# Sparsifying Correlation Clustering: Edge Coresets, Triangle Witnesses, and Observation Lower Bounds

**Ibne Farabi Shihab**[1*] **Sanjeda Akter**[1*] **Anuj Sharma**[2]

[1]Department of Computer Science
[2]Department of Civil, Construction and Environmental Engineering
Iowa State University, Ames, IA 50011, USA
{ishihab, sanjeda, anujs}@iastate.edu

## Abstract

The standard metric LP for correlation clustering has $\Theta(n^2)$ pairwise marginals and $\Theta(n^3)$ triangle inequalities, while dense pairwise information is often the bottleneck in large relational pipelines. We separate three sparsification questions that are often treated together and show that they behave differently. First, to preserve the objective of every clustering, we prove that the signed-edge disagreement class induced by all partitions of $n$ vertices has VC dimension exactly $n-1$, so weighted edge sampling yields additive $\varepsilon$-coresets of size $\tilde{O}(n/\varepsilon^2)$ with optimal dependence on $n$. Second, assuming the signed weighted input and a feasible LP point are available, we ask whether LP-PIVOT rounding can run from only sampled LP marginals. Our algorithm, SPARSE-LP-PIVOT, fills each missing marginal from triangle witnesses, whose LP-feasible intervals give a deterministic per-pair error bound. Every Lipschitz LP-PIVOT rule then has an unconditional but coarse guarantee; a sharper guarantee, governed by edge-weighted interval widths, holds under a rule-specific influence certificate, for which we give a checkable sufficient condition but which no nontrivial rule is yet known to satisfy. Under uniform sampling, typical pairs acquire a witness once $m = \Theta(n^{3/2})$ marginals are sampled and all pairs at $m = O(n^{3/2}\sqrt{\log n})$, although witnesses alone do not guarantee narrow intervals. Third, when the signs and weights themselves are observed only through $o(n)$ uniform samples, every algorithm has expected approximation ratio $\Omega(n^{3/2})$ on some general weighted instance. Experiments illustrate the $n^{3/2}$ witness scaling, show the sparse-rounding gap on stochastic block models closing as the budget passes this scale, and measure $4.7$–$11.6\times$ lower post-LP time ($1.47$–$1.51\times$ end to end) when the dense LP solve is shared.

## 1 Introduction

Correlation clustering (CC) [6] clusters data from pairwise similarity and dissimilarity judgments. Given a complete signed graph, the objective is to minimize the total weight of positive edges cut by the clustering plus negative edges placed inside clusters. The problem is central in entity resolution [13], community detection [24], duplicate detection [22], and relational learning, where pairwise labels are natural but dense pairwise supervision is expensive. The classical PIVOT algorithm gives a simple constant-factor approximation [3], LP rounding improves the approximation factor for the standard relaxation [12], preclustering combined with a stronger relaxation improves it further [16], and recent work based on stronger cluster-LP relaxations pushes the best known ratios below $1.5$ [10]; for pseudometric-weighted CC, recent LP-PIVOT analyses give a tight $10/3$ guarantee

*Equal contribution.

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

within a standard LP-rounding framework [19]. These guarantees, however, are designed for dense pairwise information. The standard metric LP has $\binom{n}{2}$ variables and $3\binom{n}{3}$ triangle constraints, and dense pairwise supervision is often the primary bottleneck in large relational pipelines. Metric embeddings [18] and ultrametric approximations [11] provide alternative structural views of the LP geometry but do not address sparsification directly.

This paper asks a deliberately narrower question than full sublinear-time CC: *which parts of LP-based correlation clustering survive sparsification, and under which access model?* Query-efficient CC with noisy oracles [23] addresses a related but distinct question where the algorithm adaptively queries pairwise labels, while our model assumes that the labels are known but LP marginals are sparse. We separate three issues that are often conflated. Edge coresets ask whether a sampled weighted edge set approximates the objective of every clustering. Sparse LP-marginal rounding asks whether an LP-based rounding algorithm can operate when only a sparse subset of LP marginals is available. Sparse edge observation asks whether the input signs and weights themselves can be recovered well enough from a small random sample. Our main results show that these regimes behave differently, and the conceptual message is that triangle inequalities allow *information propagation* only after the sampled graph contains enough length-two witnesses. A pair $(u, v)$ has a triangle witness when there exists $z$ such that both sampled marginals $x_{uz}$ and $x_{vz}$ are known, and uniform sampling develops such witnesses at the normalized scale $m/n^{3/2}$. Witness existence alone is not enough for a multiplicative approximation because the witness interval may still be wide, so we introduce a computable diagnostic, $\overline{\Gamma}_w$, measuring the weighted average width of the tightest observed triangle interval for each pair (formally defined in Section 4).

**Contributions.** Each result lives in one access model (Table 1). For each, we state what is new and which standard tool it uses.

- **Exact complexity of edge coresets (Section 3).** The disagreement class induced by all clusterings of an $n$-vertex signed complete graph has VC dimension exactly $n - 1$ (Theorem 3.1). The lower bound shatters a star; the upper bound uses the fact that any $n$ edges on $n$ vertices contain a cycle, on which transitivity of the same-cluster relation forbids exactly one cut edge. Standard VC uniform convergence, applied once this exact parameter is known, gives additive $\varepsilon W$ edge coresets of size $\tilde{O}(n/\varepsilon^2)$ under weighted edge sampling, with worst-case optimal dependence on $n$ (Theorem 3.2).
- **Certified completion of missing LP marginals (Section 4).** SPARSE-LP-PIVOT fills a missing marginal $x_{uv}$ from an observed triangle witness $z$, a vertex with both $x_{uz}$ and $x_{vz}$ sampled. The LP triangle inequalities turn each witness into a feasible interval, so the midpoint of the narrowest one carries a deterministic per-pair error certificate $\Gamma_S(u, v)/2$ (Lemma 4.2). Under uniform sampling, typical pairs acquire a witness at $m = \Theta(n^{3/2})$ and all pairs at $m = O(n^{3/2}\sqrt{\log n})$ (Lemma 4.7); witness existence is kept separate from interval quality.
- **A stability interface for sparse rounding (Section 4).** Every Lipschitz LP-PIVOT rule has an unconditional but coarse sparse transfer with additive loss $(LW/2)\sum_{u<v}\Gamma_S(u, v)$ (Proposition 4.3). Theorem 4.5 is a conditional reduction: a rule-specific influence certificate replaces the global charge $W$ by influence weights, and the error term becomes the computable diagnostic $W\overline{\Gamma}_w$ when those weights are dominated by edge weights. Remark 4.6 gives a checkable sufficient condition, and Table 2 records the status of each statement. The common-randomness coupling is standard; what is new is isolating what a concrete rule must prove beyond Lipschitz attachment probabilities.
- **An access-model lower bound (Section 5).** When the input edges themselves are observed only through $m = o(n)$ uniform samples, every algorithm has expected approximation ratio $\Omega(n^{3/2})$ on some general weighted instance (Theorem 5.1). The hard distribution balances cutting positive pairs on an almost-all-positive instance against leaving an unseen heavy negative clique uncut; Yao's principle is standard. The bound is specific to this model and does not rule out adaptive sampling, structural assumptions, graph sketching, or sparse LP-marginal access.

Experiments (Section 6) illustrate the predicted $n^{3/2}$ witness scaling, report $\overline{\Gamma}_w$ as a diagnostic rather than a certificate, and measure post-LP runtime.

Our edge-coreset result is closest in spirit to uniform-convergence coresets for clustering [20], but the exact VC calculation is specific to the transitivity constraints of clusterings. The witness-

Table 1: Access models separated in this paper. The edge-coreset result preserves the objective of every integral clustering; sparse LP rounding assumes known signs and weights but only partial access to a feasible LP solution; sparse edge observation hides the input itself and is therefore strictly more restrictive.

| Model | Signs and weights | LP marginals | Main result |
|---|---|---|---|
| Edge coreset | known through weighted sampling | not used | exact VC dimension, $\tilde{O}(n/\varepsilon^2)$ coreset (Thms. 3.1, 3.2) |
| Sparse LP marginals | known | only $x_{uv}$ for $uv \in S$ | coarse transfer (Prop. 4.3), conditional sharp (Thm. 4.5) |
| Sparse edge observation | only sampled edges observed | not available | $\Omega(n^{3/2})$ ratio for $m = o(n)$ (Thm. 5.1) |

threshold analysis differs from classical cut and spectral sparsification, where the goal is to preserve cut values or quadratic forms [8, 9, 26]. Sublinear, streaming, sketching, de-sparsification, and parallel algorithms for correlation clustering address computational scalability through different access models [4, 7, 2, 15, 17, 5]; our focus is narrower in that we ask which LP-rounding guarantees can be transferred when the signed instance is known but only sparse LP marginals are available. The model is also distinct from neighborhood-sampling methods used for scalable graph representation learning, such as GraphSAGE and Cluster-GCN [21, 14]: here the sampled structure is used to reconstruct LP marginals for a rounding algorithm rather than to train a message-passing model. On the chromatic side, Xiu et al. [28] gave a 2.5-approximation for CCC via LP rounding, and Fan et al. [19] recently improved this to 2.15; our sparsification results apply to the standard binary signed CC LP and do not by themselves imply a CCC sparsification theorem.

