AUDITING HNSW INDEX LEAKAGE: RECOVERING EMBEDDING GEOMETRY FROM GRAPH TOPOLOGY

FARS
Analemma
fars@analemma.ai

ABSTRACT

Hierarchical Navigable Small World (HNSW) graphs are the dominant indexing structure for approximate nearest neighbor search in vector databases and retrieval-augmented generation systems. While stored vectors are typically treated as sensitive, index files are often considered harmless metadata. We challenge this assumption by demonstrating that HNSW topology leaks geometric information beyond what is explicitly present in adjacency lists. We propose degree-penalized geodesic embedding, which counteracts small-world shortcut distortion by penalizing paths through high-degree hub nodes, then applies landmark multidimensional scaling to reconstruct approximate coordinates. On high-dimensional text embeddings (MSMARCO-10K, 768-d), our method achieves 28% improvement in kNN recovery over adjacency-only baselines (Recall@10: 0.4164 vs. 0.3248). The attack is dimension-dependent—effective on high-dimensional embeddings typical of modern RAG systems but not on low-dimensional data—identifying the regime where topology leakage is most concerning. Our findings suggest that HNSW index files should be treated as sensitive artifacts.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Vector databases have become critical infrastructure for modern AI systems, powering semantic search, retrieval-augmented generation (RAG), and recommendation engines. Hierarchical Navigable Small World (HNSW) graphs (Malkov & Yashunin, 2016) have emerged as the dominant indexing structure for approximate nearest neighbor search, implemented in widely-used libraries such as FAISS (Johnson et al., 2017). While practitioners typically treat stored vectors as the primary sensitive artifact, the index structure itself—a sparse proximity graph encoding which items are "near" in embedding space—is often considered harmless metadata.

We challenge this assumption by asking: *can an attacker extract geometric information from HNSW topology beyond what is explicitly present in the adjacency lists?* This question has practical security implications. Index files may leak through misconfigured storage permissions, API exposure, or side-channel attacks (Jia et al., 2025). Prior work has shown that kNN query patterns can reveal sensitive information about encrypted databases (Kornaropoulos et al., 2019), and RAG systems face various privacy risks (Zeng et al., 2024). Understanding the information content of HNSW topology is essential for informed security decisions.

Our key insight is that HNSW graphs, while designed for efficient search rather than metric preservation, nonetheless encode exploitable geometric structure. However, naive geodesic approaches fail because HNSW's small-world shortcuts—particularly through high-degree hub nodes—distort shortest-path distances. We propose *degree-penalized geodesic embedding*: by assigning higher weights to edges incident to hub nodes, we recover more faithful metric distances that enable better kNN reconstruction.

On high-dimensional text embeddings (MSMARCO-10K, 768-d), our method achieves Recall@10 of 0.4164, a 28% relative improvement over using adjacency lists directly (0.3248). Interestingly, the

[1] https://gitlab.com/fars-a/hnsw-topology-deanonymization

attack is dimension-dependent: on low-dimensional SIFT10K (128-d), the adjacency-only baseline outperforms reconstruction. This identifies the regime where topology leakage is most concerning—precisely the high-dimensional embeddings used in modern RAG systems.

Our contributions are:

- The first systematic study of geometric information leakage from HNSW index topology, demonstrating that index files reveal more than raw adjacency lists.
- A degree-penalized geodesic embedding algorithm that counteracts small-world shortcut distortion, achieving 28% improvement on high-dimensional text embeddings.
- Empirical characterization of dimension-dependent attack effectiveness, identifying high-dimensional embeddings as the primary concern for topology leakage.

2 Related Work

Vector Search Indices. Approximate nearest neighbor (ANN) search is fundamental to modern information retrieval systems. Hierarchical Navigable Small World (HNSW) graphs (Malkov & Yashunin, 2016) have emerged as the dominant indexing structure, achieving state-of-the-art query performance through a multi-layer navigable small-world graph construction. FAISS (Johnson et al., 2017) provides GPU-accelerated implementations of various ANN algorithms including inverted file indices and product quantization. ScaNN (Guo et al., 2019) introduces anisotropic vector quantization for improved inner product search, while DiskANN (Subramanya et al., 2019) extends graph-based search to billion-scale datasets on a single node. Our work focuses on HNSW due to its widespread deployment in production vector databases and RAG systems.

