Novelty Assessment Report

Paper: Sublinear Spectral Clustering Oracle with Little Memory
PDF URL: https://openreview.net/pdf?id=0GpolO2auw
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

We study the problem of designing sublinear spectral clustering oracles for well-clusterable graphs. Such an oracle is an algorithm that, given query access to the adjacency list of a graph G, first constructs a compact data structure $\mathcal{D}$ that captures the clustering structure of G. Once built, $\mathcal{D}$ enables sublinear time responses to \textsc{WhichCluster}(G,x) queries for any vertex X. A major limitation of existing oracles is that constructing $\mathcal{D}$ requires $\Omega(\sqrt{n})$ memory, which becomes a bottleneck for massive graphs and memory-limited settings. In this paper, we break this barrier and establish a memory-time trade-off for sublinear spectral clustering oracles. Specifically, for well-clusterable graphs, we present oracles that construct $\mathcal{D}$ using much smaller than $O(\sqrt{n})$ memory (e.g., $O(n^{0.01})$) while still answering membership queries in sublinear time. We also characterize the trade-off frontier between memory usage S and query time T, showing, for example, that $S\cdot T=\widetilde{O}(n)$ for clusterable graphs with a logarithmic conductance gap, and we show that this trade-off is nearly optimal (up to logarithmic factors) for a natural class of approaches. Finally, to complement our theory, we validate the performance of our oracles through experiments on synthetic networks.

Core Task Landscape

This paper addresses: **Sublinear Spectral Clustering with Memory Constraints**

A total of **15 papers** were analyzed and organized into a taxonomy with **9 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Sublinear-Time Clustering Oracles**
- **Sequential and Streaming Approaches**
- **Memory-Efficient Batch Methods**
- **Theoretical Foundations and Cross-Domain Extensions**

Complete Taxonomy Tree

- Sublinear Spectral Clustering with Memory Constraints Survey Taxonomy
- Sublinear-Time Clustering Oracles
  - Memory-Time Trade-off Oracles ★ (2 papers)
  - [0] Sublinear Spectral Clustering Oracle with Little Memory (Anon et al., 2026) View paper
  - [15] SUBLINEAR SPECTRAL CLUSTERING ORACLE WITH (MEMORY, n.d.) View paper
  - Standard Sublinear Oracles (2 papers)
  - [3] Spectral Clustering Oracles in Sublinear Time (Grzegorz GÅuch, 2021) View paper
  - [5] A Sublinear-Time Spectral Clustering Oracle with Improved Preprocessing Time (Shen, 2023) View paper
- Sequential and Streaming Approaches
  - Streaming Spectral Clustering (2 papers)
  - [2] Streaming spectral clustering (Shinjae Yoo, 2016) View paper
  - [8] Streaming, memory limited algorithms for community detection (Yun, 2014) View paper
  - Sequential Batch Processing (1 papers)
  - [7] Scalable sequential spectral clustering (Yeqing Li, 2016) View paper
- Memory-Efficient Batch Methods
  - Column Sampling and Nyström Approximation (1 papers)
  - [9] Time and space efficient spectral clustering via column sampling (Mu Li, 2011) View paper
  - Incremental and Cosine Similarity Methods (3 papers)
  - [4] A fast incremental spectral clustering algorithm with cosine similarity (Ran Li, 2023) View paper
  - [6] On The Memory Scalability of Spectral Clustering Algorithms (Ran Li, 2023) View paper
  - [11] Fast, Memory-Efficient Spectral Clustering with Cosine Similarity (R. F. Li, 2023) View paper
  - Multiscale and Hierarchical Decomposition (1 papers)
  - [13] Image Segmentation Based on Multiscale Fast Spectral Clustering (Chongyang Zhang, 2018) View paper
- Theoretical Foundations and Cross-Domain Extensions
  - General Sublinear Algorithm Theory (1 papers)
  - [1] Sublinear Algorithms for Matrices: Theory and Applications (Ray, 2024) View paper
  - Specialized Extensions (3 papers)
  - [10] A Fast and Memory-Efficient Spectral Library Search Algorithm Using Locality-Sensitive Hashing. (Lei Wang, 2021) View paper
  - [12] DIFFERENTIALLY PRIVATE SUBLINEAR ALGORITHMS (Mukherjee, 2023) View paper
  - [14] What and How: Generalized Lifelong Spectral Clustering via Dual Memory. (Gan Sun, n.d.) View paper