## 2 Preliminaries

Let $V$ be a vertex set of size $n$. We write $uv$ for an unordered pair and use the symmetric notation $x_{uv} = x_{vu}$ and $w_{uv} = w_{vu}$ whenever convenient. A weighted signed CC instance is a complete graph on $V$ with a sign map $s_{uv} \in \{+, -\}$ and nonnegative weights $w_{uv} \geq 0$ for every unordered pair $uv$. A clustering $\mathcal{C}$ is a partition of $V$, and we let $\sigma_{\mathcal{C}}(u, v)$ be the indicator that $u, v$ lie in different clusters of $\mathcal{C}$. The CC objective is

$$\mathrm{obj}_G(\mathcal{C}) = \sum_{uv:s_{uv}=+} w_{uv}\sigma_{\mathcal{C}}(u, v) + \sum_{uv:s_{uv}=-} w_{uv}(1 - \sigma_{\mathcal{C}}(u, v)), \tag{1}$$

and we write $W = \sum_{u<v} w_{uv}$ and $\mathrm{OPT} = \min_{\mathcal{C}} \mathrm{obj}_G(\mathcal{C})$. The standard metric LP relaxation introduces variables $x_{uv} \in [0, 1]$ intended to represent whether $u$ and $v$ are separated:

$$\min_x \quad \mathrm{LP}(x) = \sum_{uv:s_{uv}=+} w_{uv}x_{uv} + \sum_{uv:s_{uv}=-} w_{uv}(1 - x_{uv}) \tag{CC-LP}$$
$$\text{s.t.} \quad x_{uv} \leq x_{uz} + x_{vz} \quad \forall \text{ distinct } u, v, z \in V, \qquad 0 \leq x_{uv} \leq 1 \quad \forall u < v.$$

Every integral clustering induces a feasible LP point by setting $x_{uv} = \sigma_{\mathcal{C}}(u, v)$, so $\mathrm{LP}^* \leq \mathrm{OPT}$. A weight function is a pseudometric if $w_{uv} \leq w_{uz} + w_{vz}$ for all distinct $u, v, z$. Pseudometric weights enter only through the exact-marginal guarantee of Fan et al. [19] discussed below; the triangle-imputation mechanism uses the triangle inequalities of the LP variables $x$, not those of the weights, and neither the universal coarse stability proposition nor the conditional sparse-rounding theorem below assumes pseudometric weights.

The three access models used in the paper are summarized in Table 1: positive and negative results live in different models, and the sparse LP-marginal model assumes that the signed weighted instance is known and that a feasible LP point exists, but it does not solve the end-to-end problem of computing all LP marginals quickly. The rounding consumer receives and retains $m$ sampled raw LP marginals rather than $\binom{n}{2}$ values, saving raw-marginal transfer and storage. Building the witness index costs $O(\sum_z d_S(z)^2)$ time, where $d_S(z)$ is the number of sampled pairs containing $z$, and the recursion still considers every active pivot-vertex pair, so we do not claim subquadratic rounding time. The signed input and LP solve remain required, so these access savings alone do not imply lower end-to-end computation or total memory; measured post-LP and end-to-end times are reported in Appendix F.1. The lower bound in Section 5 applies only to the stricter sparse edge-observation model. Several LP-based pivot algorithms choose a pivot $v$, attach $u$ to the pivot cluster with a probability determined by $x_{uv}$ and the sign $s_{uv}$, and recurse, motivating the following abstraction.

**Assumption 2.1** (Exact-marginal LP-PIVOT guarantee)**.** On an instance class $\mathcal{I}$, suppose an LP-PIVOT rounding rule with functions $f^+, f^- : [0,1] \to [0,1]$ satisfies two properties: each $f^s$ is $L$-Lipschitz, and for every instance $G \in \mathcal{I}$ and feasible LP solution $x$, $\mathbb{E}[\mathrm{obj}_G(\mathcal{C}_{\mathrm{exact}})] \le \alpha\,\mathrm{LP}(x)$.

For pseudometric-weighted CC, Fan et al. [19] establish a $10/3$ exact-marginal guarantee. Their published rounding function jumps from 0 to $2/3$ at $t = 0.4$, so it is not Lipschitz: it is not an instance of Assumption 2.1, and neither Proposition 4.3 nor Theorem 4.5 below applies to it. A nontrivial instance is the complete-graph rule of Chawla et al. [12] on signed complete graphs with unit weights, which uses $f^-(t) = t$ and an $f^+$ that rises continuously and piecewise quadratically from 0 at $t = 0.19$ to 1 at $t = 0.5095$; it satisfies Assumption 2.1 with $\alpha = 2.06$ and $L = 2/(0.5095 - 0.19) < 6.3$. Throughout the paper, $f^s(t)$ denotes the probability that a vertex is separated from the pivot across an edge of sign $s$ whose LP value is $t$, so the attachment probability in Algorithm 1 is $1 - f^s(t)$, and this convention is used consistently in the coupling arguments below.

**Example 2.2** (The three access models on four vertices)**.** Let $V = \{a, b, c, d\}$ have unit weights, positive pairs $ab, cd$, and all other pairs negative. A feasible LP point is

$$(x_{ab}, x_{ac}, x_{ad}, x_{bc}, x_{bd}, x_{cd}) = (0.3, 0.2, 0.5, 0.4, 0.6, 0.4).$$

For $S = \{ac, bc, ad, bd\}$, the narrowest witness intervals for $ab$ and $cd$ are $[0.2, 0.6]$ and $[0.3, 0.7]$, giving midpoint imputations $0.4$ and $0.5$. An edge-coreset draw of these four pairs retains their signs and assigns weight $W/m = 1.5$ per draw. Sparse LP rounding knows all signs and weights but only the four sampled LP values; sparse edge observation sees only the sampled signs and weights and no LP values. This toy draw illustrates information access, not a coreset accuracy guarantee.

## 3 Exact VC Dimension and Edge Coresets

For a clustering $\mathcal{C}$, define the sign-dependent disagreement indicator $h_{\mathcal{C}}(uv) = \sigma_{\mathcal{C}}(u, v)$ if $s_{uv} = +$ and $h_{\mathcal{C}}(uv) = 1 - \sigma_{\mathcal{C}}(u, v)$ if $s_{uv} = -$, and let $\mathcal{H}_G = \{h_{\mathcal{C}} : \mathcal{C} \text{ a partition of } V\}$. Then $\mathrm{obj}_G(\mathcal{C}) = \sum_{uv} w_{uv} h_{\mathcal{C}}(uv)$, which makes the objective a weighted average of indicator values from $\mathcal{H}_G$ and brings the problem squarely within the scope of VC theory.

**Theorem 3.1** (Exact VC dimension)**.** *For every signed complete graph $G$ on $n \ge 3$ vertices,* $\mathrm{VC}(\mathcal{H}_G) = n - 1$.

For the lower bound, fix a center $v_1$ and consider the $n-1$ star edges $v_1 v_i$: for any desired disagreement pattern, place $v_i$ in the cluster of $v_1$ exactly when that realizes the requested value of $h_{\mathcal{C}}(v_1 v_i)$, and otherwise make $v_i$ a singleton; this shatters the star. For the upper bound, any $n$ edges on $n$ vertices contain a cycle, and on a cycle the same-cluster relation is transitive, so it is impossible for exactly one cycle edge to be cut while all others are uncut. Since the sign pattern only XORs this cut vector with a fixed vector, at least one labeling of the cycle edges is unrealizable, hence no $n$ edges are shattered. The full proof is in Appendix A.

**Theorem 3.2** (Additive weighted edge coreset)**.** *Let $G$ be a weighted signed complete graph with total weight $W > 0$. Draw $m$ edges independently with replacement from $p(uv) = w_{uv}/W$, give each sampled edge weight $W/m$, and merge duplicate sampled pairs by summing their sampled weights. If $m \ge C\big((n-1)\log(1/\varepsilon) + \log(1/\delta)\big)/\varepsilon^2$ for a universal constant $C$, then with probability at least $1-\delta$ the sampled weighted instance $H$ satisfies $|\mathrm{obj}_H(\mathcal{C}) - \mathrm{obj}_G(\mathcal{C})| \le \varepsilon W$ simultaneously for all clusterings $\mathcal{C}$.*

The proof is a direct application of VC uniform convergence to Theorem 3.1 (Appendix B). The rate is optimal in $n$ because any uniform convergence guarantee for a class of VC dimension $n-1$ needs $\Omega(n/\varepsilon^2)$ samples in the worst case, and if an $\alpha$-approximation algorithm is run on $H$, its output has cost at most $\alpha\mathrm{OPT} + (\alpha+1)\varepsilon W$ on $G$. Thus the guarantee is additive: a relative overhead of at most $\tau$ above $\alpha$ needs $\varepsilon \le \tau\mathrm{OPT}/((\alpha+1)W)$, so the result is a useful multiplicative guarantee only when $\mathrm{OPT} = \Omega(\varepsilon W)$, and when $\mathrm{OPT} \ll \varepsilon W$ (including $\mathrm{OPT} = 0$) it remains a valid additive bound but implies no finite approximation ratio.

## 4 Sparse LP-Marginal Rounding

The coreset result concerns objectives of integral clusterings and does not use LP geometry. We now move to LP-based rounding, where the object being sparsified is no longer the edge-weighted ob-

**Algorithm 1** SPARSE-LP-PIVOT

**Require:** Active vertex set $U$, signs $s$, weights $w$, sampled marginals $\{x_{ab} : ab \in S\}$, rounding functions $f^+, f^-$
**Ensure:** A clustering of $U$
1: **if** $U = \emptyset$ **then return** $\emptyset$
2: **end if**
3: Choose pivot $v \in U$ uniformly at random; let $C \leftarrow \{v\}$
4: **for** $u \in U \setminus \{v\}$ **do**
5: Compute $\tilde{x}_{uv}$ by min-width triangle imputation; with probability $1 - f^{s_{uv}}(\tilde{x}_{uv})$, add $u$ to $C$
6: **end for**
7: **return** $\{C\} \cup$ SPARSE-LP-PIVOT$(U \setminus C, s, w, S, f^+, f^-)$

jective itself but the set of LP marginals visible to a rounding procedure. In the sparse LP-marginal model, the signed weighted instance $G = (V, s, w)$ is known, and an LP solver or oracle has produced a feasible point $x$ for (CC-LP); the rounding algorithm, however, may directly read only the marginals $\{x_{uv} : uv \in S\}$ for a sampled set $S \subseteq \binom{V}{2}$, and for pairs outside $S$ it must infer a surrogate value $\hat{x}_{uv}$ from the observed marginals. This isolates the information requirement of LP rounding from the separate computational problem of solving the LP, and the mechanism for inference is triangle witnesses.

**Definition 4.1** (Triangle witness and interval)**.** For a pair $uv$ and a third vertex $z$, $z$ is an observed triangle witness for $uv$ if $uz, vz \in S$. Its LP-feasible interval is $L_z(uv) = |x_{uz} - x_{vz}|$ and $U_z(uv) = \min\{x_{uz} + x_{vz}, 1\}$, with width $\gamma_z(uv) = U_z(uv) - L_z(uv)$.