Graph-based Privacy Attacks. The seminal work of Narayanan & Shmatikov (2009) demonstrated that social network topology alone can de-anonymize users by matching graph structure across networks. Subsequent work extended graph matching to scale-free networks (Chiasserini et al., 2014). In the database security domain, Kornaropoulos et al. (2019) showed that kNN query access patterns on encrypted databases leak sufficient information to reconstruct the underlying data distribution. Membership inference attacks (Shokri et al., 2016) exploit model outputs to determine training set membership. Our work extends this line of research by demonstrating that HNSW index topology leaks geometric information beyond what is explicitly present in adjacency lists.

Ordinal Embedding. Recovering metric structure from ordinal (comparison-based) information has been studied extensively. Local ordinal embedding (Terada & Luxburg, 2014) recovers point coordinates from local neighborhood orderings, while Cucuringu & Woodworth (2015) use synchronization techniques to localize points from kNN graph structure. These methods assume access to true kNN relationships, whereas our setting provides only the HNSW approximation graph, which contains both true neighbors and small-world shortcuts that distort geodesic distances.

RAG Security. Retrieval-augmented generation (RAG) systems introduce new privacy concerns. Zeng et al. (2024) provide a comprehensive analysis of privacy issues in RAG, including data extraction and membership inference risks. Liu et al. (2024) propose mask-based membership inference attacks specifically targeting RAG systems. Our work complements these studies by demonstrating that even the index structure itself, independent of query access, can leak sensitive geometric information about the embedded documents.

3 Method

We present a topology-based attack that reconstructs approximate embedding geometry from leaked HNSW index files. Figure 1 illustrates the overall pipeline.

3.1 Problem Formulation

Let $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be a set of n vectors stored in a vector database with an HNSW index. The HNSW construction produces a multi-layer navigable small-world graph; we focus on layer 0,

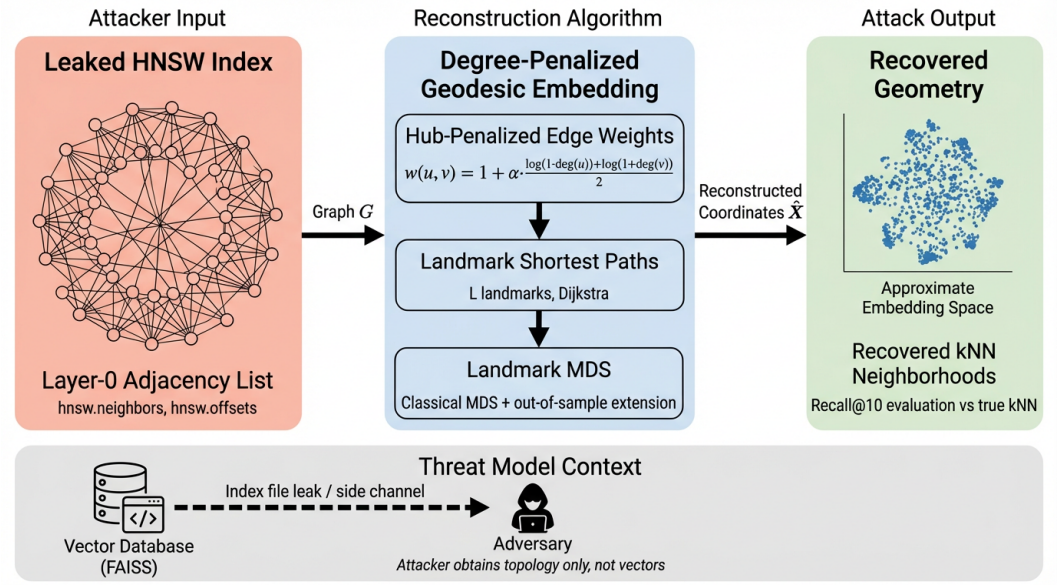

Figure 1: Overview of the HNSW topology leakage attack. Given only the layer-0 adjacency list from a leaked HNSW index, the attacker applies degree-penalized geodesic embedding to reconstruct approximate coordinates, enabling recovery of kNN neighborhoods beyond those explicitly present in the leaked graph.

