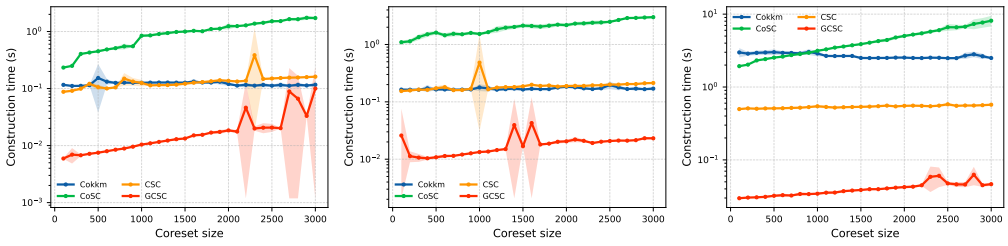


(d) USPS (e) Letters (f) MNIST

Figure 3: Full-graph NCut across coreset sizes, using each method's selected affinity setting.



(a) Spambase (b) Waveform3 (c) PenDigits



(d) USPS (e) Letters (f) MNIST

Figure 4: Coreset construction time across coreset sizes, using each method's selected affinity setting.

7 Conclusion

GCSC constructs and solves a spectral-clustering coreset directly in the graph domain, without a weighted kernel k-means intermediary, while preserving retained-edge support and weights. For fixed S and q, the theory characterizes top-edge optimality, endpoint coverage, active-graph distortion, tolerance-aware rounding, and robust lifting, yielding conditional two-sided NCut fidelity to a fixed full-graph reference. Experiments show broad gains within the coreset family and identify overly aggressive edge retention as the main limitation at the tested budget.

AI USE STATEMENT

Generative AI assisted with language editing, literature and reference checks, and the checking of theoretical statements and proofs. This introduced no new experimental measurements. The authors retain responsibility for the final manuscript, including verification of all AI-assisted content.

ETHICS STATEMENT

This work concerns unsupervised clustering of benchmark feature matrices. No new human-subject data were collected. Clustering can amplify representation bias, so deployment in consequential settings requires dataset-specific fairness, privacy, and error analysis.

REFERENCES

E. Alpaydin and Fevzi. Alimoglu. Pen-Based Recognition of Handwritten Digits. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5MG6K.

Leo Breiman and Charles J. Stone. Waveform Database Generator (Version 1). UCI Machine Learning Repository, 1984. DOI: https://doi.org/10.24432/C5CS3C.

Ke Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.

Anna Choromanska, Tony Jebara, Hyungtae Kim, Mahesh Mohan, and Claire Monteleoni. Fast spectral clustering via the nyström method. In *International Conference on Algorithmic Learning Theory*, pp. 367–381. Springer, 2013.

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 551–556, 2004.

Ling Ding, Chao Li, Di Jin, and Shifei Ding. Survey of spectral clustering based on graph theory. *Pattern Recognition*, 151:110366, 2024.

Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 569–578, 2011.

Peter W Frey and David J Slate. Letter recognition using holland-style adaptive classifiers. *Machine learning*, 6:161–182, 1991.

Alexander J Gates and Yong-Yeol Ahn. The impact of random models on clustering similarity. *Journal of Machine Learning Research*, 18(87):1–28, 2017.

Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pp. 291–300, 2004.

Mark Hopkins, Erik Reeber, George Forman, and Jaap Suermondt. Spambase. UCI Machine Learning Repository, 1999. DOI: https://doi.org/10.24432/C53G6X.

Dong Huang, Chang-Dong Wang, Jian-Sheng Wu, Jian-Huang Lai, and Chee-Keong Kwoh. Ultra-scalable spectral clustering and ensemble clustering. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1212–1226, 2019.

Lingxiao Huang and Nisheeth K Vishnoi. Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1416–1429, 2020.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. DOI: https://doi.org/10.1109/34.291440.

Shaofeng H-C Jiang, Robert Krauthgamer, Jianing Lou, and Yubo Zhang. Coresets for kernel clustering. *Machine Learning*, 113(8):5891–5906, 2024.

Ben Jourdan, Gregory Schwartzman, Peter Macgregor, and He Sun. Coreset spectral clustering. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. DOI: https://doi.org/10.1109/5.726791.

Zhenguo Li, Xiao-Ming Wu, and Shih-Fu Chang. Segmentation using superpixels: A bipartite graph partitioning approach. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 789–796. IEEE, 2012.

Peter Macgregor. Fast and simple spectral clustering in theory and practice. *Advances in Neural Information Processing Systems*, 36:34410–34425, 2023.

Aaron F McDaid, Derek Greene, and Neil Hurley. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*, 2011.

Farhad Pourkamali-Anaraki. Scalable spectral clustering with nyström approximation: Practical and theoretical aspects. *IEEE Open Journal of Signal Processing*, 1:242–256, 2020.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Robert J Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pp. 39–48, 1974.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

A Algorithm and Complexity Details

The pseudocode of GCSC is shown in Algorithm 1. The source code is avaliable in `https://anonymous.4open.science/r/GCSC/`.

