Preserving Manifold Structure in Landmark Coresets

Morse-Seeded Sampling on kNN Graphs

Tirth Joshi
Katz School of Science and Health,
Yeshiva University
tirth.joshi@yu.edu

Core result at 10× compression ($n = 2500, m = 250$): MorseSeeded-FF preserves FarthestFirst-level coverage on Swiss Roll and Two Moons, while SCC-aware variants increase recall of high-boundary structural nodes. The method targets **coverage + graph structure**, not only minimum average distortion.

1. Problem and Definitions

Landmark coreset. Given points $X = \{x_i\}_{i=1}^n$, build a weighted kNN graph $G = (V, E)$. A landmark coreset is a subset

$$L \subset V, \qquad |L| = m \ll n.$$

In the main experiment, $m = 250$ for $n = 2500$, i.e., a **10× reduction**. The sweep tests $m \in \{75, 125, 175, 250, 350\}$, or about **3%-14%** of the nodes.

Coverage. A coreset should keep every point close to a representative:

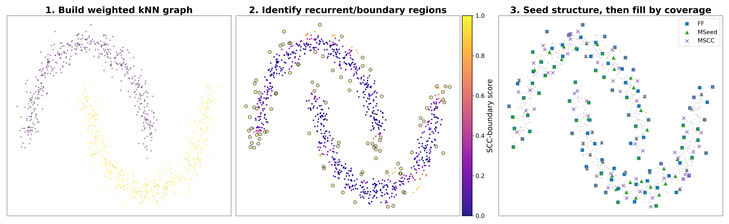
$$r(L) = \max_{i \in V} \min_{\ell \in L} \|x_i - x_\ell\|_2.$$

Structure-preserving. Let $d_G(i, j)$ be the shortest-path distance on the full kNN graph and $\widehat{d}_L(i, j)$ the landmark approximation. We evaluate

$$E_{\text{geo}}(i, j) = \left| \frac{\widehat{d}_L(i, j) - d_G(i, j)}{d_G(i, j)} \right|.$$

Structure is preserved when coverage remains competitive *and* selected landmarks represent recurrent/boundary graph regions.

2. Conceptual Method: From Graph Structure to Seeds



The sampler first exposes graph structure, then uses structural seeds before a global farthest-first fill. This replaces a purely distance-based selection with a structure-aware initialization.

3. Algorithm and Metrics

Morse/SCC proxy. Define a local density potential by the k-NN radius

$$\Phi(i) = \|x_i - x_{i,k}\|_2,$$

where $x_{i,k}$ is the k-th nearest neighbor of x_i. Lower $\Phi(i)$ indicates a locally denser region.
Construct a directed flow $G^{\downarrow}$ on the same vertex set by adding $i \to j$ when

$$j \in \text{kNN}(i) \quad \text{and} \quad (\Phi(j) \le \Phi(i)) \text{ or } (|\Phi(j) - \Phi(i)| \le \varepsilon).$$

The first condition moves toward denser regions; the plateau condition connects near-equal-density neighbors. Strongly connected components (SCCs) of $G^{\downarrow}$ are used as a discrete Morse-style proxy for recurrent graph regions.

MorseSeeded-FF. Let $\mathcal{C}$ be the SCC partition. For each node, define a boundary score

$$b(i) = \frac{|\{j \in \text{kNN}(i) : C(j) \neq C(i)\}|}{|\text{kNN}(i)|}.$$

We choose seed landmarks from large SCCs using high boundary score, then run global farthest-first initialized by these seeds until $|L| = m$:

$$S \leftarrow \text{TopBoundary}(\mathcal{C}), \qquad L \leftarrow \text{FF}(X, S, m).$$

This keeps the coverage behavior of farthest-first while forcing early representation of recurrent or transition regions.

Landmark geodesic approximation. Let

$$a(i) = \arg\min_{\ell \in L} \|x_i - x_\ell\|_2$$

be the nearest landmark to x_i. We approximate full-graph geodesics by

$$\widehat{d}_L(i, j) = \|x_i - x_{a(i)}\|_2 + d_{G_L}(a(i), a(j)) + \|x_j - x_{a(j)}\|_2,$$

where d_{G_L} is shortest-path distance on the landmark kNN graph.

Heuristic error decomposition.

$$E_{\text{geo}}(i, j) = \left| \frac{\widehat{d}_L(i, j) - d_G(i, j)}{d_G(i, j)} \right| \approx \frac{\underbrace{\|x_i - x_{a(i)}\|_2 + \|x_j - x_{a(j)}\|_2}_{\text{coverage attachment}} + \underbrace{|d_{G_L}(a(i), a(j)) - d_G(i, j)|}_{\text{structural mismatch}}}{d_G(i, j)}.$$

Thus, low distortion requires both small attachment cost and a landmark graph that preserves the relevant kNN connectivity.