Narrative

Core task: sublinear spectral clustering with memory constraints. The field addresses the challenge of partitioning large-scale graphs or datasets when both computational time and memory are severely limited. The taxonomy organizes research into four main branches. Sublinear-Time Clustering Oracles focus on query-based models that answer cluster membership questions without processing the entire input, often trading preprocessing cost against per-query efficiency. Sequential and Streaming Approaches handle data that arrives incrementally or cannot be stored in full, exemplified by works like Streaming Spectral Clustering[2] and Sequential Spectral Clustering[7]. Memory-Efficient Batch Methods develop techniques such as column sampling and locality-sensitive hashing to reduce space usage while processing the whole dataset, as seen in Column Sampling Clustering[9] and Locality Sensitive Hashing[10]. Theoretical Foundations and Cross-Domain Extensions explore privacy-preserving variants, multiscale strategies, and algorithmic guarantees that span multiple problem settings.

Within the oracle-based branch, a particularly active line of work investigates memory-time trade-offs: how much preprocessing and storage one must invest to achieve fast query responses. Sublinear Spectral Oracle[0] sits squarely in this cluster, emphasizing oracles that balance preprocessing overhead with per-query complexity under strict memory budgets. It shares thematic ground with Spectral Clustering Oracles[3] and Improved Preprocessing Oracle[5], both of which also study preprocessing strategies and query efficiency. Compared to these neighbors, Sublinear Spectral Oracle[0] appears to push further on explicit memory constraints, exploring regimes where even moderate storage is unavailable. Meanwhile, works like Memory Scalability Spectral[6] and Dual Memory Clustering[14] tackle related scalability questions in batch or hybrid settings, highlighting an ongoing tension between oracle flexibility and the practicality of storing auxiliary structures.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. SUBLINEAR SPECTRAL CLUSTERING ORACLE WITH

Authors: L MEMORY | **URL**: View paper

Abstract

â¦ Thus, developing memory-efficient clustering algorithms is not â¦ local clustering oracles that run in sublinear time and space. â¦ paper does not explicitly state the space complexity, it can be â¦

⚠ Similarity Notice

This paper appears to be the same as or a near-duplicate of the original paper. Both papers have identical titles ('Sublinear Spectral Clustering Oracle with Little Memory'), identical abstracts describing the same memory-time trade-off results ($S \cdot T = \tilde{O}(n)$), identical technical content including the same algorithms (ESTRWDOT, ESTCOLLIPROB, INITORACLE, QUERYDOT), identical theorem statements (Theorem 1.1, 3.1, 3.2), and identical experimental validation claims. The papers share the same core contribution of breaking the $\Omega(\sqrt{n})$ space barrier and achieving sublinear memory usage (e.g., O(n^0.01)) while maintaining sublinear query time.

Contributions Analysis

Overall novelty summary. The paper introduces sublinear spectral clustering oracles that break the $\Omega(\sqrt{n})$ memory barrier, achieving memory usage as low as O(n^0.01) while maintaining sublinear query time. Within the taxonomy, it resides in the 'Memory-Time Trade-off Oracles' leaf under 'Sublinear-Time Clustering Oracles'. This leaf contains only two papers total, indicating a relatively sparse research direction. The sibling paper in this leaf shares the focus on explicit memory-time trade-offs, suggesting this is an emerging rather than crowded area of investigation.