Algorithm 1 Graph-Coreset Spectral Clustering (GCSC)

Require: $G = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, clusters k, budget p, strata count m
Ensure: labels $\widehat{\mathbf{y}} \in [k]^{|\mathcal{V}|}$ or an insufficient-active-core failure flag
1: Compute weighted degrees, sort the vertices, and partition them into m strata.
2: Sample $s = \lfloor p/m \rfloor$ vertices without replacement from each stratum; let their union be S.
3: $\mathcal{E}_S^+ \leftarrow \{\{u, v\} \in \mathcal{E} : u, v \in S,\ u \prec v,\ W_{uv} > 0\}$.
4: $E_H \leftarrow \mathrm{Top}_{\min\{\lfloor s^2/4 \rfloor, |\mathcal{E}_S^+|\}}(\mathcal{E}_S^+)$ under the fixed global edge order.
5: $V_H \leftarrow \bigcup_{\{u,v\} \in E_H} \{u, v\}$ and $H \leftarrow (V_H, E_H, W|_{E_H})$.
6: If $|V_H| < k$, terminate and report an insufficient active core.
7: Form $\mathcal{L}_H = \mathbf{I} - \mathbf{D}_H^{-1/2} \mathbf{A}_H \mathbf{D}_H^{-1/2}$.
8: $\mathbf{U}_H \leftarrow$ the bottom-k eigenvectors of $\mathcal{L}_H$; $C_H \leftarrow k$-means$(\mathbf{U}_H, k)$.
9: Set $\widehat{y}_u \leftarrow c$ for every $u \in C_{H,c}$ and $c \in [k]$.
10: **for** each $u \in \mathcal{V} \setminus V_H$ **do**
11: $\quad \mathrm{score}(u, c) \leftarrow |C_{H,c}|^{-1} \sum_{v \in C_{H,c}} W_{uv}$ for $c \in [k]$.
12: $\quad \widehat{y}_u \leftarrow \arg\max_c \mathrm{score}(u, c)$.
13: **end for**
14: **return** $\widehat{\mathbf{y}}$

Let $n = |\mathcal{V}|$, $h = |V_H|$, $e_H = |E_H|$,

$$M_S = \big|\{\{u, v\} \in \mathcal{E} : u, v \in S\}\big|, \quad M_\times = \big|\{\{u, v\} \in \mathcal{E} : u \in V_H, v \notin V_H\}\big|.$$

For sparse adjacency-list input, the principal cost is

$$\begin{aligned} &O\big(|\mathcal{E}| + n \log n + M_S \log k_E + T_{\mathrm{eig}}(e_H k + h k^2) \\ &\quad + T_{\mathrm{km}} h k^2 + M_\times + nk\big), \end{aligned} \tag{19}$$

where a size-k_E heap identifies the largest sampled edges. For dense input, the lift costs $O((n-h)h)$. The eigensolve and lift can dominate despite the contraction from n vertices to $h \le p$.

B Additional Theoretical Properties and Proofs

Proposition B.1 (Strong-coreset solve transfer)**.** *The map Γ_H is surjective. If (H, Γ_H) is a strong $(1 \pm \epsilon)$ graph-coreset, $a \ge 1$, $\xi \ge 0$, and $J_H(Q) \le a \min_{P \in \mathcal{P}_k(V_H)} J_H(P) + \xi$, then every $\widetilde{C} \in \mathcal{P}_k(\mathcal{V})$ with $\Gamma_H(\widetilde{C}) = Q$ obeys*

$$J_G(\widetilde{C}) \le a \frac{1+\epsilon}{1-\epsilon} \mathrm{OPT}_k(G) + \frac{\xi}{1-\epsilon}. \tag{20}$$

A lift that leaves labels on V_H unchanged is such a preimage.

B.1 Proof of Proposition B.1

Proof. For any $Q = (Q_1, \ldots, Q_k) \in \mathcal{P}_k(V_H)$, assign the vertices outside V_H arbitrarily to its k cells. The resulting full partition has ordinary restriction Q, so no repair is invoked and Γ_H is surjective. The same argument applies to any lift that preserves Q on V_H.

Let $C^\star$ minimize J_G and let $P^\star = \Gamma_H(C^\star)$. Strong coreset preservation gives

$$\min_{P \in \mathcal{P}_k(V_H)} J_H(P) \le J_H(P^\star) \le (1+\epsilon)\, \mathrm{OPT}_k(G).$$

For a preimage $\widetilde{C}$ of Q, the other side of the strong inequality and the assumed solver accuracy yield

$$\begin{aligned} (1-\epsilon) J_G(\widetilde{C}) &\le J_H(Q) \\ &\le a \min_P J_H(P) + \xi \le a(1+\epsilon)\, \mathrm{OPT}_k(G) + \xi. \end{aligned}$$

Division by $1 - \epsilon$ proves Eq. (20). □

B.2 Proof of Proposition 5.1

Proof. Suppose a q-edge candidate F omits an earlier edge e_i and retains a later edge e_j. Since $W_{e_i} \geq W_{e_j}$, replacing e_j by e_i cannot decrease $\sum_{e\in F} \varphi(W_e)$ for any nondecreasing φ. Repeating this exchange reaches E_H. Taking $\varphi(w) = w^r$ shows that the omitted rth-power mass is minimized. Each omitted undirected edge appears twice in the adjacency matrix, giving Eq. (6). The case $r = 1$ maximizes retained edge weight, and $r = 2$ gives the Frobenius statement.

Every simple graph with h vertices has at most $\binom{h}{2}$ edges, so $q \leq \binom{h}{2}$ implies the lower bound in Eq. (7). The union of q pairs contains at most $2q$ vertices and lies in S, proving the upper bound. Each endpoint is incident to a retained positive edge, hence its retained degree is at least the minimum retained weight. If $h \leq k-1$, then $q \leq \binom{k-1}{2}$, proving the stated contrapositive. Finally, taking longer prefixes of one fixed total edge order makes both edge sets and their endpoint unions nested. □

B.3 Endpoint-coverage statement and proof

Let $\mathcal{E}^+ = \{\{u, v\} \in \mathcal{E} : u \prec v,\ W_{uv} > 0\}$, $K = \lfloor s^2/4 \rfloor$, and let $E_K^\star$ be the first $\min\{K, |\mathcal{E}^+|\}$ edges of $\mathcal{E}^+$ under the fixed global order. For $u \in T_a$, put $P_K(u) = \{v : \{u, v\} \in E_K^\star\}$ and $M_{uj} = |P_K(u) \cap T_j|$, and write $j(u)$ for the unique stratum containing u. Assuming $1 \leq s_j \leq N_j$, define

$$q_u^{\mathrm{prot}} = \frac{\binom{N_a-1-M_{ua}}{s_a-1}}{\binom{N_a-1}{s_a-1}} \prod_{j\neq a} \frac{\binom{N_j-M_{uj}}{s_j}}{\binom{N_j}{s_j}}, \qquad \chi_c = \sum_{u\in C_c^{\mathrm{ref}}} \frac{s_{j(u)}}{N_{j(u)}} q_u^{\mathrm{prot}}. \tag{21}$$

Infeasible numerator coefficients are interpreted as zero.