Boundary recall at top 10%. Let B_{10} be the top 10% of all nodes ranked by SCC-boundary score $b(i)$. We report

$$R_{10}(L) = \frac{|L \cap B_{10}|}{|B_{10}|}.$$

Here, 10% refers to **nodes ranked by boundary score**, not the coreset size. Higher R_{10} means the sampler better captures transition, interface, or recurrent-boundary regions.

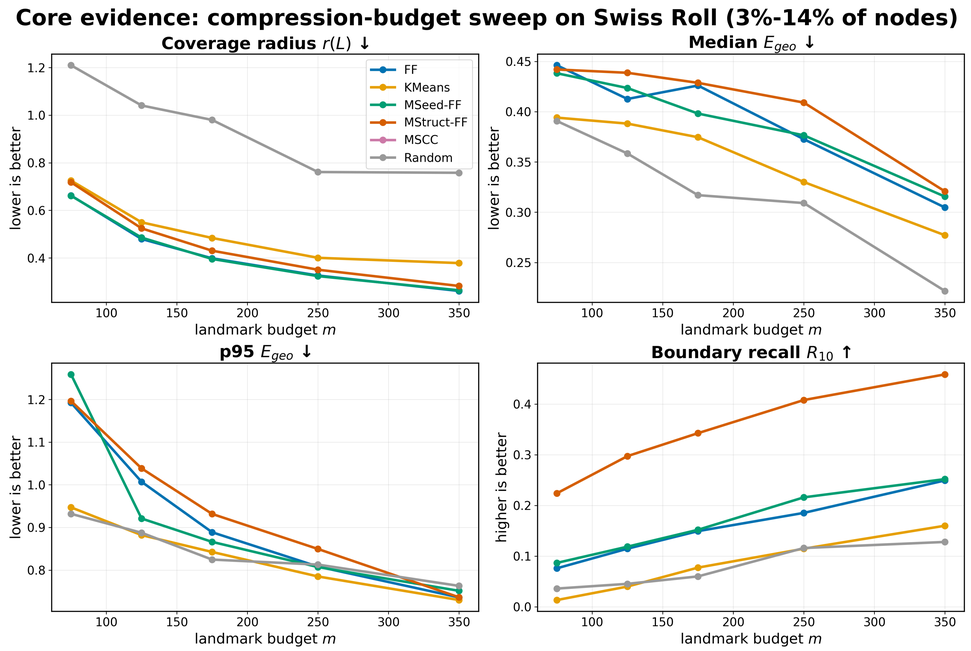
4. Main Empirical Result

Claim supported by the data. At 10× reduction, MorseSeeded-FF matches FarthestFirst coverage on Swiss Roll (0.324 vs. 0.326) and Two Moons (0.118 vs. 0.118). SCC-aware variants improve boundary recall to 0.417 on Swiss Roll and 0.492 on Two Moons, revealing structural nodes that coverage-only selection often misses.

Dataset	FF cover	MSeed cover	Best median E_{geo}	Best p95 E_{geo}	Best R_{10}	Interpretation
Swiss Roll	0.326	**0.324**	Random 0.310	KMeans 0.781	MSCC 0.417	Coverage preserved; structure recalled
Two Moons	0.118	**0.118**	Random 0.124	Random 0.601	MSCC 0.492	Strongest boundary signal
Digits PCA	5.257	**5.257**	KMeans 0.153	KMeans 0.625	MSCC 0.167	Real-data signal is weaker

Means across seeds. Lower is better for coverage/error. Higher is better for R_{10}. The table intentionally separates structural recall from pure geodesic error, because they are not the same objective.

5. Core Evidence: Compression-Budget Sweep



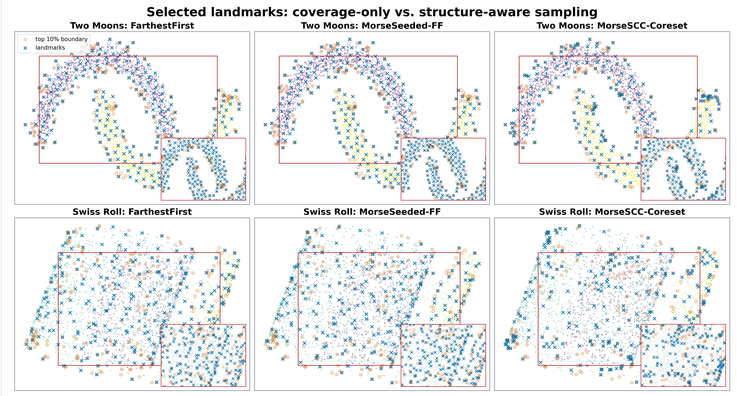
This is the main performance story: as the landmark budget grows, MorseSeeded-FF stays close to FarthestFirst in coverage, while SCC-aware methods increase boundary recall. Distortion decreases with budget but does not uniformly favor structural sampling.