The taxonomy reveals that the broader 'Sublinear-Time Clustering Oracles' branch contains a 'Standard Sublinear Oracles' leaf with papers accepting $O(\sqrt{n})$ or higher memory usage. Neighboring branches include 'Sequential and Streaming Approaches' (processing data incrementally) and 'Memory-Efficient Batch Methods' (using sampling or decomposition). The paper's oracle-based query model distinguishes it from streaming methods that process continuous arrivals, and from batch methods that require full dataset access. Its focus on preprocessing into compact structures for repeated queries positions it between pure streaming and traditional batch paradigms.

Among 24 candidates examined, the batch-based collision probability estimation algorithm (Contribution 2) shows one refutable candidate among 10 examined, suggesting some overlap with prior sampling techniques. The core oracle contribution (Contribution 1) examined 8 candidates with none refutable, and the lower bound result (Contribution 3) examined 6 candidates with none refutable. The limited search scope means these statistics reflect top-semantic-match coverage rather than exhaustive field analysis. The collision estimation component appears to have more substantial prior work than the memory-constrained oracle framework itself.

Given the sparse taxonomy leaf and limited refutation signals across most contributions, the work appears to occupy a relatively novel position within the constrained search scope. The explicit memory-time trade-off characterization and sub-$\sqrt{n}$ memory regimes represent a clear departure from standard oracle designs. However, the analysis covers approximately 24 papers from semantic search, leaving open whether deeper literature exploration would reveal additional overlapping work in related algorithmic frameworks or theoretical computer science venues.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Sublinear spectral clustering oracle with memory-time trade-off

Description: The authors design a spectral clustering oracle that breaks the previous $\Omega(\sqrt{n})$ memory barrier by introducing a flexible trade-off between memory usage S and query time T, achieving $S \cdot T = \tilde{O}(n)$ for well-clusterable graphs. This allows constructing the data structure D with substantially less memory while maintaining sublinear query times.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. SBSC: A fast Self-tuned Bipartite proximity graph-based Spectral Clustering

URL: View paper

Brief Assessment

Bipartite Spectral Clustering[31] focuses on efficient spectral clustering through bipartite graph construction for large datasets, not on sublinear oracles with memory-time trade-offs for query-based clustering.

---

2. Variational Perspective on Local Graph Clustering

URL: View paper

Brief Assessment

Local Graph Clustering[35] focuses on local spectral graph clustering methods that identify clusters around seed nodes without accessing the entire graph, using personalized PageRank. This differs from the original paper's focus on constructing a global data structure D for arbitrary vertex queries with explicit memory-time trade-offs.

---

3. A Sublinear Time Tester for Max-Cut on Clusterable Graphs

URL: View paper

Brief Assessment

MaxCut Clusterable Tester[33] focuses on distinguishing max-cut values in clusterable graphs using spectral properties, not on designing spectral clustering oracles with memory-time trade-offs for cluster membership queries.

---

4. Learning Hierarchical Cluster Structure of Graphs in Sublinear Time

URL: View paper

Brief Assessment

Hierarchical Cluster Learning[34] focuses on hierarchical clustering structures in graphs, not on memory-time trade-offs for flat spectral clustering oracles. The candidate addresses a different problem setting with hierarchical models.

---

5. Spectral Clustering Oracles in Sublinear Time

URL: View paper

Brief Assessment

Spectral Clustering Oracles[3] focuses on achieving sublinear query time through dot product access to spectral embeddings, but does not explicitly address memory-time trade-offs or breaking the $\Omega(\sqrt{n})$ memory barrier that is central to the original paper's contribution.

---

6. SUBLINEAR SPECTRAL CLUSTERING ORACLE WITH

URL: View paper

Brief Assessment

Sublinear Spectral Oracle[15] is the same paper as the original submission being reviewed. The candidate paper title explicitly indicates this is the same work, making comparison meaningless for novelty assessment.

---

7. Scalable Constrained Clustering: A Generalized Spectral Method

URL: View paper

Brief Assessment

Constrained Spectral Method[36] addresses constrained clustering with must-link/cannot-link constraints using spectral methods, not sublinear oracles with memory-time trade-offs for query-based clustering on well-clusterable graphs.