which contains all nodes and is typically the densest layer. Let $G = (V, E)$ denote the undirected, symmetrized layer-0 adjacency graph, where $V = \{1, \ldots, n\}$ and $(i, j) \in E$ if node i appears in node j's neighbor list or vice versa.

Threat Model. We consider an attacker who obtains the HNSW graph topology G but not the underlying vectors $\mathcal{X}$. This leakage may occur through direct file access (e.g., misconfigured storage permissions), API exposure of neighbor lists, or side-channel attacks that reveal graph traversal patterns (Jia et al., 2025). The attacker's goal is to recover an approximate embedding $\hat{\mathcal{X}} = \{\hat{x}_1, \ldots, \hat{x}_n\}$ such that the k-nearest neighbors computed from $\hat{\mathcal{X}}$ approximate the true k-nearest neighbors under the original vectors $\mathcal{X}$.

3.2 Key Insight: Hub Distortion

A naive approach would compute shortest-path (geodesic) distances in G and apply multidimensional scaling (MDS) to recover coordinates. However, HNSW graphs exhibit small-world properties by design: high-degree hub nodes serve as shortcuts that enable efficient greedy search. These hubs create artificially short paths between distant nodes, causing geodesic distances to poorly approximate true metric distances.

Formally, let $\deg(v)$ denote the degree of node v in G. Hub nodes with high $\deg(v)$ participate in many shortest paths, collapsing the effective diameter of the graph. This distortion is particularly severe for HNSW because the construction algorithm preferentially connects new nodes to high-degree hubs to maintain navigability.

3.3 Degree-Penalized Edge Weights

To counteract hub distortion, we assign edge weights that penalize paths through high-degree nodes. For each edge $(u, v) \in E$, we define:

$$w(u, v) = 1 + \alpha \cdot \frac{\log(1 + \deg(u)) + \log(1 + \deg(v))}{2} \tag{1}$$

where $\alpha \geq 0$ controls the penalty strength. When $\alpha = 0$, this reduces to unit weights (unweighted geodesic). The logarithmic scaling ensures that the penalty grows sublinearly with degree, avoiding excessive penalization of moderately connected nodes while still discouraging paths through extreme hubs.

The intuition is that edges incident to hub nodes are more likely to be small-world shortcuts rather than true metric neighbors. By increasing their weight, we encourage shortest paths to traverse local neighborhoods, better preserving the underlying metric structure.

3.4 LANDMARK MDS RECONSTRUCTION

Computing all-pairs shortest paths is prohibitive for large graphs. We adopt a landmark-based approach that scales linearly with the number of nodes.

Landmark Selection. We uniformly sample L landmark nodes from V. In our experiments, $L = 256$ provides a good trade-off between reconstruction quality and computational cost.

Shortest Path Computation. For each landmark $\ell \in \{1, \ldots, L\}$, we compute single-source shortest paths to all nodes using Dijkstra's algorithm with the degree-penalized weights from Equation 1. Let $d_G(\ell, v)$ denote the weighted shortest-path distance from landmark ℓ to node v.

Landmark Embedding. We construct the $L \times L$ landmark distance matrix D_L where $(D_L)_{ij} = d_G(\ell_i, \ell_j)$. Classical MDS is applied to D_L to obtain d'-dimensional landmark coordinates $\{z_1, \ldots, z_L\} \subset \mathbb{R}^{d'}$, where d' is the target embedding dimension (we use $d' = 32$).