6. Complete Error Profile: Mean, Median, p95, Max

Swiss Roll	mean	median	p95	max
FF	0.411	0.382	0.801	9.193
KMeans	0.388	0.346	0.788	6.254
MSeed-FF	0.413	0.379	0.792	**3.654**
MSCC	0.408	0.346	0.840	8.069
Random	0.390	0.305	0.798	13.380

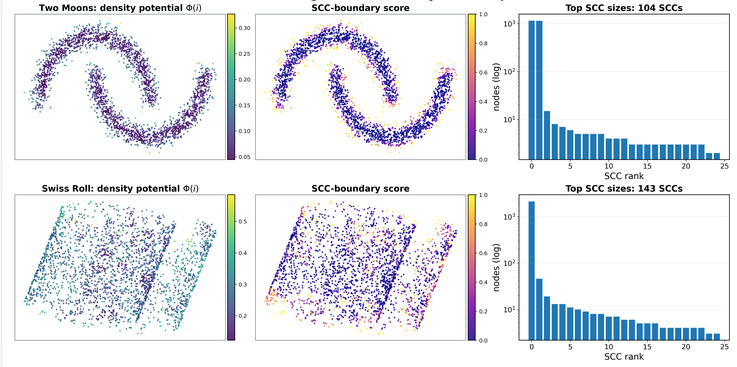
Relative to what? Error is computed against the full kNN shortest-path distance $d_G(i, j)$.
Why report max? Median alone hides outliers. On Swiss Roll, MSeed-FF has competitive median/p95 behavior and the lowest observed maximum error, suggesting fewer catastrophic landmark-geodesic failures in this run.
Reading the table. Mean reflects average distortion, median captures the typical case, p95 highlights tail behavior, and max shows worst-case failures. Reporting all four avoids over-claiming from a single summary statistic.

7. Visual Evidence: What Gets Selected?



Orange rings mark the top 10% boundary-score nodes. Insets zoom into boundary-heavy regions. Structural methods place more landmarks near recurrent/boundary regions, while FarthestFirst mainly spreads points for coverage.

8. Morse/SCC Diagnostics: What to Look For



- **Boundary signature:** high boundary-score nodes appear near sparse transitions, manifold ends, and interfaces between SCCs.
- **Recurrent signature:** large SCCs indicate dominant recurrent basins; small SCCs often capture local boundary or transition fragments.
- **Selection goal:** raise R_{10} without losing FarthestFirst-level coverage.

9. Discussion and Takeaways

Contribution. We propose a structure-aware landmarking method for kNN graphs: use a Morse/SCC proxy to seed graph-relevant regions, then use farthest-first traversal to keep coverage competitive.
Empirical signal. On structured datasets, the method preserves coverage while improving representation of high-boundary regions. The strongest evidence is the budget sweep plus visual boundary recall: structural sampling changes *which graph roles* are represented.
How to read the evidence.

- **Where it helps most:** Swiss Roll and Two Moons show the clearest structural benefit because boundary or transition regions are visually and graph-topologically meaningful.
- **What improves:** SCC-aware methods raise R_{10} by placing more landmarks near recurrent/boundary regions, even when average geodesic error does not uniformly improve.
- **What weakens on real data:** Digits PCA has a weaker structural signal, suggesting that density-flow SCCs are most helpful when the manifold geometry is clearer.

Limitation. Structural recall and geodesic distortion are not identical objectives. The current method exposes this trade-off rather than hiding it. Future work should use multi-scale Morse summaries, adaptive budget allocation, and a reproducible public code package.

10. Final Contribution, Empirical Signal, and Limitation

Contribution

MorseSeeded-FF turns landmarking into a graph-structural sampling problem: compute SCCs of a directed density flow, seed graph-relevant regions, then fill by farthest-first coverage.

Empirical signal

At **10× compression**, coverage remains near FarthestFirst on Swiss Roll and Two Moons, while SCC-aware variants improve recall of recurrent/boundary nodes.

Limitation

Structural recall and geodesic distortion are different objectives. Next: multi-scale Morse summaries, adaptive budget allocation, and a public reproducible package.

Tirth Joshi
Katz School of Science and Health
Yeshiva University

Preserving manifold structure in landmark coresets on kNN graphs
tirth.joshi@yu.edu

Code