Proposition B.2 (Explicit endpoint-coverage bound)**.** *Let $Y_c = |V_H \cap C_c^{\mathrm{ref}}|$ and $F_c(t) = \Pr[X_c \leq t]$. For $s_0 \geq 1$ and arbitrary integers $r_c \geq 0$,*

$$\Pr\Big[\min_c Y_c < s_0\Big] \leq \delta_{\mathrm{cov}}(\mathbf{r}) := \min\left\{1, \sum_{c=1}^{k} \left[F_c(s_0 + r_c - 1) + \min\left\{1, \frac{\chi_c}{r_c+1}\right\}\right]\right\}. \tag{22}$$

Moreover, with $\psi(x) = 1 - x + x\log x$ and $\mu_c > 0$,

$$F_c(t) \leq \exp[-\mu_c \psi(t/\mu_c)] \quad (0 \leq t \leq \mu_c), \qquad \Pr[X_c = 0] \leq e^{-\mu_c}.$$

Thus $s_0 = 1, r_c = 0$ yields Eq. (14)*.*

Proof. Write $X_{jc} = |S \cap T_j \cap C_c^{\mathrm{ref}}|$. Independent sampling across strata gives

$$X_{jc} \sim \mathrm{Hypergeom}(N_j, M_{jc}, s_j), \qquad X_c = \sum_j X_{jc},$$

and hence the displayed formula for μ_c. The probability of no reference cell-c vertex in the sample is

$$\Pr[X_c = 0] = \prod_j \frac{\binom{N_j-M_{jc}}{s_j}}{\binom{N_j}{s_j}} \leq \exp\left(-\sum_j \frac{s_j M_{jc}}{N_j}\right) = e^{-\mu_c}. \tag{23}$$

The sampling indicators within each stratum are negatively associated, and independent unions preserve this property. Thus, for the nonpositive exponential parameter used in a lower-tail bound, their product Laplace transform is bounded by the independent-indicator counterpart (Serfling, 1974). Exponential Markov optimization then gives, for $\mu_c > 0$ and $0 \leq t \leq \mu_c$,

$$\Pr[X_c \leq t] \leq \exp[-\mu_c\{1 - t/\mu_c + (t/\mu_c)\log(t/\mu_c)\}],$$

with the continuous convention $0 \log 0 = 0$.

It remains to bridge S to V_H. If $e \in E_K^\star$ and both endpoints of e are sampled, fewer than K sampled edges can precede e in the restricted global order. Thus $e \in E_H$, including the case in which fewer than K sampled positive edges exist. Conditional on $u \in T_a$ being sampled, Eq. (21) is exactly the probability that no protected neighbor of u is sampled: the same stratum contributes a draw of $s_a - 1$ from $N_a - 1$ remaining vertices, while every other stratum retains its original quota. Therefore

$$\Pr[u \in S \setminus V_H] \leq \frac{s_a}{N_a} q_u^{\mathrm{prot}}.$$

For $D_c = |(S \setminus V_H) \cap C_c^{\mathrm{ref}}|$, linearity of expectation gives $\mathbb{E}D_c \le \chi_c$, while $Y_c = X_c - D_c$. For every integer $r_c \ge 0$,

$$\{Y_c < s_0\} \subseteq \{X_c \le s_0 + r_c - 1\} \cup \{D_c \ge r_c + 1\}.$$

Markov's inequality bounds the second probability by $\min\{1, \chi_c/(r_c+1)\}$. Summing over c proves Eq. (22); Eq. (23) gives the stated $s_0 = 1$ specialization. No independence between X_c and D_c was used. □

B.4 Proof of Lemma 5.3

Proof. Let

$$\mathbf{P}_R = \mathbf{D}_R^{-1/2}\mathbf{A}_R\mathbf{D}_R^{-1/2}, \quad \mathbf{B} = \mathbf{D}_R^{-1/2}\mathbf{R}_H\mathbf{D}_R^{-1/2}, \quad \mathbf{T} = \mathbf{D}_R^{1/2}\mathbf{D}_H^{-1/2}.$$

Because both degree matrices are diagonal on V_H,

$$\mathbf{P}_H = \mathbf{D}_H^{-1/2}\mathbf{A}_H\mathbf{D}_H^{-1/2} = \mathbf{T}(\mathbf{P}_R - \mathbf{B})\mathbf{T}.$$

For any $\mathbf{x}$,

$$\begin{aligned}|\mathbf{x}^\top\mathbf{B}\mathbf{x}| &\le \frac{1}{2}\sum_{u,v}(R_H)_{uv}\left(\frac{x_u^2}{d_R(u)} + \frac{x_v^2}{d_R(v)}\right)\\ &= \sum_u \frac{r_u}{d_R(u)}x_u^2 \le \rho_H\|\mathbf{x}\|_2^2,\end{aligned}$$

so $\|\mathbf{B}\|_2 \le \rho_H$. Moreover,

$$\|\mathbf{T}\|_2 \le (1-\rho_H)^{-1/2}, \qquad \|\mathbf{T} - \mathbf{I}\|_2 \le (1-\rho_H)^{-1/2} - 1.$$

Since $\|\mathbf{P}_R\|_2 \le 1$,

$$\begin{aligned}\|\mathbf{T}\mathbf{P}_R\mathbf{T} - \mathbf{P}_R\|_2 &\le \|(\mathbf{T}-\mathbf{I})\mathbf{P}_R\mathbf{T}\|_2 + \|\mathbf{P}_R(\mathbf{T}-\mathbf{I})\|_2\\ &\le \frac{\rho_H}{1-\rho_H},\end{aligned}$$

and $\|\mathbf{T}\mathbf{B}\mathbf{T}\|_2 \le \rho_H/(1-\rho_H)$. Therefore

$$\|\mathbf{P}_H - \mathbf{P}_R\|_2 \le \frac{2\rho_H}{1-\rho_H}.$$

Because $\mathcal{L} = \mathbf{I} - \mathbf{P}$, this is the same bound for η_H. Both normalized Laplacians lie between $\mathbf{0}$ and $2\mathbf{I}$, which also gives $\eta_H \le 2$. □

B.5 Discarded-weight and active-graph certificates

Throughout this subsection, assume $|V_H| \ge k$.