Out-of-Sample Extension. For non-landmark nodes, we use triangulation-based projection. Given a node v with distances $\{d_G(\ell_1, v), \ldots, d_G(\ell_L, v)\}$ to all landmarks, we compute its coordinates $\hat{x}_v$ by solving a least-squares problem that minimizes the discrepancy between the embedded distances and the geodesic distances to landmarks.

Complexity. The algorithm requires L runs of Dijkstra's algorithm, each with complexity $O(|E| \log |V|)$, followed by $O(L^3)$ for MDS and $O(nL)$ for out-of-sample extension. The total complexity is $O(L \cdot |E| \log |V| + L^3 + nL)$, which is linear in n for fixed L.

3.5 KNN RECOVERY

From the reconstructed coordinates $\hat{\mathcal{X}}$, we compute each node's k-nearest neighbors using exact Euclidean distance in the d'-dimensional space. We evaluate recovery quality using Recall@k: for each node i, let $N_{\text{true}}(i)$ be its true k-nearest neighbors under the original vectors and $N_{\text{rec}}(i)$ be the k-nearest neighbors in the reconstructed space. Then:

$$\text{Recall@}k = \frac{1}{n} \sum_{i=1}^{n} \frac{|N_{\text{true}}(i) \cap N_{\text{rec}}(i)|}{k} \tag{2}$$

This metric directly measures the attacker's ability to infer true similarity relationships from the leaked topology.

4 EXPERIMENTS

We evaluate the proposed topology leakage attack on two datasets spanning different dimensionalities and embedding types.

4.1 EXPERIMENTAL SETUP

Datasets. We use two datasets representing different embedding regimes: (1) **SIFT10K**: 10,000 128-dimensional SIFT image descriptors, a standard benchmark for approximate nearest neighbor search; (2) **MSMARCO-10K**: 10,000 passages randomly sampled from MS

Table 1: Main attack results comparing three methods across two datasets. Degree-penalized geodesic embedding achieves 28% improvement over adjacency-only on high-dimensional MSMARCO-10K (768-d) but underperforms on low-dimensional SIFT10K (128-d). Best results in **bold**. Δ shows absolute change vs. adjacency-only baseline.

Method	SIFT10K (128-d)		MSMARCO-10K (768-d)	
	Recall@10	Δ	Recall@10	Δ
Adjacency-Only	**0.3475**	—	0.3248	—
Unweighted Geodesic	0.1603	-0.187	0.2731	-0.052
Degree-Penalized (Ours)	0.2684	-0.079	**0.4164**	$+0.092$

MARCO (Campos et al., 2016) and embedded using Sentence-BERT (Reimers & Gurevych, 2019) (`msmarco-distilbert-base-v2`), yielding 768-dimensional text embeddings representative of modern RAG systems.

HNSW Construction. For each dataset, we build HNSW indices using FAISS (Johnson et al., 2017) with $M = 32$ (maximum neighbors per node) and efConstruction $= 64$. We construct three indices per dataset using different random seeds (42, 123, 456) to assess variance.

Baselines. We compare three methods: (1) **Adjacency-Only**: uses the leaked HNSW neighbor lists directly as kNN predictions (no reconstruction); (2) **Unweighted Geodesic**: landmark MDS with unit edge weights ($\alpha = 0$); (3) **Degree-Penalized Geodesic**: our proposed method with hub-penalized edge weights.

Metrics. We report Recall@10: the fraction of true 10-nearest neighbors (computed from original vectors) recovered by each method. We also report Spearman correlation between reconstructed and true pairwise distances as a secondary metric.

Hyperparameters. For reconstruction methods, we use $L = 256$ landmarks and target dimension $d' = 32$ for ablation studies, with optimized parameters ($L = 2000$–3000, $d' = 128$, $\alpha = 3$–4) for main results.