---

8. Distributed graph clustering by load balancing

URL: View paper

Brief Assessment

Load Balancing Clustering[32] focuses on distributed graph clustering through load balancing techniques, not on sublinear spectral clustering oracles with memory-time trade-offs for well-clusterable graphs.

Contribution 2: Batch-based collision probability estimation algorithm

Description: The authors introduce a new batch-based estimation strategy (ESTCOLLIPROB) that partitions random walks into batches to estimate collision probabilities of random walk distributions. This technique enables the memory-efficient dot product oracle that underlies their clustering oracle, achieving the desired space-time trade-off.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Collisions or Adsorption: An Electrochemical Random Walk Decides

URL: View paper

Brief Assessment

Electrochemical Random Walk[28] focuses on modeling electrochemical droplet collisions and adsorption events using random walk simulations, not on batch-based estimation strategies for clustering algorithms in graph analysis.

---

2. Sampling hypergraphs via joint unbiased random walk

URL: View paper

Brief Assessment

Hypergraph Random Walk[21] focuses on sampling hypergraphs via random walks and collision-based estimation for hypergraph structures, not on batch-based partitioning strategies for memory-efficient dot product oracles in graph clustering contexts.

---

3. Peer counting and sampling in overlay networks: random walk methods

URL: View paper

Brief Assessment

Peer Counting Sampling[22] focuses on peer counting in overlay networks using random walks for sampling and collision detection, not on batch-based estimation of collision probabilities in random walk distributions for clustering oracles. The technical contexts are fundamentally different.

---

4. Random-walk modeling of turbulent impaction to a smooth wall

URL: View paper

Brief Assessment

Turbulent Impaction Modeling[29] addresses turbulent particle impaction on walls using random-walk simulations in fluid dynamics, not graph clustering or collision probability estimation for random walk distributions on graphs.

---

5. Discrete-Event Simulation of Aggregation Processes in Batch and Flow Tubular Reactors Based on the Stochastic Lattice Models.

URL: View paper

Brief Assessment

The candidate paper (Lattice Aggregation Simulation[25]) focuses on discrete-event simulation of particle aggregation in chemical reactors using stochastic lattice models, not on collision probability estimation for random walk distributions in graph clustering contexts.

---

6. Elimination of the reaction rate â scale effectâ: Application of the Lagrangian reactive particleâ tracking method to simulate mixingâ limited, fieldâ scale biodegradation at â¦

URL: View paper

Brief Assessment

Lagrangian Particle Tracking[26] focuses on simulating mixing-limited biodegradation using particle-tracking methods for reactive transport, not on estimating collision probabilities in random walk distributions for graph clustering algorithms.

---

7. Estimating sizes of social networks via biased sampling

URL: View paper

Brief Assessment

Biased Network Sampling[24] focuses on estimating social network sizes through collision counting in random walk samples, not on batch-based estimation strategies for clustering oracles or memory-efficient dot product computation in spectral clustering contexts.

---

8. On the optimal analysis of the collision probability tester

URL: View paper

Brief Assessment

Collision Probability Tester[30] focuses on analyzing collision probability estimation for distribution testing (distinguishing uniform from non-uniform distributions), not on batch-based strategies for random walk distributions in graph clustering contexts.

---

9. Stochastic streams: Sample complexity vs. space complexity

URL: View paper

Prior Art Analysis

Stochastic Streams Complexity[27] demonstrates that batch-based estimation strategies for collision probability were already established in prior work. The candidate paper presents a batch-based approach that partitions samples into batches to estimate collision probability with space-time trade-offs, achieving $s \cdot t \approx \tilde{O}(n)$. This directly refutes the novelty claim that the original paper was first to introduce batch-based estimation for collision probabilities in random walk distributions, as the candidate shows this technique was already known in the stochastic streaming literature.