Corollary B.3 (Discarded-weight certificate)**.** *Let* $\Omega_H = \sum_{e \in E(G_R)\setminus E_H} W_e$ *and* $d_{\min}^R = \min_{u \in V_H} d_R(u)$. *Then*

$$\rho_H \le \frac{\Omega_H}{d_{\min}^R}.$$

If $\theta_H := \Omega_H/d_{\min}^R$ *satisfies* $\theta_H < 1$, *then it is a directly computable certificate for* $\eta_H \le \min\{2, 2\theta_H/(1-\theta_H)\}$.

Proposition B.4 (Uniform active-graph NCut distortion)**.** *When* $|V_H| \ge k$, *for every* $C \in \mathcal{P}_k(V_H)$,

$$J_{G_R}(C) - k\rho_H \le J_H(C) \le \frac{J_{G_R}(C)}{1-\rho_H}. \tag{24}$$

Consequently, if $Q \in \mathcal{P}_k(V_H)$, $a \ge 1$, $\xi \ge 0$, *and* $J_H(Q) \le a\min_{C\in\mathcal{P}_k(V_H)} J_H(C) + \xi$, *then*

$$J_{G_R}(Q) \le \frac{a}{1-\rho_H}\,\mathrm{OPT}_k(G_R) + \xi + k\rho_H. \tag{25}$$

Proof of Corollary B.3. For each u, the removed incident mass r_u is the weight of a subset of all removed undirected edges, hence $r_u \le \Omega_H$. Therefore $r_u/d_R(u) \le \Omega_H/d_{\min}^R$ uniformly in u. Substitution into Lemma 5.3 proves the certificate. □

Proof of Proposition B.4. For any nonempty $A \subseteq V_H$, the definition of ρ_H gives

$$(1-\rho_H)\,\mathrm{vol}_{G_R}(A) \le \mathrm{vol}_H(A) \le \mathrm{vol}_{G_R}(A).$$

The deleted weight crossing A is at most $\sum_{u\in A} r_u \le \rho_H \,\mathrm{vol}_{G_R}(A)$, so

$$\mathrm{cut}_{G_R}(A, V_H \setminus A) - \rho_H \,\mathrm{vol}_{G_R}(A) \le \mathrm{cut}_H(A, V_H \setminus A) \le \mathrm{cut}_{G_R}(A, V_H \setminus A).$$

Let $\phi_R(A) = \mathrm{cut}_{G_R}(A, A^c)/\,\mathrm{vol}_{G_R}(A)$. If $\phi_R(A) - \rho_H \ge 0$, the upper volume bound gives the desired lower ratio; if it is negative, nonnegativity of the NCut term gives it trivially. The lower volume bound gives the upper ratio. Hence, cell by cell,

$$\max\{0, \phi_R(A) - \rho_H\} \le \frac{\mathrm{cut}_H(A, A^c)}{\mathrm{vol}_H(A)} \le \frac{1}{1-\rho_H}\phi_R(A).$$

Dropping the maximum and summing proves Eq. (24). If $C_R^\star$ minimizes J_{G_R}, then

$$\min_{C\in\mathcal{P}_k(V_H)} J_H(C) \le J_H(C_R^\star) \le \mathrm{OPT}_k(G_R)/(1-\rho_H).$$

Combining this with $J_{G_R}(Q) \le J_H(Q) + k\rho_H$ proves Eq. (25). □

B.6 Proof of Lemma 5.4

Proof. Weyl's inequality gives $|\lambda_i(\mathcal{L}_H) - \lambda_i(\mathcal{L}_R)| \le \eta_H$ for every i. Applying it to indices k and $k+1$ proves $g_H \ge g_R - 2\eta_H$, and summing its first k instances proves the last inequality in Eq. (12).

The gap and perturbation conditions separate the bottom-k eigenspaces. Weyl's inequality also gives $\lambda_{k+1}(\mathcal{L}_R) - \lambda_k(\mathcal{L}_H) \ge g_R - \eta_H$. Davis–Kahan therefore yields $\|\sin\Theta\|_F \le \sqrt{k}\eta_H/(g_R-\eta_H)$, and orthogonal Procrustes alignment adds at most a factor $\sqrt{2}$, proving Eq. (11). Under the optimal alignment, write $\mathbf{U}_H\mathbf{O} = \mathbf{U}_R + \mathbf{E}$ with $\|\mathbf{E}\|_F \le e_{\mathrm{emb}}$. Right multiplication by $\mathbf{O}$ does not change the projection objective. For every C,

$$\begin{aligned}|\Phi_{\mathbf{U}_H}(C) - \Phi_{\mathbf{U}_R}(C)| &\le 2\|(\mathbf{I}-\mathbf{P}_C)\mathbf{U}_R\|_F\|\mathbf{E}\|_F + \|\mathbf{E}\|_F^2 \\ &\le 2\sqrt{k}\,e_{\mathrm{emb}} + e_{\mathrm{emb}}^2.\end{aligned}$$