The LP triangle inequalities imply that every feasible $x$ obeys $x_{uv} \in [L_z(uv), U_z(uv)]$ for every witness $z$, so SPARSE-LP-PIVOT uses the midpoint of the tightest observed witness interval. For $uv \notin S$ let $z^*(uv) \in \arg\min_{z:\, uz, vz \in S} \gamma_z(uv)$ when a witness exists, and set $\tilde{x}_{uv} = (L_{z^*}(uv) + U_{z^*}(uv))/2$; if no witness exists, set $\tilde{x}_{uv} = 1/2$, and for observed pairs set $\tilde{x}_{uv} = x_{uv}$. The associated weighted average imputation width is

$$\Gamma_S(u,v) = \begin{cases} 0, & uv \in S, \\ \min\limits_{z:\, uz, vz \in S} \gamma_z(uv), & uv \notin S \text{ and a witness exists}, \\ 1, & uv \notin S \text{ and no witness exists}, \end{cases}$$

$$\overline{\Gamma}_w(S) = \frac{1}{W} \sum_{u<v} w_{uv} \Gamma_S(u,v).$$

**Lemma 4.2** (Imputation error)**.** *For every feasible LP point $x$ and every sampled set $S$, $|\tilde{x}_{uv} - x_{uv}| \le \Gamma_S(u,v)/2$ for all $u < v$.*

*Proof.* If $uv \in S$, the error is zero. If $uv \notin S$ has a witness $z$, LP feasibility gives $x_{uv} \in [L_z(uv), U_z(uv)]$, so the midpoint of the minimum-width interval is within half its width. If no witness exists, $\tilde{x}_{uv} = 1/2$ and $x_{uv} \in [0,1]$, so the error is at most $1/2 = \Gamma_S(u,v)/2$. □

SPARSE-LP-PIVOT (Algorithm 1) runs the standard LP-PIVOT recursion using $\tilde{x}$ in place of $x$ in the attachment probabilities.

The next two results separate two levels of stability. The first is a coarse perturbation bound for every Lipschitz LP-PIVOT rule, which is always valid but can be loose because it charges every marginal perturbation by the total edge weight $W$. The second explains the additional rule-specific certificate needed to recover an error term controlled by the weighted triangle-width diagnostic rather than by the unweighted sum of all imputation widths. Table 2 summarizes the resulting hierarchy.

**Proposition 4.3** (Universal but coarse influence stability)**.** *Consider any LP-PIVOT rule of the form in Algorithm 1, where the separation probability for a pivot-pair $uv$ is $q_{uv}(y) = f^{s_{uv}}(y_{uv})$ and each $f^s$ is $L$-Lipschitz. Then for any two matrices $x, y \in [0,1]^{\binom{V}{2}}$,*

$$\left|\mathbb{E}[\mathrm{obj}_G(\mathcal{C}(y))] - \mathbb{E}[\mathrm{obj}_G(\mathcal{C}(x))]\right| \le LW \sum_{u<v} |y_{uv} - x_{uv}|,$$

Table 2: Logical status of the sparse-rounding guarantees. Each row adds one hypothesis to the row above; rows two to four give the term added to $\alpha\,\mathrm{LP}(x)$ under Assumption 2.1. None of the rows applies to the discontinuous rule of Fan et al. [19].

| Statement | Added hypothesis | Status | Error term |
|---|---|---|---|
| Lemma 4.2 | feasible LP point $x$ | unconditional | $\Gamma_S(u,v)/2$ per pair |
| Proposition 4.3 | $f^+, f^-$ are $L$-Lipschitz | unconditional | $(LW/2)\sum_{u<v}\Gamma_S(u,v)$ |
| Theorem 4.5 | Assumption 4.4, e.g., via Remark 4.6 | conditional | $(KL/2)\,\Gamma_\Phi(S)$ |
| Same-edge form | $\Phi_{uv} \le c_\Phi w_{uv}$ | conditional; not known for any nontrivial rule | $O(L)\,W\overline{\Gamma}_w(S)$ |

*so every such rule satisfies an influence-stability statement with constant $K = 1$ and $\Phi_{uv} = W$.*

The proof (Appendix C) couples the two executions through a common pivot ordering and shared attachment randomness, observes that each unordered pair is queried at most once during the recursion, and bounds the divergence probability by $L\sum_{u<v}|y_{uv} - x_{uv}|$; since two output clusterings can differ by at most $W$ in objective, the displayed bound follows. This universal statement uses $\Phi_{uv} = W$, and replacing $W$ by a tighter influence vector requires extra structural information about the rule.

**Assumption 4.4** (Weighted influence stability)**.** Fix an exact-marginal LP-PIVOT rule satisfying Assumption 2.1. We say that this rule has weighted influence stability with constant $K$ and influence weights $\Phi_{uv} = \Phi_{uv}(G,x) \ge 0$ if, for every surrogate matrix $y \in [0,1]^{\binom{V}{2}}$ used in place of the feasible LP point $x$,

$$\mathbb{E}[\mathrm{obj}_G(\mathcal{C}(y))] \le \mathbb{E}[\mathrm{obj}_G(\mathcal{C}(x))] + KL\sum_{u<v}\Phi_{uv}\,|y_{uv} - x_{uv}|.$$

Here $\mathcal{C}(x)$ denotes the output distribution of the exact-marginal pivot rule and $\mathcal{C}(y)$ denotes the output distribution obtained by running the same rule with $y$ in the attachment probabilities.

The influence weights are part of the rule-specific analysis and need not equal the edge weights, but when a concrete LP-PIVOT analysis proves $\Phi_{uv} \le c_\Phi w_{uv}$ for all pairs, the simpler same-edge diagnostic $\overline{\Gamma}_w$ controls the perturbation through the influence-weighted imputation width $\Gamma_\Phi(S) = \sum_{u<v}\Phi_{uv}\Gamma_S(u,v)$, which then satisfies $\Gamma_\Phi(S) \le c_\Phi W\overline{\Gamma}_w(S)$.

**Theorem 4.5** (Conditional sparse LP-PIVOT under weighted influence stability)**.** *Assume that the exact-marginal pivot rule satisfies Assumption 2.1 with parameters $(\alpha, L)$ on an instance class $\mathcal{I}$ and also has weighted influence stability with constant $K$ and influence weights $\Phi$. Then for every $G \in \mathcal{I}$, every feasible LP point $x$, and every sampled marginal set $S$,*

$$\mathbb{E}[\mathrm{obj}_G(\mathcal{C}_{\mathrm{sparse}})] \le \alpha\,\mathrm{LP}(x) + \frac{KL}{2}\Gamma_\Phi(S).$$

*In particular, if the certificate satisfies $\Phi_{uv} \le c_\Phi w_{uv}$ for all pairs, the perturbation term reduces to $O(L)\,W\overline{\Gamma}_w(S)$, where the hidden constant depends on the certificate but not on $n$, $w$, $x$, or $S$.*

*Remark* 4.6 (A checkable sufficient condition). In the common-randomness coupling used for Proposition 4.3 (Appendix C), let $D_e$ be the event that the first divergent attachment decision occurs on pair $e$, let $\pi_e(x)$ be the probability that $e$ is queried in the exact-marginal execution, and let $B_e \ge 0$ bound $\mathbb{E}[\mathrm{obj}_G(\mathcal{C}(y)) - \mathrm{obj}_G(\mathcal{C}(x)) \mid D_e]$ uniformly over surrogate matrices $y$. Then Assumption 4.4 holds with $K = 1$ and $\Phi_e = \pi_e(x)B_e$. Proposition 4.3 corresponds to $B_e = W$ and $\pi_e(x) \le 1$, and the same-edge form of Theorem 4.5 needs $\pi_e(x)B_e = O(w_e)$.

Bounding $B_e$ itself by $O(w_e)$ is usually too strong: if the first divergence moves $u$ out of the pivot's cluster $C$, every pair between $u$ and $C$ changes status, so $B_e$ can be as large as the weight between $u$ and $C$. The exposure probability $\pi_e(x)$ can offset this, because inside a cluster only the pairs containing its pivot are ever queried, but proving $\pi_e(x)B_e = O(w_e)$ requires controlling the downstream cascade after a changed attachment. We do not know such a bound for any nontrivial LP-PIVOT rule, including the Lipschitz rule of Chawla et al. [12]; we expect it to be provable first for restricted variants, such as component-local or bounded-depth recursions in which one changed

attachment affects only a region of controlled weight. Theorem 4.5 is therefore a conditional reduction, while Proposition 4.3 holds unconditionally for every Lipschitz rule.

The proof (Appendix C) runs the sparse algorithm with surrogate $y = \tilde{x}$, applies Lemma 4.2 to bound $|\tilde{x}_{uv} - x_{uv}|$ by $\Gamma_S(u,v)/2$, and then invokes the influence-stability hypothesis. Specialising to the universal certificate $K = 1$ and $\Phi_{uv} = W$ yields the unconditional coarse corollary $\mathbb{E}[\mathrm{obj}_G(\mathcal{C}_{\mathrm{sparse}})] \le \alpha\,\mathrm{LP}(x) + (LW/2)\sum_{u<v}\Gamma_S(u,v)$. It improves on the trivial objective bound $W$ only when

$$\alpha\,\mathrm{LP}(x) + \frac{LW}{2}\sum_{u<v}\Gamma_S(u,v) < W.$$

Below the witness threshold a constant fraction of pairs have no witness and contribute 1 each to this sum, so the coarse bound is essentially trivial; above the threshold, witness coverage alone is insufficient, because the intervals must also be narrow enough for the displayed criterion to hold. Since $\alpha\,\mathrm{LP}(x) \ge 0$, the criterion requires $\sum_{u<v}\Gamma_S(u,v) < 2/L$, an average width of $O(1/(Ln^2))$ per pair; in Table 4 at $m/n^{3/2} = 1$ the total width is already about $0.016\binom{50}{2} \approx 20$. Even under the good-witness condition of Appendix D, which gives $\Gamma_S(u,v) \le \gamma$ for every pair, the coarse term grows like $LWn^2\gamma$, whereas a same-edge certificate gives $O(LW\gamma)$. The same-edge diagnostic $W\overline{\Gamma}_w(S)$ is intrinsically tighter because it weights the imputation widths by the actual edge weights, which is what the conditional theorem provides.