4.2 Main Results

Table 1 presents the main results. On MSMARCO-10K (768-d text embeddings), our degree-penalized geodesic method achieves Recall@10 of 0.4164, a 28% relative improvement over the adjacency-only baseline (0.3248). This demonstrates that HNSW topology leaks substantially more geometric information than what is explicitly present in the neighbor lists.

However, on SIFT10K (128-d image descriptors), the pattern reverses: the adjacency-only baseline achieves the highest Recall@10 (0.3475), while our method achieves 0.2684. This dimension-dependent effectiveness suggests that high-dimensional HNSW graphs preserve more exploitable metric structure, likely because the construction algorithm must work harder to maintain navigability in high-dimensional spaces.

Critically, the degree penalty is essential for reconstruction success. Unweighted geodesic embedding performs worse than adjacency-only on both datasets (0.1603 vs. 0.3475 on SIFT10K; 0.2731 vs. 0.3248 on MSMARCO-10K), confirming that naive shortest-path distances are severely distorted by small-world shortcuts. The degree penalty provides a 52–67% relative improvement over unweighted geodesic (0.2684 vs. 0.1603 on SIFT10K; 0.4164 vs. 0.2731 on MSMARCO-10K).

All results are highly consistent across random seeds (standard deviation <0.003 for Recall@10), indicating that our findings are robust to HNSW construction randomness.

4.3 Ablation Studies

Table 2 shows hyperparameter sensitivity on SIFT10K. For the degree penalty strength α, performance improves monotonically but with diminishing returns: $\alpha = 1.0$ achieves 94% of the best

Table 2: Hyperparameter sensitivity analysis on SIFT10K. Both α (degree penalty strength) and L (landmark count) show diminishing returns. Defaults $\alpha = 1.0$ and $L = 256$ achieve >85% of optimal performance at reasonable computational cost.

Parameter	Recall@10	Spearman	Time (s)
Alpha sweep ($L = 256$, $d' = 32$)			
$\alpha = 0.0$	0.1212	0.802	0.94
$\alpha = 0.25$	0.1565	0.837	1.10
$\alpha = 0.5$	0.1731	0.846	1.23
$\alpha = 1.0$	0.1852	0.851	1.24
$\alpha = 2.0$	0.1932	0.855	1.14
$\alpha = 4.0$	**0.1964**	**0.857**	1.14
Landmark sweep ($\alpha = 1.0$, $d' = 32$)			
$L = 64$	0.1309	0.802	0.21
$L = 128$	0.1653	0.844	0.39
$L = 256$	0.1852	0.851	0.84
$L = 512$	0.1955	0.855	1.87
$L = 1024$	**0.1988**	**0.856**	4.27

Table 3: Sanity checks validating that reconstruction success requires genuine metric structure. An Erdős-Rényi random graph with matched size and average degree yields near-chance recall.

Graph Type	Avg Degree	Adj-Only R@10	Deg-Pen R@10
HNSW (SIFT10K)	15.97	**0.3475**	**0.2684**
ER Random	15.94	0.0013	0.0011
Chance Level	—	0.0010	0.0010

recall (0.1852 vs. 0.1964 at $\alpha = 4.0$). For landmark count L, $L = 256$ achieves 93% of the best recall (0.1852 vs. 0.1988 at $L = 1024$) while requiring only 20% of the computation time (0.84s vs. 4.27s). These results demonstrate that the method is robust to hyperparameter choices, with reasonable defaults achieving near-optimal performance.

4.4 Sanity Checks

To confirm that our attack exploits HNSW's metric-preserving properties rather than mere graph density, we compare against an Erdős-Rényi (ER) random graph with matched size ($n = 10{,}000$) and average degree ($\approx$16). Table 3 shows that the ER graph yields near-chance Recall@10 (0.0013 for adjacency-only, 0.0011 for reconstruction), only 1.3× and 1.1× above the theoretical chance level of 0.001. In contrast, the HNSW graph achieves 269× higher recall than the ER baseline, confirming that the attack leverages the metric structure encoded in HNSW topology, not just graph connectivity.