This is Eq. (13) and is uniform because the bound does not depend on C. □

B.7 Tolerance event and discrete stability

On E_{cov}^H, write $A_c = V_H \cap C_c^{\mathrm{ref}}$. For $C = (C_1, \dots, C_k) \in \mathcal{P}_k(V_H)$, let $y_C(u) = c$ for $u \in C_c$, let $\mathcal{S}_k$ be the permutations of $[k]$, and choose the lexicographically first permutation

$$\pi_C \in \arg\min_{\pi\in\mathcal{S}_k} \sum_{u\in V_H} d_G(u)\mathbf{1}\{\pi(y_C(u)) \ne y_u^{\mathrm{ref}}\}.$$

Set $\widehat{y}_C = \pi_C \circ y_C$, $\widehat{A}_c(C) = C_{\pi_C^{-1}(c)}$, and

$$q_H(C) = \sum_{u\in V_H} d_G(u)\mathbf{1}\{\widehat{y}_C(u) \ne y_u^{\mathrm{ref}}\}.$$

Use this fixed alignment throughout and define the uniform-measure discrepancy

$$\theta_c(C) = 1 - \frac{|A_c \cap \widehat{A}_c(C)|}{\max\{|A_c|, |\widehat{A}_c(C)|\}}. \tag{26}$$

Fix $b_H \ge 0$, $\bar{\theta}_{1:k} \in [0,1]^k$, and $\xi \ge 0$ before the random draw, and let

$$\mathcal{G}_R(b_H, \bar{\theta}_{1:k}) = \left\{C : q_H(C) \le b_H \,\mathrm{vol}_G(\mathcal{V}), \theta_c(C) \le \bar{\theta}_c \;\forall c\right\},$$

and, with $C_A = (A_1, \dots, A_k)$, set

$$\kappa_R(b_H, \bar{\theta}_{1:k}) = \min_{C\notin\mathcal{G}_R(b_H,\bar{\theta}_{1:k})} \left[\Phi_{\mathbf{U}_R}(C) - \Phi_{\mathbf{U}_R}(C_A)\right], \tag{27}$$

where the minimum over an empty set is $+\infty$.

Proposition B.5 (Tolerance-aware discrete stability)**.** *On E^H_{cov}, suppose $h \geq k+1$, $g_R > 0$, and $\eta_H < g_R/2$. If*

$$\kappa_R(b_H, \bar{\theta}_{1:k}) > 2d_{\mathrm{emb}} + \xi$$

and the realized ordered solver output C^{raw}_H satisfies

$$\Phi_{\mathbf{U}_H}(C^{\mathrm{raw}}_H) \leq \min_{C \in \mathcal{P}_k(V_H)} \Phi_{\mathbf{U}_H}(C) + \xi,$$

then $C^{\mathrm{raw}}_H \in \mathcal{G}_R(b_H, \bar{\theta}_{1:k})$.

The proposition permits active error through a tolerance profile: the same-vertex reference embedding $\mathbf{U}_R$ of $G[V_H]$ need only separate partitions outside the admissible error set. Exact active recovery is the special case $b_H = 0$ and $\bar{\theta}_c = 0$ for every c. On $E^H_{\mathrm{cov}} \cap \{h \geq k+1\}$, define

$$\begin{aligned} E^H_{\mathrm{core}} = \{ & g_R > 0,\ \eta_H < g_R/2, \\ & \kappa_R(b_H, \bar{\theta}_{1:k}) > 2d_{\mathrm{emb}} + \xi, \\ & \Phi_{\mathbf{U}_H}(C^{\mathrm{raw}}_H) \leq \min_{C \in \mathcal{P}_k(V_H)} \Phi_{\mathbf{U}_H}(C) + \xi\}. \end{aligned} \tag{28}$$

Set E^H_{core} false outside $E^H_{\mathrm{cov}} \cap \{h \geq k+1\}$. After solving, apply $\pi_{C^{\mathrm{raw}}_H}$ to the active cells and to every label produced by the lift. Write the aligned active partition as $\widehat{C}_H = (\widehat{A}_1, \ldots, \widehat{A}_k)$, where $\widehat{A}_c := \widehat{A}_c(C^{\mathrm{raw}}_H)$, and retain $\widehat{C}$ for the aligned full output. This relabeling leaves all NCut values unchanged.

B.8 Proof of Proposition B.5

Proof. The proof of Lemma 5.4 establishes the uniform bound

$$|\Phi_{\mathbf{U}_H}(C) - \Phi_{\mathbf{U}_R}(C)| \leq d_{\mathrm{emb}} := 2\sqrt{k}\, e_{\mathrm{emb}} + e^2_{\mathrm{emb}} \quad \text{for every } C.$$

Because $C_A \in \mathcal{P}_k(V_H)$ on E^H_{cov}, solver accuracy gives $\Phi_{\mathbf{U}_H}(C^{\mathrm{raw}}_H) \leq \Phi_{\mathbf{U}_H}(C_A) + \xi$. Conversely, for every $C \notin \mathcal{G}_R(b_H, \bar{\theta}_{1:k})$,

$$\Phi_{\mathbf{U}_H}(C) - \Phi_{\mathbf{U}_H}(C_A) \geq \kappa_R(b_H, \bar{\theta}_{1:k}) - 2d_{\mathrm{emb}} > \xi.$$

No such C can be the returned partition. Hence $C^{\mathrm{raw}}_H \in \mathcal{G}_R(b_H, \bar{\theta}_{1:k})$. This argument does not require C_A to be the unique global minimizer; it only requires separation from the complement of the declared tolerance set. □

B.9 Proofs of Lemmas 5.5 and 5.6

Proof of Lemma 5.5. Let μ_A and $\mu_{\widehat{A}}$ be the uniform measures on two nonempty finite sets A and $\widehat{A}$. Directly summing their pointwise difference gives

$$d_{\mathrm{TV}}(\mu_A, \mu_{\widehat{A}}) = 1 - \frac{|A \cap \widehat{A}|}{\max\{|A|, |\widehat{A}|\}}.$$