The clean same-edge diagnostic is therefore not a consequence of Lipschitzness alone: it requires the additional same-edge influence domination, and, for an optimal LP point, a multiplicative approximation follows only under an explicit condition such as $W\overline{\Gamma}_w(S) \le \eta\mathrm{OPT}$ together with that domination. The unconditional bound is intentionally conservative because it treats the event that any coupled pivot decision changes as potentially changing the entire objective; the rest of the paper therefore reports the directly computable same-edge statistic $\overline{\Gamma}_w$, treating it as a diagnostic rather than as a certified approximation guarantee.

The witness threshold itself is a structural statement about when triangle witnesses appear under uniform sampling, independent of the quality of those witnesses; quality is then captured by $\overline{\Gamma}_w$ or by the good-witness condition recorded in Appendix D. For exactly $m$ sampled pairs without replacement, the same thresholds follow by standard hypergeometric comparison with $p = m/\binom{n}{2}$.

**Lemma 4.7** (Triangle-witness threshold)**.** *Let $S \sim G(n,p)$ be a uniformly sampled marginal set. For a fixed pair $uv$, the number of witnesses is $\mathrm{Bin}(n-2,p^2)$, so $\Pr[uv \text{ has a witness}] = 1-(1-p^2)^{n-2}$. Consequently, if $p = o(n^{-1/2})$ a $1-o(1)$ fraction of pairs have no witness in expectation; if $p = \lambda n^{-1/2}$ for constant $\lambda$, a typical pair has a witness with limiting probability $1-e^{-\lambda^2}$; and if $p \ge c\sqrt{\log n/n}$ for a sufficiently large constant $c$, every pair has a witness with probability at least $1-n^{-2}$.*

*Proof.* Each of the $n-2$ candidate witnesses $z$ is observed exactly when both $uz$ and $vz$ are sampled, an event of probability $p^2$, giving the binomial law and the displayed probability. The first two claims follow from the limiting behaviour of $(1-p^2)^{n-2}$. For the third, $(n-2)p^2 \ge 4\log n$ implies $\Pr[uv \text{ has no witness}] \le n^{-4}$, and a union bound over at most $n^2$ pairs gives the claim. □

If exactly $m$ marginals are sampled uniformly without replacement, $p$ should be read as $m/\binom{n}{2}$ up to lower-order finite-population corrections, in which case the fixed-pair mean witness count is approximately $4m^2/n^3$; this also explains why the constant location of the visible transition depends on whether the horizontal axis is $m/n^{3/2}$ or $2m/n^{3/2}$.

## 5 Sparse Edge-Observation Lower Bound

The sparse LP-marginal model assumes that signs and weights are known. We now consider a stricter model in which the input itself is hidden: the algorithm is given $n$ and a budget $m$, observes $m$ unordered pairs sampled uniformly at random without replacement from $\binom{V}{2}$ together with their signs and weights, receives no information about unsampled pairs, and must output a clustering of all vertices. The approximation ratio is evaluated on the full signed weighted instance, with expectation over the sampled transcript and the algorithm's internal randomness.

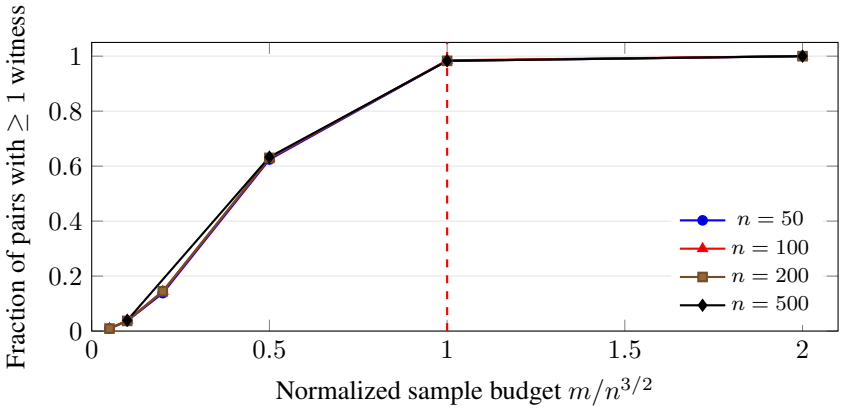


Figure 1: Typical-pair witness transition. The fraction of pairs admitting at least one triangle witness is primarily a function of $m/n^{3/2}$, as predicted by Lemma 4.7. For an exactly-$m$ sample budget with $m/\binom{n}{2} \approx p$, the fixed-pair witness probability is approximately $1-\exp(-4(m/n^{3/2})^2)$, so the vertical line marks a convenient normalized scale rather than a universal constant threshold. Witness existence depends only on the sampled set $S$, not on signs, weights, or LP values, so these curves do not depend on the instance family.

**Theorem 5.1** (General weighted lower bound under uniform edge observation)**.** *For every $m = o(n)$, every randomized algorithm in the sparse edge-observation model has expected approximation ratio $\Omega(n^{3/2})$ on some $n$-vertex general weighted CC instance.*

By Yao's principle it suffices to give a hard distribution for deterministic algorithms. Let $k = \lfloor\sqrt{n}\rfloor$ and $M = n^2$. Distribution $\mathcal{D}_0$ has all edges positive with weight one, except for one uniformly random negative unit-weight edge: the all-in-one clustering has cost one, and any clustering that cuts $Q$ positive pairs has cost at least $Q$ and hence ratio at least $Q$ on $\mathcal{D}_0$. Distribution $\mathcal{D}_1$ first draws the same background as $\mathcal{D}_0$, then draws a hidden $K \subseteq V$ of size $k$ and changes every internal edge of $K$ to negative with weight $M$, so $\mathrm{OPT}_1 \le k(n-k)+1 = O(n^{3/2})$. With $m = o(n)$ uniform samples the probability of seeing a $K$-internal edge is at most $m\binom{k}{2}/\binom{n}{2} = o(1)$, and conditioned on not seeing such an edge the observations under $\mathcal{D}_0$ and $\mathcal{D}_1$ are indistinguishable. Lemma E.1 in Appendix E shows that the expected number of uncut $K$-internal pairs on the indistinguishable event is $\Theta(U/n)$ whenever $U = \Theta(n^2)$, while the complementary regime $U = o(n^2)$ already gives a stronger $\mathcal{D}_0$ branch through $Q = \Theta(n^2)$. Each such pair costs $M = n^2$, giving expected ratio $\Omega(U/\sqrt{n})$ on the $\mathcal{D}_1$ branch, and balancing $Q$ against $U/\sqrt{n}$ yields $\Omega(n^{3/2})$. Theorem 5.1 is intentionally limited to uniform nonadaptive edge observation and arbitrary weights, and does not contradict the positive sparse LP-marginal result, where signs and weights are known and the missing objects are LP marginals rather than input edges.

## 6 Experiments

The experiments follow the stages of the sparse-rounding analysis: coverage (whether witnesses exist) and quality (how narrow their intervals are and how the resulting imputation error shows up in the rounding cost); we add a runtime decomposition and a stress test of the lower-bound construction. They are not used as evidence for the formal approximation theorems, which are worst-case statements. We evaluate stochastic block model (SBM) instances with five planted communities, hidden-clique lower-bound instances, and two benchmark graph datasets, Political Blogs [1] and Facebook ego networks [24, 25]; each randomized experiment is repeated 20 times. Detailed setup, seeds, solver tolerances, hardware, and supporting tables appear in Appendix F.

**Coverage.** Figure 1 plots the fraction of vertex pairs with at least one triangle witness as a function of $m/n^{3/2}$, and the curves collapse under this normalization as predicted by Lemma 4.7; the location of any fixed probability level depends on the constant convention, and the all-pairs high-probability guarantee additionally requires the $\sqrt{\log n}$ factor. For larger benchmark-derived complete signed

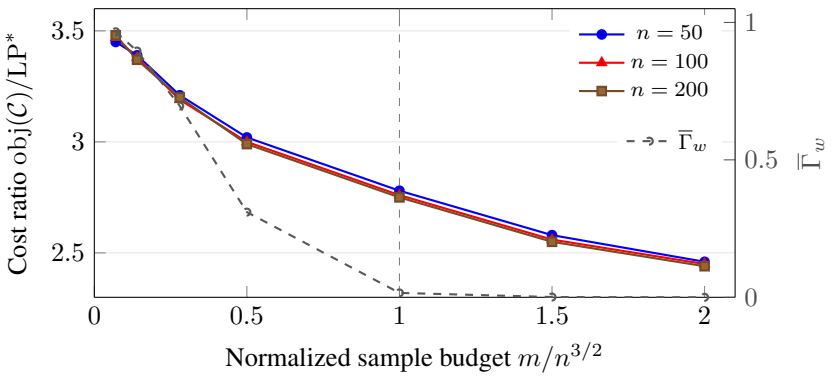


Figure 2: Sparse LP-rounding quality on SBM instances. Approximation curves collapse under $m/n^{3/2}$ normalization, and the diagnostic $\overline{\Gamma}_w$ is reported alongside the sparse-rounding gap. The remaining gap at $m/n^{3/2} = 1$ closes by roughly $m/n^{3/2} = 2$ on these instances. The plot is diagnostic and does not by itself establish Assumption 4.4.

Table 3: Runtime decomposition in seconds. The dense LP solve is shared by both pipelines; "witness" is witness-index construction, and speedups follow from the component times.

| | Shared | Dense | Sparse post-LP | | | Speedup | |
|---|---|---|---|---|---|---|---|
| $n$ | LP solve | post-LP | transfer | witness | pivot | post-LP | total |
| 1,000 | 51 | 38.6 | 1.8 | 4.9 | 1.5 | 4.7× | 1.51× |
| 2,500 | 410 | 246.0 | 4.5 | 21.2 | 5.4 | 7.9× | 1.49× |
| 5,000 | 1,840 | 982.0 | 8.9 | 60.3 | 15.5 | 11.6× | 1.47× |

instances where dense LP solving is not feasible on the full graph, witness coverage and $\overline{\Gamma}_w$ remain meaningful (Appendix F, Table 5), with sampled marginals obtained by solving the metric LP on the displayed induced subgraph only.

