Morse-Seeded Coresets for Structure-Preserving Landmarking on kNN Graphs

Anonymous submission to NSIA (NetSci 2026 Satellite)

Motivation. Coresets/landmarks are often selected to optimize geometric coverage (e.g., farthest-first), yet many downstream tasks depend on *graph-geodesic* structure induced by data connectivity. We ask: can we bias landmark selection using a coarse dynamical/topological summary of a directed flow on the kNN graph, while keeping near-optimal coverage?

Setup. Given points $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$, build a weighted undirected kNN graph $G = (V, E)$ with edge weights $w_{ij} = \|x_i - x_j\|$. Let $d_G(i, j)$ denote shortest-path distance in G. We seek a landmark set $L \subset V$ ($|L| = m \ll n$) such that a landmark-based approximation $\hat{d}(i, j)$ preserves $d_G(i, j)$.

Directed flow and Morse proxy. Define a scalar potential using a local density proxy

$$\Phi(i) \;=\; \text{distance from } x_i \text{ to its } k\text{th nearest neighbor.}$$

Construct a directed "flow" graph $G^{\downarrow}$ on the same vertex set by adding edges

$$i \to j \quad \text{if} \quad j \in \mathrm{kNN}(i) \;\text{ and }\; \big(\Phi(j) \le \Phi(i)\big) \;\text{ or }\; |\Phi(j) - \Phi(i)| \le \varepsilon,$$

where the plateau condition ($\varepsilon > 0$) induces recurrent components. We take strongly connected components (SCCs) of $G^{\downarrow}$ as a *combinatorial Morse-decomposition proxy* (not a full Conley index computation).

MorseSeeded-FF. Let $\mathcal{C}$ be the SCC partition. We select seeds S by choosing one representative from each of the largest SCCs until covering 95% of nodes, ignoring SCCs of size < 30 (up to 20 seeds). Representatives are chosen using a boundary score: fraction of kNN neighbors outside the SCC. We then run global farthest-first initialized by S to fill the remaining budget to m. This keeps coverage close to farthest-first while forcing representation across recurrent regions.

Landmark geodesic approximation. Let $a(i) = \arg\min_{\ell \in L} \|x_i - x_\ell\|$ and let $d_L(\cdot, \cdot)$ be shortest-path distance on the landmark kNN graph. We use

$$\hat{d}(i, j) \;=\; \|x_i - x_{a(i)}\| \;+\; d_L\big(a(i), a(j)\big) \;+\; \|x_j - x_{a(j)}\|.$$

We report absolute relative error

$$|e(i, j)| = \left| \frac{\hat{d}(i, j) - d_G(i, j)}{d_G(i, j)} \right|$$

on sampled pairs.

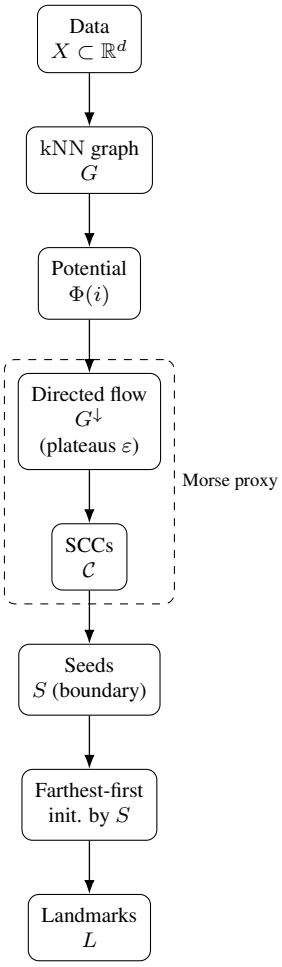

Figure 1. Compact pipeline for MorseSeeded-FF.

Experiments. We evaluate on Swiss Roll with n=2500, m=250 (compression 10×), k=12, plateau ε=0.02, and a fixed pair set (900 pairs) for distortion evaluation. Baselines: Random, global Farthest-First (FF), and k-means centers (mapped to nearest data points). Coverage is measured by the covering radius $\max_{i \in V} \min_{\ell \in L} \|x_i - x_\ell\|$.

Method	Cover radius ↓	median $\|e\|$ ↓	p95 $\|e\|$ ↓
Random	0.7017	0.0539	0.5813
FF (global)	0.3233	0.0815	0.7595
k-means centers	0.4525	0.0412	0.6908
MorseSCC-Coreset (ablation; per-SCC budgeting)	0.4523	0.0129	0.7940
MorseSeeded-FF (ours)	**0.3209**	**0.0460**	**0.7007**

Table 1. Swiss Roll (k=12, ε=0.02, n=2500, m=250). Absolute relative error $|e|$ on 900 fixed pairs. Lower is better.

Key observation. MorseSeeded-FF improves geodesic fidelity while preserving coverage: median absolute relative error drops from 8.15% (FF) to 4.60% with essentially unchanged covering radius ($0.3233 \to 0.3209$), and tail error improves (p95 $|e| : 0.7595 \to 0.7007$). In contrast, per-SCC budgeting (MorseSCC-Coreset) can reduce typical distortion but degrades coverage due to many small SCCs, motivating seeded integration with global farthest-first.

Relevance to NSIA. The method treats learning-relevant geometry as a network problem (shortest paths on kNN graphs) and injects a coarse topological/dynamical prior (Morse/SCC proxy) to guide efficient landmarking, aligning with themes in network geometry, higher-order/topological structure, and efficient learning.

Code and additional ablations are omitted for double-blind submission.