For any real function f on $A \cup \widehat{A}$, subtracting its minimum and using the variational characterization of total variation yields

$$|\mathbb{E}_{\mu_A} f - \mathbb{E}_{\mu_{\widehat{A}}} f| \leq (\max f - \min f) d_{\mathrm{TV}}(\mu_A, \mu_{\widehat{A}}).$$

Taking $A = A_c$, $\widehat{A} = \widehat{A}_c$, and $f(v) = W_{uv}$ proves Eq. (16). On the tolerance event its right-hand side is at most $\Omega^H_u \bar{\theta}_c$. Thus, for $y = y^{\mathrm{ref}}_u$ and every $c \neq y$, the lifted score difference is at least

$$a^H_{uy} - a^H_{uc} - \Omega^H_u(\bar{\theta}_y + \bar{\theta}_c) \geq m^H_u - \bar{r}^H_u.$$

Outside $B^H_{\Delta,\mathrm{rob}}$ this is at least $\Delta > 0$, so ties cannot alter the reference label. □

Proof of Lemma 5.6. Changing a set on a symmetric difference of volume t_c changes both its volume and its cut weight by at most t_c. With $\phi_c = \mathrm{cut}_G(C_c, C_c^{\mathrm{c}})/v_c$ and $\widetilde{\phi}_c = \mathrm{cut}_G(\widetilde{C}_c, \widetilde{C}_c^{\mathrm{c}})/\mathrm{vol}_G(\widetilde{C}_c)$, this gives

$$\frac{\max\{0, \phi_c - r_c\}}{1 + r_c} \leq \widetilde{\phi}_c \leq \frac{\phi_c + r_c}{1 - r_c}.$$

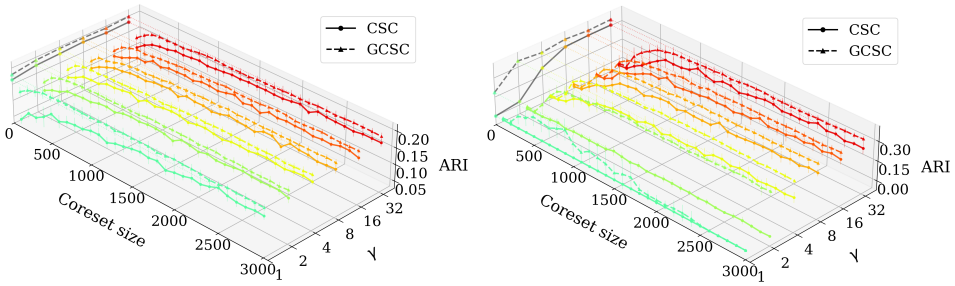


Figure 5: Parameter sensitivity of GCSC and CSC on Letters (left) and MNIST (right). Each curve varies the affinity parameter and coreset size; points and error bars show the mean and variability across repeated trials.

Summing, using $r_c \le r$, gives the two objective bounds in Eq. (17). Since $0 \le \phi_c \le 1$, the positive increase of cell c is at most $r_c(1+\phi_c)/(1-r_c) \le 2r_c/(1-r)$; its positive decrease is at most $2r_c$. Summing proves the absolute bound.

Finally, a vertex in the fixed-alignment disagreement set leaves exactly one reference cell and enters exactly one predicted cell. Its degree is therefore counted in exactly two cell symmetric differences, whereas an agreed vertex is counted in none. Hence $\sum_c t_c = 2\,\mathrm{vol}_G(M)$. □

B.10 Proof of Theorem 5.7

Proof. On $E^H_{\mathrm{cov}} \cap E^H_{\mathrm{core}}$, the aligned active disagreement set has volume $q_H(C^{\mathrm{raw}}_H) \le b_H\,\mathrm{vol}_G(\mathcal{V})$ and all uniform-measure discrepancies satisfy $\theta_c(C^{\mathrm{raw}}_H) \le \bar{\theta}_c$. Lemma 5.5 then shows that, on E^H_{mar}, every remaining disagreement lies in $B^H_{\Delta,\mathrm{rob}}$. Thus the full fixed-alignment disagreement set M satisfies

$$q := \mathrm{vol}_G(M) \le (b_H + b_L)\,\mathrm{vol}_G(\mathcal{V}) = b\,\mathrm{vol}_G(\mathcal{V}).$$

For $t_c = \mathrm{vol}_G(\widehat{C}_c \triangle C^{\mathrm{ref}}_c)$ and $v_c = \mathrm{vol}_G(C^{\mathrm{ref}}_c)$, each symmetric difference is contained in M, so

$$r := \max_c \frac{t_c}{v_c} \le \frac{b}{\beta} = \rho, \qquad R := \sum_c \frac{t_c}{v_c} \le \frac{\sum_c t_c}{\beta\,\mathrm{vol}_G(\mathcal{V})} = \frac{2q}{\beta\,\mathrm{vol}_G(\mathcal{V})} \le 2\rho.$$

Applying Lemma 5.6 with $C = C^{\mathrm{ref}}$ gives all three inequalities in Eq. (18). Proposition B.2 bounds the first failure event, and a union bound over the coverage, core, and margin failures gives probability at least $1 - \delta_{\mathrm{tot}}$. No independence among these events is required. The upper inequality has the form in Definition 5.2 with the stated α and ζ. □