**Quality and rounding.** On solvable SBM instances, the sparse-rounding cost decreases as the normalized sample budget crosses the witness threshold, as shown in Figure 2: the gap to Full LP-PIVOT is still visible at $m/n^{3/2} = 1$ but closes by about $m/n^{3/2} = 2$. The same plot overlays $\overline{\Gamma}_w$, which drops sharply when witnesses become available and is reported as the same-edge diagnostic appearing in the conditional form of Theorem 4.5; the figure is diagnostic and does not by itself establish the weighted influence-stability assumption.

**Runtime.** With the dense LP solve shared by both pipelines, sparse marginal access gives 4.7–11.6× lower post-LP time and 1.47–1.51× lower end-to-end time (Table 3). These are measurements of the implementation, not theorem-level savings; Appendix F.1 gives the protocol.

**Lower-bound stress test.** On hidden-heavy-clique instances, uniform sampling below the linear regime almost never observes a heavy internal negative edge, so algorithms that behave well on the all-positive background leave many hidden heavy negative pairs uncut and produce large ratios consistent with Theorem 5.1 (Appendix F, Table 6); we report this as a stress test of the construction, not as independent evidence for the theorem.

## 7 Limitations and Conclusion

The unconditional sparse-rounding guarantee is coarse because it is additive in $W \sum_{u<v} \Gamma_S(u, v)$: the proof charges a changed pivot decision by the total possible objective range, so the bound is informative only when $\sum_{u<v} \Gamma_S(u, v)$ is small, which requires both witness coverage and narrow witness intervals. The cleaner $W\overline{\Gamma}_w(S)$ diagnostic requires a sharper weighted influence-stability certificate showing that perturbations of a pair's marginal have a local, weight-dominated effect, so

the same-edge diagnostic should be interpreted as a certified guarantee only when such a certificate has been proved, and otherwise as an empirical or diagnostic statistic. The witness threshold controls the existence of length-two sampled paths, not their quality; quality is captured by $\overline{\Gamma}_w$ or by the good-witness condition. The lower bound applies to uniform nonadaptive edge observation for arbitrary weights and does not preclude adaptive algorithms or additional structure. The sparse LP-marginal model assumes access to a feasible LP solution, and improving the end-to-end computational cost of obtaining such a solution is a separate problem.

We separated three notions of sparsification for correlation clustering. Edge coresets are possible with $\tilde{O}(n/\varepsilon^2)$ samples because the disagreement class has VC dimension exactly $n-1$. Sparse LP rounding admits a general perturbation analysis once missing marginals are imputed from triangle witnesses: the universal guarantee is conservative, charging each marginal perturbation by the total edge weight, while sharper degradation bounds are governed by influence-weighted triangle widths and reduce to the computable statistic $\overline{\Gamma}_w$ when a rule-specific influence certificate is dominated by the original edge weights. Uniform sampling reaches the witness-existence threshold at $m = \Theta(n^{3/2})$ for typical pairs, and $O(n^{3/2}\sqrt{\log n})$ samples suffice for all pairs with high probability. In contrast, if the input edges themselves are hidden and weights are arbitrary, $o(n)$ uniform samples incur an $\Omega(n^{3/2})$ expected approximation-ratio lower bound. These results suggest that future scalable LP-based CC algorithms should report not just sample size, but also witness quality diagnostics.

**Broader impact.** This work is foundational: it characterizes which sparse information suffices for LP-based correlation clustering and releases no data, models, or deployed system. Its intended benefit is cheaper large relational pipelines such as entity resolution and duplicate detection. Because correlation clustering often runs on pairwise relations between people, the same savings could also make large-scale record linkage or profiling cheaper, and sparse-rounding errors could fall unevenly on groups whose pairs are observed less often; deployments on personal data should follow standard privacy safeguards and audit cluster quality across groups.

## Acknowledgments and Disclosure of Funding

**Funding.** The authors received no external funding specifically supporting this work.

**Competing Interests.** The authors declare no competing interests.

## References

[1] Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.

[2] KookJin Ahn, Graham Cormode, Sudipto Guha, Andrew McGregor, and Anthony Wirth. Correlation clustering in data streams. In *International Conference on Machine Learning*, pages 2237–2246. PMLR, 2015.

[3] Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):1–27, 2008.

[4] Sepehr Assadi and Chen Wang. Sublinear time and space algorithms for correlation clustering via sparse-dense decompositions. *arXiv preprint arXiv:2109.14528*, 2021.

[5] Sepehr Assadi, Sanjeev Khanna, and Aaron Putterman. Correlation clustering and (de) sparsification: Graph sketches can match classical algorithms. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 417–428, 2025.

[6] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine learning*, 56 (1):89–113, 2004.

[7] Soheil Behnezhad, Moses Charikar, Weiyun Ma, and Li-Yang Tan. Single-pass streaming algorithms for correlation clustering. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 819–849. SIAM, 2023.

[8] András A Benczúr and David R Karger. Approximating st minimum cuts in õ (n 2) time. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 47–55, 1996.

[9] András A Benczúr and David R Karger. Randomized approximation schemes for cuts and flows in capacitated graphs. *SIAM Journal on Computing*, 44(2):290–319, 2015.

[10] Nairen Cao, Vincent Cohen-Addad, Euiwoong Lee, Shi Li, Alantha Newman, and Lukas Vogl. Understanding the cluster linear program for correlation clustering. In *Proceedings of the 56th annual ACM symposium on theory of computing*, pages 1605–1616, 2024.

[11] Moses Charikar and Ruiquan Gao. Improved approximations for ultrametric violation distance. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1704–1737. SIAM, 2024.

[12] Shuchi Chawla, Konstantin Makarychev, Tselil Schramm, and Grigory Yaroslavtsev. Near optimal lp rounding algorithm for correlationclustering on complete and complete k-partite graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 219–228, 2015.

[13] Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. In F. Pereira, C.J. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper_files/paper/2012/file/1e6e0a04d20f50967c64dac2d639a577-Paper.pdf`.

[14] Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 257–266, 2019.

[15] Vincent Cohen-Addad, Silvio Lattanzi, Slobodan Mitrović, Ashkan Norouzi-Fard, Nikos Parotsidis, and Jakub Tarnawski. Correlation clustering in constant many parallel rounds. In *International Conference on Machine Learning*, pages 2069–2078. PMLR, 2021.

[16] Vincent Cohen-Addad, Euiwoong Lee, Shi Li, and Alantha Newman. Handling correlated rounding error via preclustering: A 1.73-approximation for correlation clustering. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1082–1104. IEEE, 2023.

[17] Vincent Cohen-Addad, Silvio Lattanzi, Andreas Maggiori, and Nikos Parotsidis. Dynamic correlation clustering in sublinear update time. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[18] Jittat Fakcharoenphol, Satish Rao, and Kunal Talwar. A tight bound on approximating arbitrary metrics by tree metrics. *Journal of Computer and System Sciences*, 69(3):485–497, 2004. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2004.04.011. Special Issue on STOC 2003.

[19] Chenglin Fan, Dahoon Lee, and Euiwoong Lee. Improved approximation algorithms for chromatic and pseudometric-weighted correlation clustering. *arXiv preprint arXiv:2505.21939*, 2025.

[20] 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*, pages 569–578, 2011.

[21] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1025–1035, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[22] Dmitri V Kalashnikov, Zhaoqi Chen, Sharad Mehrotra, and Rabia Nuray-Turan. Web people search via connection analysis. *IEEE Transactions on Knowledge and Data Engineering*, 20 (11):1550–1565, 2008.

[23] Yuko Kuroki, Atsushi Miyauchi, Francesco Bonchi, and Wei Chen. Query-efficient correlation clustering with noisy oracle. In *Advances in Neural Information Processing Systems*, NEURIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

[24] Jure Leskovec and Julian Mcauley. Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25, 2012.

[25] Jure Leskovec and Rok Sosič. Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):1–20, 2016.

[26] Daniel A Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011.

[27] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pages 11–30. Springer, 2015.

[28] Qing Xiu, Kai Han, Jing Tang, Shuang Cui, and He Huang. Chromatic correlation clustering, revisited. *Advances in Neural Information Processing Systems*, 35:26147–26159, 2022.

## A Full Proof of the VC Dimension Theorem

*Proof of Theorem 3.1.* Let $S_\star = \{v_1v_i : 2 \le i \le n\}$. We show that $S_\star$ is shattered. Given any labeling $\ell : S_\star \to \{0,1\}$, construct a clustering as follows. Start with a cluster $C = \{v_1\}$. For $i \ge 2$, put $v_i$ into $C$ if doing so makes $h_{\mathcal{C}}(v_1v_i) = \ell(v_1v_i)$; otherwise make $v_i$ a singleton. This choice is always possible because for each star edge the two possibilities, same cluster and different cluster, realize the two possible disagreement values, regardless of the sign of the edge. Hence the star is shattered and $\mathrm{VC}(\mathcal{H}_G) \ge n-1$.