Evidence

Evidence 1 - **Rationale**: Both papers describe partitioning samples into batches/groups to estimate collision probability. The candidate paper's approach of partitioning into t/w groups and computing estimates per group directly parallels the original paper's batch-based strategy of partitioning r walks into b=r/m batches. - **Original**: we propose a batch-based estimation strategy. The idea behind this approach is inspired by canonne & yang (2024), where a similar batching technique is used to design memory-efficient algorithms for uniformity testing under memory constraints. While the underlying technique is inspired by prior work... - **Candidate**: our algorithm partitions the input stream intot/w groups ofw consecutive samples. we now analyze one specific group ofw samples. suppose the samples area1,...,aw. for each pair⁄i= j ∈{1,2,...,w}, letxi,j be an indicator random variable which is1 iff ai = aj. let x = 1 (w 2) ∑⁄i=jxi,j. then e[x] is...

Evidence 2 - **Rationale**: The original paper's Algorithm 1 shows the batch-based approach with b batches computing estimates zb and averaging. The candidate describes running the algorithm on independent groups of samples and taking median, which is the same batch-based estimation principle for collision probability. - **Original**: algorithm 1:estrwdot (g, r, t, m, x, y) 1z := 0, b:= r m ▷ b: number of batch 2forb= 1tobdo 3 run m independent random walks of lengthtstarting fromx(resp. from y) 4define bpx(i)(resp. bpy(i)) as the fraction of randoms walks fromx (resp. fromy) that end ati 5z b :=⟨ bpx, bpy⟩,z :=z+z b 6z := z b 7r... - **Candidate**: we can assumeδis a constant, since the algorithm generalizes to smallerδsimply by increasing the sample complexity by a factor oflog(1/δ), running our algorithmlog(1/δ) times in sequence on each oflog(1/δ) independent groups of samples, and taking the median. note t= ω(n1/2) was shown necessary by ...

Evidence 3 - **Rationale**: Both establish space-time trade-offs for collision probability estimation using batch-based approaches. The original paper's Lemma 3.1 shows the batch algorithm achieves error bounds with specific space o(m·logn) and time o(rt), while the candidate's Theorem 1 shows similar space-time trade-offs for estimating f2 (collision probability). - **Original**: lemma 3.1.let k≥2 be an integer and φ, ε∈(0,1) . let g= (v, e) be a d-regular and (k, φ, ε)-clusterable graph. let m be the random walk transition matrix of g. let z be the output ofestrwdot (g, r, t, m, x, y)(alg. 1). let σerr >0 . let c >1 be a large enough constant. for any t≥ 20 logn φ2 and any ... - **Candidate**: ▶ theorem 1. for t= ˜ω ϵ,δ(n1/2), estimatingf2 up to a factor(1 ±ϵ) given t samples with probability at least1 -δis possible in˜oϵ,δ(1 + n/t) bits of space. proof. we can assumeδis a constant, since the algorithm generalizes to smallerδsimply by increasing the sample complexity by a factor oflog(1/δ...

---

10. Ant-inspired density estimation via random walks

URL: View paper

Brief Assessment

Ant Density Estimation[23] focuses on density estimation via encounter rates in ant colonies using random walks on grids, not on batch-based estimation strategies for collision probabilities in spectral clustering contexts. The candidate addresses biological computation and collision rates between agents, while the original paper develops memory-efficient algorithms for graph clustering oracles.

Contribution 3: Nearly tight lower bound for 1-cluster vs. 2-cluster distinction

Description: The authors establish a space-time lower bound $S \cdot T = \Omega(n)$ for distinguishing between single-cluster expanders and two-cluster graphs using random walk queries. This lower bound demonstrates that their upper bound trade-off is nearly optimal up to polylogarithmic factors for collision-probability-based approaches.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Balancing a complete signed graph by editing edges and deleting nodes

URL: View paper

Brief Assessment

Signed Graph Balancing[19] addresses a completely different problem domain (signed graph balancing through edge editing and node deletion) and does not discuss space-time lower bounds for distinguishing cluster structures in graphs.

---

2. Testing cluster structure of graphs

URL: View paper

Brief Assessment

Testing Cluster Structure[16] focuses on property testing for clusterable graphs with a different problem formulation (testing (k,ϕ)-clusterability), not on space-time lower bounds for distinguishing single-cluster versus two-cluster graphs using random walk queries.