C Additional Experimental Results

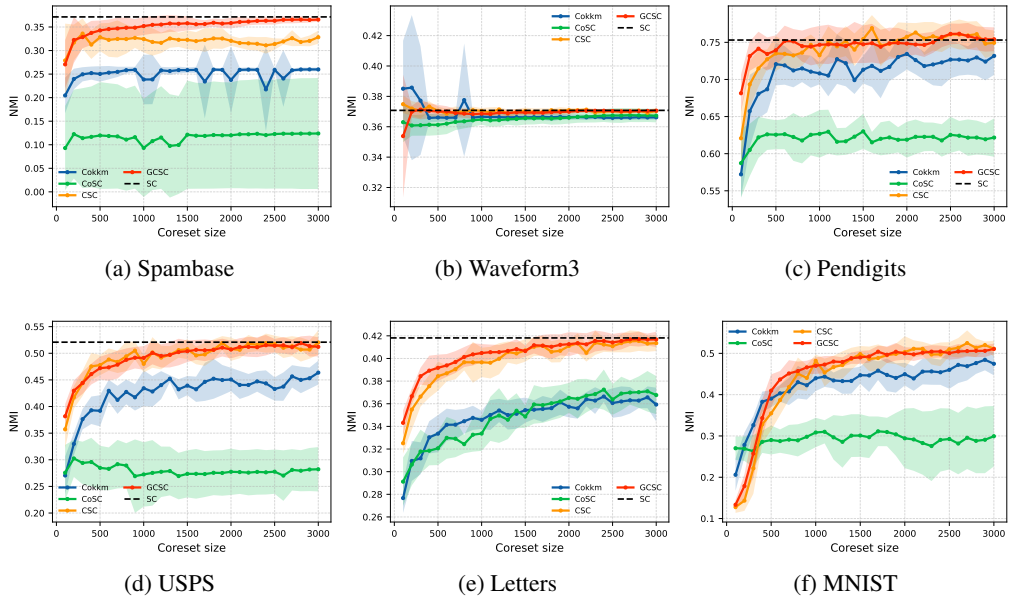
C.1 Parameter settings

The common search grid is

$$\gamma \in \{1/32, 1/16, 1/8, 1/4, 1/2, 1, 2, 4, 8, 16, 32\}.$$

For every coreset method–dataset pair, we select the value maximizing ten-trial mean ARI at $p = 3000$ and then hold it fixed across the displayed budget curve. In the order Letters, MNIST, PenDigits, Spambase, USPS, and Waveform3, the selected values are GCSC: $(32, 8, 1, 32, 16, 1/2)$; CSC: $(32, 16, 1, 8, 32, 16)$; CoSC: $(1/32, 1/4, 1/8, 1/16, 1/32, 1/16)$; and Cokkm: $(4, 1/32, 2, 1/16, 1/32, 1/16)$. Full SC is a single-run reference selected on the same ARI criterion, with values $(2, 1, 8, 32, 2)$ for Letters, PenDigits, Spambase, USPS, and Waveform3; it is unavailable for MNIST. The NCut comparison instead supplies every method with GCSC's selected graph on each dataset.

Figure 5 restores the two largest-dataset sensitivity study. On Letters, GCSC remains above CSC across the displayed parameter range. On MNIST, both methods deteriorate at the smallest setting, whereas GCSC is more stable and consistently stronger once the parameter is at least 2.



(a) Spambase (b) Waveform3 (c) Pendigits

(d) USPS (e) Letters (f) MNIST

Figure 6: NMI across coreset sizes.

C.2 Additional metrics and runtime

Figure 6 gives the NMI counterpart to the main-text ARI comparison. Figure 7 shows total runtime comparable to CSC and substantially below Full SC on the larger datasets. The spectral solve and affinity lift can dominate despite lightweight graph-domain coreset construction, consistent with Eq. (19); construction-only time is reported in main-text Figure 4. The coreset size p is a user-specified computational budget and is not selected with evaluation labels.

C.3 Dataset characteristics

Table 3: Benchmark characteristics.

	Spambase	Waveform3	PenDigits	USPS	Letters	MNIST
Points	4,601	5,000	10,992	11,000	20,000	70,000
Features	57	21	16	256	16	784
Clusters	2	3	10	10	26	10

C.4 Comparison with individual baselines

Table 4 shows substantial gains over CoSC and Cokkm on most datasets, consistent with the comparison against the strongest baseline in Table 1.

C.5 Propagation-rule ablation

At $p = 600$, maximizing GCSC's ten-trial mean ARI over the same grid selects 2 (Letters), 1 (PenDigits), 4 (Spambase), 32 (USPS), and $1/2$ (Waveform3). In the two oracle rows, sampled vertices receive their benchmark ground-truth labels solely to isolate the lift; the end-to-end row instead uses labels obtained by SC on the unpruned sampled graph.

Average affinity removes the systematic representative-count advantage of the sum rule. Its largest gain occurs under explicit imbalance; nearest-core assignment remains stronger in some natural-oracle cases, so the result supports an imbalance-robust default rather than universal optimality.

Table 4: ARI improvements over each coreset baseline at $p = 3000$, with Holm-adjusted two-sided p-values.

Dataset	Comparator	ΔARI	Holm p
Letters	CSC	+.019	.003
	CoSC	+.058	.010
	Cokkm	+.054	$< .001$
MNIST	CSC	+.064	$< .001$
	CoSC	+.222	$< .001$
	Cokkm	+.078	.001
PenDigits	CSC	+.023	.318
	CoSC	+.183	.001
	Cokkm	+.049	.011
Spambase	CSC	+.069	$< .001$
	CoSC	+.306	$< .001$
	Cokkm	+.108	$< .001$
USPS	CSC	+.001	1.000
	CoSC	+.247	$< .001$
	Cokkm	+.086	.001
Waveform3	CSC	-.001	.566
	CoSC	-.002	1.000
	Cokkm	-.001	.007

Table 5: Mean ARI of three propagation rules over five datasets and both sampling schemes, with three trials per setting. Minority stress retains only 25% of the sampled minority representatives.