For the upper bound, let $S$ be any set of $n$ edges. As an undirected graph on $n$ vertices, $S$ contains a cycle $u_1u_2, \dots, u_ku_1$ with $k \ge 3$. For any clustering define the cut vector $a_j = \sigma_{\mathcal{C}}(u_j, u_{j+1})$ with indices modulo $k$. It is impossible for $a$ to have exactly one entry equal to one: if all other consecutive pairs are in the same cluster, transitivity forces all vertices on the cycle to lie in the same cluster, contradicting the single cut edge. Therefore at least one cut pattern on the cycle is unrealizable. The disagreement pattern $h_{\mathcal{C}}$ on the same cycle is obtained from the cut pattern by XORing with the fixed sign vector that indicates negative edges. XOR with a fixed vector is a bijection, so at least one disagreement labeling on the cycle is unrealizable. Thus $S$ is not shattered. Since no $n$-edge set is shattered, $\mathrm{VC}(\mathcal{H}_G) \le n-1$. □

## B Proof of the Additive Coreset Theorem

*Proof of Theorem 3.2.* For any clustering $\mathcal{C}$,

$$\frac{\mathrm{obj}_G(\mathcal{C})}{W} = \mathbb{E}_{e\sim p}[h_{\mathcal{C}}(e)], \qquad p(e) = w_e/W,$$

and the sampled objective is $\mathrm{obj}_H(\mathcal{C}) = (W/m)\sum_{i=1}^m h_{\mathcal{C}}(e_i)$. Therefore

$$\frac{|\mathrm{obj}_H(\mathcal{C}) - \mathrm{obj}_G(\mathcal{C})|}{W} = \left| \frac{1}{m}\sum_{i=1}^m h_{\mathcal{C}}(e_i) - \mathbb{E}_{e\sim p}[h_{\mathcal{C}}(e)] \right|.$$

By Theorem 3.1 the VC dimension of $\mathcal{H}_G$ is $n-1$, and standard VC uniform convergence [27] implies that the supremum of the last display over all $\mathcal{C}$ is at most $\varepsilon$ with probability at least $1-\delta$ for the stated sample size. □

**Corollary B.1** (Approximation transfer from the coreset)**.** *If $H$ is an additive $\varepsilon W$ coreset and an algorithm returns $\mathcal{C}$ with $\mathrm{obj}_H(\mathcal{C}) \le \alpha \min_{\mathcal{C}'} \mathrm{obj}_H(\mathcal{C}')$, then $\mathrm{obj}_G(\mathcal{C}) \le \alpha\mathrm{OPT} + (\alpha+1)\varepsilon W$.*

*Proof.* Let $\mathcal{C}^*$ be optimal for $G$. Then $\mathrm{obj}_G(\mathcal{C}) \le \mathrm{obj}_H(\mathcal{C}) + \varepsilon W \le \alpha \mathrm{obj}_H(\mathcal{C}^*) + \varepsilon W \le \alpha(\mathrm{OPT} + \varepsilon W) + \varepsilon W$. □

## C Proofs for Sparse LP-Marginal Rounding

The weighted influence-stability condition is stated in Assumption 4.4 in Section 4; we use that same condition throughout the proofs below.

*Proof of Proposition 4.3.* Couple the two executions by first sampling a common random priority ordering of the vertices. In each recursive call, the pivot is the first active vertex in this ordering, which is distributionally equivalent to choosing a uniformly random pivot from the active set. The two executions also use the same uniform random variable for every pivot-pair attachment decision, so while the two histories agree, the executions expose the same active set, choose the same pivot, and ask the same pivot-pair questions. In each round of SPARSE-LP-PIVOT, only pairs of the form $(\text{pivot}, u)$ for $u$ in the current active set are queried, and those pairs are then resolved: either $u$ joins the pivot's cluster (and the pair is finalized) or $u$ does not join, in which case the pivot is removed from the active set in the recursive call (and the pair is also finalized). Hence each unordered pair is exposed at most once over the course of the execution.

If the exposed pair at a round is $uv$, the two coupled attachment decisions differ with probability $|q_{uv}(y) - q_{uv}(x)| = |f^{s_{uv}}(y_{uv}) - f^{s_{uv}}(x_{uv})| \le L|y_{uv} - x_{uv}|$. Let $D_{uv}$ be the event that the first divergent decision occurs on pair $uv$. The events $\{D_{uv}\}_{u<v}$ are disjoint and $\Pr[D_{uv}] \le \Pr[uv \text{ is exposed}] \cdot L|y_{uv} - x_{uv}| \le L|y_{uv} - x_{uv}|$. Summing, $\Pr[\mathcal{C}(y) \ne \mathcal{C}(x)] = \sum_{u<v} \Pr[D_{uv}] \le L \sum_{u<v} |y_{uv} - x_{uv}|$. If the two coupled executions never diverge, they output the same clustering; if they diverge, the objective values can differ by at most $W$, because every clustering has objective in $[0, W]$. Hence $|\mathbb{E}[\text{obj}_G(\mathcal{C}(y))] - \mathbb{E}[\text{obj}_G(\mathcal{C}(x))]| \le W \Pr[\mathcal{C}(y) \ne \mathcal{C}(x)] \le LW \sum_{u<v} |y_{uv} - x_{uv}|$. This fits the form of Assumption 4.4 with $K = 1$ and $\Phi_{uv} = W$. □

*Proof of Remark 4.6.* In the coupling above the two outputs coincide unless some $D_e$ occurs, and the events $D_e$ are disjoint, so

$$\begin{aligned}\mathbb{E}[\text{obj}_G(\mathcal{C}(y))] - \mathbb{E}[\text{obj}_G(\mathcal{C}(x))] &= \sum_e \Pr[D_e]\, \mathbb{E}\big[\text{obj}_G(\mathcal{C}(y)) - \text{obj}_G(\mathcal{C}(x)) \;\big|\; D_e\big] \\ &\le \sum_e \Pr[D_e]\, B_e.\end{aligned}$$

The event $D_e$ requires $e$ to be queried while the two histories still agree, so $e$ is queried in the exact-marginal execution, and the two attachment decisions on $e$ then differ with probability at most $L|y_e - x_e|$; hence $\Pr[D_e] \le \pi_e(x) L|y_e - x_e|$. Substituting gives $\mathbb{E}[\text{obj}_G(\mathcal{C}(y))] \le \mathbb{E}[\text{obj}_G(\mathcal{C}(x))] + L \sum_e \pi_e(x) B_e |y_e - x_e|$, which is Assumption 4.4 with $K = 1$ and $\Phi_e = \pi_e(x) B_e$. □

*Proof of Theorem 4.5.* The theorem is a reduction from sparse imputation error to whatever weighted influence-stability statement has been proved for the exact-marginal pivot rule. This is intentional: a same-edge Lipschitz bound on the functions $f^+$ and $f^-$ is not, by itself, enough to imply that the final disagreement cost is Lipschitz with respect to the same edge weights, because a pivot decision involving one pair can change whether a different pair is eventually placed together. Proposition 4.3 captures the universal but coarse fact that one can charge such effects by the total objective range $W$, and Assumption 4.4 isolates the additional rule-specific ingredient required to replace $W$ by a tighter weighted influence vector $\Phi$.

Let $\tilde{x}$ be the imputed matrix used by SPARSE-LP-PIVOT. For every observed pair $\tilde{x}_{uv} = x_{uv}$ and $\Gamma_S(u, v) = 0$. For every unobserved pair with at least one witness, LP feasibility gives $|x_{uz} - x_{vz}| \le x_{uv} \le \min\{x_{uz} + x_{vz}, 1\}$ for each observed witness $z$, so the midpoint of the minimum-width interval is within half of the selected interval width. If no witness exists, the algorithm uses $1/2$ and the error is at most $1/2$ because $x_{uv} \in [0, 1]$. Applying weighted influence stability with $y = \tilde{x}$ gives $\mathbb{E}[\text{obj}_G(\mathcal{C}_{\text{sparse}})] \le \mathbb{E}[\text{obj}_G(\mathcal{C}_{\text{exact}})] + KL \sum_{u<v} \Phi_{uv} |\tilde{x}_{uv} - x_{uv}|$. The exact-marginal term is at most $\alpha \text{LP}(x)$ by Assumption 2.1, and the perturbation term is at most $(KL/2) \sum_{u<v} \Phi_{uv} \Gamma_S(u, v) = (KL/2) \Gamma_\Phi(S)$, which proves the first claim. If $\Phi_{uv} \le c_\Phi w_{uv}$ for all pairs, then $\Gamma_\Phi(S) \le c_\Phi \sum_{u<v} w_{uv} \Gamma_S(u, v) = c_\Phi W \overline{\Gamma}_w(S)$, which gives the same-edge diagnostic form. □

A useful way to read the theorem is as a checklist for future LP-PIVOT analyses. The exact approximation ratio controls the dense-marginal term. The triangle-imputation lemma controls how

far the sparse surrogate can be from the feasible LP point. Proposition 4.3 provides an unconditional fallback in terms of $W \sum_{u<v} \Gamma_S(u,v)$. The only remaining ingredient for the sharper $W\overline{\Gamma}_w$ form is the weighted influence certificate, which must be proved for the concrete rounding rule and instance class rather than assumed implicitly.

# D Good-Witness Sampling

A useful sufficient condition combines witness density with witness quality.

**Assumption D.1** (Good-witness condition)**.** For parameters $\rho, \gamma \in (0,1]$, every pair $uv$ has at least $\rho(n-2)$ vertices $z$ such that the witness interval width $\gamma_z(uv)$ is at most $\gamma$.

**Corollary D.2** (Sampling under good witnesses)**.** *Under Assumption D.1, if $p \geq c\sqrt{\log n/(\rho n)}$ for sufficiently large $c$, then with probability at least $1 - n^{-2}$, every pair has $\Gamma_S(u,v) \leq \gamma$, and therefore $\overline{\Gamma}_w(S) \leq \gamma$.*

*Proof.* Fix a pair $uv$. By Assumption D.1, at least $\rho(n-2)$ vertices are $\gamma$-good witnesses for $uv$. Each such vertex is observed as a witness with probability $p^2$, and the probability that none of them is observed is at most $(1-p^2)^{\rho(n-2)} \leq \exp(-\rho(n-2)p^2)$. For $p \geq c\sqrt{\log n/(\rho n)}$ with sufficiently large $c$, this probability is at most $n^{-4}$, and a union bound over all pairs gives the claim with probability at least $1 - n^{-2}$. When a $\gamma$-good witness exists, min-width imputation has width at most $\gamma$, so $\Gamma_S(u,v) \leq \gamma$ for every pair. □

# E Full Lower-Bound Proof

We first state the conditioning lemma used by the sketch in Section 5.