3. Testing graph clusterability: Algorithms and lower bounds

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

4. Improved graph clustering

URL: View paper

Brief Assessment

Improved Graph Clustering[17] focuses on exact recovery guarantees for graph clustering under stochastic block models, not on space-time lower bounds for distinguishing cluster configurations using random walk queries.

5. SUBLINEAR SPECTRAL CLUSTERING ORACLE WITH

URL: View paper

Brief Assessment

Sublinear Spectral Oracle[15] is the same paper as the original submission. The candidate paper title confirms this is the identical work, so no prior work comparison is possible.

6. New lower bounds for Massively Parallel Computation from query complexity

URL: View paper

Brief Assessment

MPC Lower Bounds[20] focuses on massively parallel computation lower bounds for distinguishing 1-cycle vs. 2-cycle graphs, not on random walk query complexity for single-cluster vs. two-cluster expanders in the context of spectral clustering oracles.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Sublinear Spectral Clustering Oracle with Little Memory View paper
- [1] Sublinear Algorithms for Matrices: Theory and Applications View paper
- [2] Streaming spectral clustering View paper
- [3] Spectral Clustering Oracles in Sublinear Time View paper
- [4] A fast incremental spectral clustering algorithm with cosine similarity View paper
- [5] A Sublinear-Time Spectral Clustering Oracle with Improved Preprocessing Time View paper
- [6] On The Memory Scalability of Spectral Clustering Algorithms View paper
- [7] Scalable sequential spectral clustering View paper
- [8] Streaming, memory limited algorithms for community detection View paper
- [9] Time and space efficient spectral clustering via column sampling View paper
- [10] A Fast and Memory-Efficient Spectral Library Search Algorithm Using Locality-Sensitive Hashing. View paper
- [11] Fast, Memory-Efficient Spectral Clustering with Cosine Similarity View paper
- [12] DIFFERENTIALLY PRIVATE SUBLINEAR ALGORITHMS View paper
- [13] Image Segmentation Based on Multiscale Fast Spectral Clustering View paper
- [14] What and How: Generalized Lifelong Spectral Clustering via Dual Memory. View paper
- [15] SUBLINEAR SPECTRAL CLUSTERING ORACLE WITH View paper
- [16] Testing cluster structure of graphs View paper
- [17] Improved graph clustering View paper
- [18] Testing graph clusterability: Algorithms and lower bounds View paper
- [19] Balancing a complete signed graph by editing edges and deleting nodes View paper
- [20] New lower bounds for Massively Parallel Computation from query complexity View paper
- [21] Sampling hypergraphs via joint unbiased random walk View paper
- [22] Peer counting and sampling in overlay networks: random walk methods View paper
- [23] Ant-inspired density estimation via random walks View paper
- [24] Estimating sizes of social networks via biased sampling View paper
- [25] Discrete-Event Simulation of Aggregation Processes in Batch and Flow Tubular Reactors Based on the Stochastic Lattice Models. View paper
- [26] Elimination of the reaction rate â□□scale effectâ□□: Application of the Lagrangian reactive particleâ□□tracking method to simulate mixingâ□□limited, fieldâ□□scale biodegradation at â□¦ View paper
- [27] Stochastic streams: Sample complexity vs. space complexity View paper
- [28] Collisions or Adsorption: An Electrochemical Random Walk Decides View paper
- [29] Random-walk modeling of turbulent impaction to a smooth wall View paper
- [30] On the optimal analysis of the collision probability tester View paper
- [31] SBSC: A fast Self-tuned Bipartite proximity graph-based Spectral Clustering View paper
- [32] Distributed graph clustering by load balancing View paper
- [33] A Sublinear Time Tester for Max-Cut on Clusterable Graphs View paper
- [34] Learning Hierarchical Cluster Structure of Graphs in Sublinear Time View paper
- [35] Variational Perspective on Local Graph Clustering View paper
- [36] Scalable Constrained Clustering: A Generalized Spectral Method View paper