Core-label setting	Average	Sum	Nearest
Oracle, natural sample	.637	.628	.641
Oracle, minority at 25%	.614	.473	.581
Unpruned spectral core	.373	.371	.358

C.6 Fixed edge-budget sensitivity

The chosen fixed budget is adequate for Waveform3 but removes too much active structure on several other graphs. The unpruned sampled graph attains mean ARI/NMI-to-Full-SC of $.847/.869$ and a mean signed relative NCut gap of $-.008$. No single larger constant multiple of the default edge count was uniformly best across these datasets. The appropriate conclusion is that the budget controls a graph-dependent approximation error, not that a small k in the original kNN graph is intrinsically invalid.

Across these trials, $G[V_H]$ is connected and every Full-SC reference cell is represented, whereas H is disconnected and the sufficient condition $\eta_H < g_R/2$ is not met. This localizes the observed limitation to edge retention rather than missing active cells. The condition is sufficient, not necessary; nevertheless, its failure prevents invoking Theorem 5.7 for these realizations.

C.7 Assumption diagnostics

The full-graph k-gap is $\lambda_{k+1}(\mathcal{L}_G) - \lambda_k(\mathcal{L}_G)$ for the normalized Laplacian of the full graph at the listed parameter. Nonpositive-margin volume is the fraction of graph volume carried by vertices with non-positive reference-label affinity margin in this retrospective diagnostic. It omits Δ and the active-core radius $\bar{r}_u^H$ in Eq. (15), so it neither estimates nor verifies E_{mar}^H; it is not an input to GCSC.

Degree has weak label information on these graphs, yet the Full-SC fidelity in Table 2 remains high on most datasets. The label-based columns are diagnostic only and are not used by GCSC. Letters

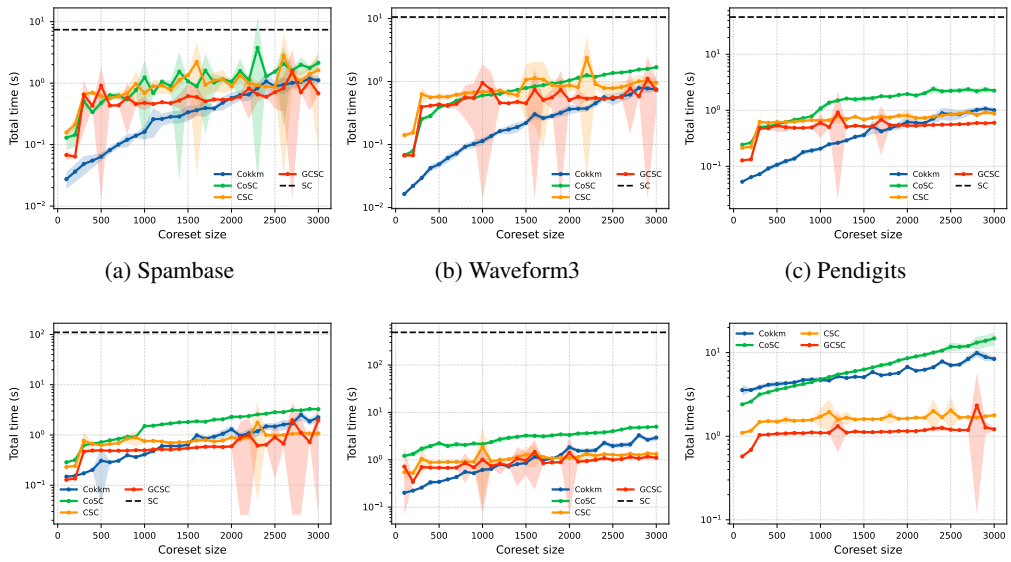
Table 6: Edge-budget sensitivity at $p = 3000$ using GCSC's selected affinity setting for each dataset. Each cell reports ARI-to-Full-SC / signed relative NCut gap, averaged over three trials. The default budget is $1\times$ and $G[S]$ is unpruned.

Dataset	$1\times$	$2\times$	$4\times$	$8\times$	$G[S]$
Letters	.364/.259	.426/.165	.532/.065	.666/.028	.666/.028
PenDigits	.474/1.573	.734/.377	.727/.098	.741/.456	.928/-.059
Spambase	.049/.721	-.003/.788	.013/.782	.034/.782	.940/-.007
USPS	.426/.186	.448/.127	.497/.126	.491/.065	.737/-.005
Waveform3	.906/.036	.927/.014	.943/.008	.955/.005	.968/.002

Table 7: Same-vertex active-core diagnostic at $p = 3000$. Here g_R is the eigengap of $G[V_H]$ and $\eta_H = \|\mathcal{L}_H - \mathcal{L}_{G[V_H]}\|_2$. Values use GCSC's selected affinity setting for each dataset and are averaged over three trials.

Dataset	$\lvert V_H\rvert/p$	Components of H	g_R	η_H	ARI$(H, G[V_H])$
Letters	.954	50.8	.009	1.085	.314
PenDigits	.896	82.1	.028	1.033	.422
Spambase	.556	252.1	.171	1.158	.033
USPS	.724	69.7	.023	1.084	.389
Waveform3	.925	4.2	.174	.968	.904

has the smallest full-graph gap and the largest nonpositive-margin volume, consistent with its greater partition variability.



(a) Spambase (b) Waveform3 (c) Pendigits

(d) USPS (e) Letters (f) MNIST

Figure 7: Total algorithm runtime after common graph construction across coreset sizes.

Table 8: Retrospective graph diagnostics at the best observed affinity parameter. The full-graph gap and nonpositive-margin volume are contextual and do not verify the events in Theorem 5.7.

Dataset	Full-graph k-gap	Degree–label NMI	Nonpositive-margin volume
Spambase	.154	.010	.116
Waveform3	.171	.003	.150
PenDigits	.030	.127	.044
USPS	.019	.065	.118
Letters	.004	.050	.329

D LIMITATIONS

Although GCSC is highly efficient during coreset construction, the subsequent spectral solution on the coreset and affinity-based label propagation can dominate total runtime. Combining GCSC with sparse-graph spectral solvers and more scalable label propagation is therefore a promising direction for fully leveraging its construction speed.