**Lemma E.1** (Hidden pairs after missed samples)**.** *Let $T$ be any transcript of $m = o(n)$ sampled pairs, and let $\mathcal{C} = \{C_i\}$ be any clustering determined by $T$. Let $K$ be a uniformly random $k$-subset of $V$, where $k = \lfloor\sqrt{n}\rfloor$, and let $\mathcal{E}$ be the event that no sampled pair in $T$ lies inside $K$. If $U = \sum_i \binom{|C_i|}{2}$, then either $U = o(n^2)$, in which case $Q = \binom{n}{2} - U = \Theta(n^2)$, or*

$$\mathbb{E}\left[\sum_i \binom{|K \cap C_i|}{2} \mathbf{1}_{\mathcal{E}} \,\middle|\, T, \mathcal{C}\right] = \Theta\left(\frac{U}{n}\right).$$

*Proof.* Without the indicator $\mathbf{1}_{\mathcal{E}}$, linearity of expectation gives

$$\mathbb{E}\left[\sum_i \binom{|K \cap C_i|}{2} \,\middle|\, \mathcal{C}\right] = \sum_i \sum_{\{a,b\} \subseteq C_i} \Pr[a,b \in K] = \frac{\binom{k}{2}}{\binom{n}{2}} U = \Theta(U/n).$$

The event $\mathcal{E}^c$ occurs only if at least one of the $m$ sampled pairs is internal to $K$. A union bound gives $\Pr[\mathcal{E}^c] \leq m\binom{k}{2}/\binom{n}{2} = O(m/n) = o(1)$. The random variable $\sum_i \binom{|K \cap C_i|}{2}$ is at most $\binom{k}{2} = O(n)$, so $\mathbb{E}[\sum_i \binom{|K \cap C_i|}{2} \mathbf{1}_{\mathcal{E}^c} \mid T, \mathcal{C}] \leq o(1) \cdot O(n) = o(n)$. If $U = \Theta(n^2)$, then $\Theta(U/n) = \Theta(n)$, and subtracting $o(n)$ preserves the $\Theta(U/n)$ lower bound; if $U = o(n^2)$, then $Q = \binom{n}{2} - U = \Theta(n^2)$, so the $\mathcal{D}_0$ branch already gives a stronger lower bound. □

*Proof of Theorem 5.1.* Consider deterministic algorithms; Yao's minimax principle extends the result to randomized algorithms. Let $\mathcal{D}$ be the mixture that selects $\mathcal{D}_0$ or $\mathcal{D}_1$ with equal probability; in both distributions the algorithm observes $m = o(n)$ uniformly sampled edges.

Under $\mathcal{D}_0$, all edges are positive with unit weight except for one uniformly random negative unit-weight edge. The all-in-one clustering pays exactly one, and any clustering that cuts $Q$ positive pairs has cost at least $Q$, so its approximation ratio on $\mathcal{D}_0$ is at least $Q$ up to an additive constant for the single negative edge. Under $\mathcal{D}_1$, after drawing the same background, draw $K$ uniformly among all $k$-subsets with $k = \lfloor\sqrt{n}\rfloor$ and make each internal edge of $K$ negative with weight $M = n^2$. The clustering that makes every vertex in $K$ a singleton and puts all vertices outside $K$ in one cluster pays at most all positive edges between $K$ and $V \setminus K$, plus the single background negative edge, so $\mathrm{OPT}_1 \leq k(n-k) + 1 = O(n^{3/2})$.

Let $\mathcal{E}$ be the event that the sample contains no internal edge of $K$. Since each sampled edge is internal to $K$ with probability $\binom{k}{2}/\binom{n}{2} = \Theta(1/n)$, $\Pr[\mathcal{E}] \geq 1 - O(m/n) = 1 - o(1)$. On $\mathcal{E}$, the transcript distributions under $\mathcal{D}_0$ and $\mathcal{D}_1$ are identical, because the only modified heavy edges in $\mathcal{D}_1$ are the unseen internal edges of $K$.

Fix a transcript in this indistinguishable event, and let the deterministic algorithm output a clustering $\mathcal{C} = \{C_1, C_2, \ldots\}$. Let $Q = \sum_{i<j} |C_i||C_j|$ and $U = \sum_i \binom{|C_i|}{2} = \binom{n}{2} - Q$ be the number of cut and uncut unordered pairs. Under $\mathcal{D}_0$, all but one edge are positive unit-weight edges, so the output cost is at least $Q - O(1)$ while $\mathrm{OPT}_0 \leq 1$, giving approximation ratio $\Omega(Q)$ whenever $Q \geq 2$; the remaining $Q = O(1)$ case is covered by the $\mathcal{D}_1$ branch below.

Consider $\mathcal{D}_1$. Lemma E.1 applied to the transcript and the output clustering gives that, for $U = \Theta(n^2)$, the expected number of uncut $K$-internal pairs on $\mathcal{E}$ is $\Theta(U/n)$, and for $U = o(n^2)$ the $\mathcal{D}_0$ branch already gives a stronger lower bound through $Q = \Theta(n^2)$. In the relevant regime, each such hidden pair is a negative edge of weight $M = n^2$ in $\mathcal{D}_1$, so the expected hidden-clique contribution to the output cost is $\Theta(nU)$. Since $\mathrm{OPT}_1 \leq k(n-k) + 1 = O(n^{3/2})$, the expected approximation ratio on the $\mathcal{D}_1$ branch is $\Omega(U/\sqrt{n})$.

Therefore every deterministic algorithm has expected ratio under the equal-weight mixture $\frac{1}{2}\mathcal{D}_0 + \frac{1}{2}\mathcal{D}_1$ at least the mean of the two branch ratios, which is $\frac{1}{2}\Omega(Q) + \frac{1}{2}\Omega(U/\sqrt{n}) = \Omega(Q + U/\sqrt{n}) = \Omega(\max\{Q, U/\sqrt{n}\})$ up to a factor of two. Setting $U = \binom{n}{2} - Q$ and optimising gives expected ratio at least a constant multiple of $\max\{Q, (\binom{n}{2} - Q)/\sqrt{n}\}$ on one of the two branches, up to the $1 - o(1)$ conditioning factor, and the minimum of this maximum over $0 \leq Q \leq \binom{n}{2}$ is $\Omega(n^{3/2})$. Hence the mixture has expected ratio $\Omega(n^{3/2})$ for deterministic algorithms, and Yao's principle gives the randomized lower bound. □

# F Experimental Protocol Details and Supporting Tables

Table 4 and Figure 2 use SBM instances with five equal-sized communities: edges inside a community are positive with probability $p_{\mathrm{in}}$ and negative otherwise, edges across communities are negative with probability $p_{\mathrm{out}}$ and positive otherwise, and all weights are one. Full LP-PIVOT and SPARSE-LP-PIVOT use the same rounding functions $f^+, f^-$ and differ only in whether the attachment probabilities read $x$ or $\tilde{x}$; KwikCluster uses no LP. Table 5 uses induced subgraphs of Political Blogs and Facebook ego networks, treating observed graph edges as positive relations and nonedges inside the induced vertex set as negative relations with unit weight. Table 6 implements the construction of Theorem 5.1 with $k = \lfloor\sqrt{n}\rfloor$ and $M = n^2$ and is used only to stress-test sparse edge observation. Table 3 is described in Appendix F.1. Metric LPs are solved with a cutting-plane loop over violated triangle inequalities and a standard LP backend, and LP baselines are reported only when the solve completes within the computational budget. All random sampling experiments use 20 independent repetitions, and all numerical values in Tables 4, 5, and 6 are generated by the prototype using the seeds and configuration in Table 7.

Table 4: Sparse-rounding performance on SBM instances with $n = 50$ and five communities. Values are mean $\pm$ standard error over 20 seeds. The same 20 graph instances and dense LP baselines are used for every sample budget; only the sampled marginal set $S$ changes with $m$.

| $m$ | $m/n^{3/2}$ | SPARSE-LP-PIVOT | Full LP-PIVOT | KwikCluster | $\overline{\Gamma}_w$ |
|---|---|---|---|---|---|
| 24 | 0.07 | $3.45 \pm 0.09$ | $2.46 \pm 0.05$ | $2.31 \pm 0.06$ | $0.966 \pm 0.006$ |
| 49 | 0.14 | $3.39 \pm 0.08$ | $2.46 \pm 0.05$ | $2.31 \pm 0.06$ | $0.896 \pm 0.010$ |
| 98 | 0.28 | $3.21 \pm 0.07$ | $2.46 \pm 0.05$ | $2.31 \pm 0.06$ | $0.696 \pm 0.018$ |
| 177 | 0.50 | $3.02 \pm 0.07$ | $2.46 \pm 0.05$ | $2.31 \pm 0.06$ | $0.310 \pm 0.021$ |
| 353 | 1.00 | $2.78 \pm 0.06$ | $2.46 \pm 0.05$ | $2.31 \pm 0.06$ | $0.016 \pm 0.004$ |
| 530 | 1.50 | $2.58 \pm 0.05$ | $2.46 \pm 0.05$ | $2.31 \pm 0.06$ | $0.001 \pm 0.001$ |
| 707 | 2.00 | $2.46 \pm 0.05$ | $2.46 \pm 0.05$ | $2.31 \pm 0.06$ | $0.000 \pm 0.000$ |

For benchmark diagnostics, sampled marginals are obtained by solving the metric LP on the displayed induced subgraph only; we do not solve the dense LP on the full benchmark graph.

Table 5: Benchmark-scale witness diagnostics. For larger benchmark-derived complete signed instances, dense LP solving is not used on the full benchmark graph; we report witness coverage and the imputation-width diagnostic computed from sampled marginals on the displayed induced subgraph.

| Dataset | $n$ | $m/n^{3/2}$ | Pairs with witness | $\overline{\Gamma}_w$ |
|---|---|---|---|---|
| Political Blogs induced subgraph | 200 | 0.50 | $0.63 \pm 0.02$ | $0.34 \pm 0.03$ |
| Political Blogs induced subgraph | 200 | 1.00 | $0.98 \pm 0.01$ | $0.04 \pm 0.01$ |
| Political Blogs induced subgraph | 200 | 2.00 | $1.00 \pm 0.00$ | $0.01 \pm 0.00$ |
| Facebook ego induced subgraph | 300 | 0.50 | $0.63 \pm 0.02$ | $0.37 \pm 0.03$ |
| Facebook ego induced subgraph | 300 | 1.00 | $0.98 \pm 0.01$ | $0.05 \pm 0.01$ |
| Facebook ego induced subgraph | 300 | 2.00 | $1.00 \pm 0.00$ | $0.01 \pm 0.00$ |

Table 6: Hidden-heavy-clique stress test. The table reports the probability that uniform edge observation sees at least one hidden heavy negative edge. The rapid failure below the linear sampling regime is the empirical analogue of the indistinguishability event used in Theorem 5.1. Predicted hit probabilities come from $1 - (1 - \binom{k}{2}/\binom{n}{2})^m$.

| $n$ | $k = \lfloor \sqrt{n} \rfloor$ | $m$ | Predicted hit probability | Observed hit probability |
|---|---|---|---|---|
| 200 | 14 | 20 | 0.088 | $0.09 \pm 0.02$ |
| 200 | 14 | 50 | 0.204 | $0.21 \pm 0.03$ |
| 200 | 14 | 100 | 0.366 | $0.36 \pm 0.04$ |
| 500 | 22 | 50 | 0.088 | $0.09 \pm 0.02$ |
| 500 | 22 | 100 | 0.169 | $0.17 \pm 0.03$ |
| 500 | 22 | 250 | 0.371 | $0.38 \pm 0.04$ |

### F.1 Protocol for the runtime decomposition

**Runtime decomposition.** Both pipelines start from the same feasible LP point, so the dense LP solve is shared and reported once. Dense post-LP time covers materializing all $\binom{n}{2}$ marginals and dense-marginal rounding; sparse post-LP time is the sum of sampled-marginal transfer, witness-index construction, and sparse pivot rounding. Post-LP speedup is dense post-LP time divided by sparse post-LP time, and total speedup is (LP + dense post-LP)/(LP + sparse post-LP). On every reported instance the cutting-plane solve ended at a feasible point rather than at the time limit; the $1{,}840$-second LP entry for $n = 5{,}000$ exceeds the $1{,}800$-second limit in Table 7 because it is pipeline wall-clock time, while the limit applies to the backend solver call only. The runtime instances are ones on which the dense LP solve completed, unlike the full benchmark graphs behind Table 5, which are handled through induced subgraphs. These gains are measurements of the current implementation and are not implied by the theorems.

Table 7: Default experimental configuration for the SBM, hidden-clique, and benchmark experiments. The same configuration is used for all reported LP-based experiments unless otherwise stated.

| Parameter | Value |
|---|---|
| Random seeds | $0, 1, \ldots, 19$ |
| Number of repetitions | 20 |
| SBM communities | 5 equal-sized communities |
| SBM within-cluster positive probability | 0.85 |
| SBM across-cluster negative probability | 0.85 |
| Default weights for SBM | $w_{uv} = 1$ |
| Hidden-clique size | $k = \lfloor \sqrt{n} \rfloor$ |
| Hidden-clique weight | $M = n^2$ |
| Marginal sample budgets | $m \in \{0.07, 0.14, 0.28, 0.50, 1.00, 1.50, 2.00\} n^{3/2}$ |
| LP backend | Gurobi 11.0.3 |
| LP feasibility tolerance | $10^{-7}$ |
| Triangle-separation tolerance | $10^{-6}$ |
| Cutting-plane stopping rule | no violated triangle inequality above $10^{-6}$ |
| Cutting-plane iteration cap | 200 rounds |
| LP time limit per instance | 1800 seconds |
| Sparse rounding repetitions per LP | 100 pivot-rounding runs |
| Hardware | 16-core AMD EPYC 7B13 CPU, 64GB RAM |
| Software | Python 3.11, NumPy 1.26, NetworkX 3.2, GurobiPy 11.0 |

## NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the main theoretical and empirical claims: the exact VC dimension $\mathrm{VC}(\mathcal{H}_G) = n - 1$ and additive $\tilde{O}(n/\varepsilon^2)$ edge coreset (Theorems 3.1, 3.2); the universal coarse and conditional sharp sparse-rounding guarantees for SPARSE-LP-PIVOT (Proposition 4.3, Theorem 4.5); the $m = \Theta(n^{3/2})$ triangle-witness threshold (Lemma 4.7); and the $\Omega(n^{3/2})$ lower bound for sparse edge observation with general weights (Theorem 5.1). The abstract explicitly distinguishes the unconditional coarse perturbation bound from the sharper influence-weighted bound that requires a rule-specific stability certificate; the main text labels the empirical use of $\overline{\Gamma}_w$ as a diagnostic rather than an unconditional certificate, and presents the runtime results of Section 6 as measurements on the tested instances.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 7 discusses three explicit limitations: the unconditional sparse-rounding bound is coarse because it charges every imputation perturbation by the total edge weight $W$, so it is informative only when $\sum_{u<v} \Gamma_S(u,v)$ is small; the same-edge diagnostic $\overline{\Gamma}_w$ becomes a certified guarantee only under the rule-specific weighted influence-stability certificate of Assumption 4.4 together with the domination condition $\Phi_{uv} \le c_\Phi w_{uv}$, and otherwise should be read as an empirical diagnostic; and the lower bound of Theorem 5.1 applies to uniform nonadaptive edge observation for arbitrary weights and does not preclude adaptive sampling, structural assumptions, or sparse LP-marginal access.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: Every theorem, lemma, and proposition is numbered and cross-referenced. Assumptions are stated in the relevant theorem statements or surrounding text: the LP-PIVOT abstraction in Assumption 2.1, the sparse LP-marginal access model in Section 4, the weighted influence-stability certificate in Assumption 4.4, the good-witness condition in Assumption D.1, and the uniform nonadaptive edge-observation model in Section 5. Proof sketches appear in the main text; complete proofs are in Appendix A (VC dimension), Appendix B (coreset and approximation transfer), Appendix C , Appendix D (good-witness corollary), and Appendix E (full proof of the $\Omega(n^{3/2})$ lower bound including the hidden-pairs lemma).

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Section 6 describes the three instance families (SBM, hidden-clique, and benchmark graphs) and the repetition count, and Appendix F reports the full experimental protocol: instance generation parameters, SBM community structure and edge probabilities, hidden-clique parameters $k = \lfloor \sqrt{n} \rfloor$ and $M = n^2$, the cutting-plane LP procedure, the 20 random seeds used, the LP solver and tolerances, and the per-LP rounding repetition count. Tables 4, 5, 6, and 3 report all numerical values, Appendix F.1 describes the runtime decomposition, Appendix F states which instances each table uses, and Table 7 fixes every remaining configuration parameter.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: The paper specifies the algorithms, instance generators, baselines, seeds, and solver configuration in detail (Section 6 and Appendix F, including Table 7), but it does not include a code release, packaged data-preparation scripts, or exact run commands. The benchmark datasets used (Political Blogs and Facebook ego networks) are public and individually citable.

   Guidelines:

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

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 6 specifies the three instance families, baselines (SPARSE-LP-PIVOT, full LP-PIVOT, KwikCluster), normalized sample budgets, and the diagnostic $\overline{\Gamma}_w$ that is reported. Appendix F expands on instance generation, the cutting-plane LP loop, and the per-instance pivot-rounding repetition count. Table 7 lists the SBM community parameters, hidden-clique parameters, sample-budget grid, LP feasibility and triangle-separation tolerances, cutting-plane stopping rule and iteration cap, and per-LP rounding repetition count. The paper does not train neural models; the auxiliary citation to GraphSAGE in the introduction is contextual rather than experimental.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: All randomized experiments are repeated 20 times with the seeds listed in Table 7, and Tables 4, 5, and 6 report mean $\pm$ standard error of the mean over those 20 seeds. The variability captured is over the random instance draw and the marginal sample $S$; the LP solve and dense baseline cost are deterministic given the instance, while the sparse rounding additionally averages over 100 pivot-rounding runs per LP. The $n = 50$ curve in Figure 2 matches Table 4, which reports standard errors; Figure 1 plots means without error bands. Table 3 reports wall-clock times without error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: Table 7 reports the hardware (16-core AMD EPYC 7B13 CPU, 64GB RAM), the software stack (Python 3.11, NumPy 1.26, NetworkX 3.2, GurobiPy 11.0, with Gurobi 11.0.3 as the LP backend), and a per-instance LP time limit of 1800 seconds together with a 200-round cap on the cutting-plane loop. The experiments do not require GPUs. The reported LP-based baselines are run only when the dense LP or cutting-plane solve completes within these limits, which together with the 20 repetitions and the 100 rounding runs per LP determines the total compute budget. Table 3 reports per-stage wall-clock times for $n \in \{1{,}000, 2{,}500, 5{,}000\}$.

Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: The work is theoretical and experimental, uses only public benchmark datasets (Political Blogs and Facebook ego networks via SNAP), does not collect or release new human-subject data, and does not release pretrained models, scraped corpora, or generative models. It does not introduce procedures with foreseeable safety, dual-use, or surveillance concerns.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: Section 7 includes a broader-impact paragraph. The contribution is foundational and access-model-specific; its intended benefit is cheaper large relational pipelines such as entity resolution. The paragraph notes that the same savings could make large-scale linkage or profiling of personal data cheaper and that sparse-rounding errors could fall unevenly on groups whose pairs are observed less often, and it names standard privacy safeguards and group-level audits of cluster quality as mitigations.

Guidelines:

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

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release pretrained language models, image generators, scraped corpora, or other high-risk artifacts. The contributed objects are theorems, an algorithm description, and an evaluation protocol on public benchmarks, so release safeguards do not apply.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The Political Blogs dataset [1], the Facebook ego-network data via SNAP [24, 25], and the LP backend (Gurobi 11.0.3, accessed via GurobiPy) are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release a new dataset, pretrained model, benchmark asset, or packaged software library. Its contributions are theorems, the SPARSE-LP-PIVOT algorithm description, and an experimental protocol, all documented in the main text and appendices.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, participant studies, or newly collected human-subject data. All datasets used are public, previously released graph benchmarks.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject experiments requiring IRB or equivalent review.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The core methodology (VC-dimension analysis, the additive coreset, the SPARSE-LP-PIVOT algorithm and its analysis, the witness-density threshold, and the lower bound) does not use LLMs as an important, original, or non-standard component. Any ordinary writing, editing, or formatting assistance does not require declaration under this checklist item.

    Guidelines:

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