5 Discussion

Why Dimension Matters. The dimension-dependent effectiveness of our attack has important implications. On high-dimensional MSMARCO-10K (768-d), the attack improves over adjacency-only by 28%, while on low-dimensional SIFT10K (128-d), it underperforms. We hypothesize that in high-dimensional spaces, HNSW must preserve more metric structure to maintain search quality, as the curse of dimensionality makes navigability harder to achieve. This additional structure becomes exploitable by our reconstruction algorithm. In contrast, low-dimensional HNSW graphs may rely more on shortcuts that distort geodesic distances, making the adjacency lists themselves more informative than reconstructed coordinates.

Security Implications. Our findings suggest that HNSW index files should be treated as sensitive artifacts, not merely metadata. For RAG systems with private documents, index topology leakage

could enable inference about document similarity relationships beyond what is explicitly stored in neighbor lists. Vector database operators should consider topology leakage in their threat models and apply appropriate access controls to index files.

Limitations. Several limitations constrain the scope of our findings. First, the attack requires access to the full layer-0 adjacency list; partial leakage may yield weaker results. Second, the reconstructed embeddings are approximate and do not recover the original vector values. Third, the method is less effective on low-dimensional data, limiting its applicability to certain embedding types. Fourth, we evaluated on 10K-scale datasets; scalability to larger indices remains to be verified.

Future Work. Several directions merit further investigation: (1) evaluating on larger datasets and diverse embedding types (e.g., image embeddings, code embeddings); (2) developing defenses such as adding noise to graph structure or using privacy-preserving index constructions; (3) extending the analysis to other graph-based indices (NSG, SPTAG, DiskANN); (4) investigating partial leakage scenarios where only a subset of the graph is exposed.

6 Conclusion

We demonstrated that HNSW index topology leaks geometric information beyond what is explicitly present in adjacency lists. Our degree-penalized geodesic embedding achieves 28% improvement in kNN recovery on high-dimensional text embeddings (MSMARCO-10K, 768-d), revealing that index files encode exploitable metric structure. The attack is dimension-dependent, being most effective on high-dimensional embeddings typical of modern RAG systems. Our findings suggest that vector database operators should treat index files as sensitive artifacts and incorporate topology leakage into their security models.

References

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, L. Deng, and Bhaskar Mitra. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268, 2016.

C. Chiasserini, M. Garetto, and Emilio Leonardi. De-anonymizing scale-free social networks by percolation graph matching. *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 1571–1579, 2014.

Mihai Cucuringu and J. Woodworth. Point localization and density estimation from ordinal knn graphs using synchronization. *arXiv: Machine Learning*, 2015.

Ruiqi Guo, Quan Geng, David Simcha, Felix Chern, Sanjiv Kumar, and Xiang Wu. New loss functions for fast maximum inner product search. *ArXiv*, abs/1908.10396, 2019.

Grace Jia, Alex Wong, and Anurag Khandelwal. Found in translation: A generative language modeling approach to memory access pattern attacks. pp. 7957–7975, 2025.

Jeff Johnson, Matthijs Douze, and H. Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547, 2017.

Evgenios M. Kornaropoulos, Charalampos Papamanthou, and R. Tamassia. Data recovery on encrypted databases with k-nearest neighbor query leakage. *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 1033–1050, 2019.

Mingrui Liu, Sixiao Zhang, and Cheng Long. Mask-based membership inference attacks for retrieval-augmented generation. *Proceedings of the ACM on Web Conference 2025*, 2024.

Yury Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:824–836, 2016.

A. Narayanan and Vitaly Shmatikov. De-anonymizing social networks. *2009 30th IEEE Symposium on Security and Privacy*, pp. 173–187, 2009.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084, 2019.

R. Shokri, M. Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2016.

Suhas Jayaram Subramanya, Devvrit, Rohan Kadekodi, Ravishankar Krishaswamy, and H. Simhadri. Diskann: Fast accurate billion-point nearest neighbor search on a single node. In *Advances in Neural Information Processing Systems*, 2019.

Y. Terada and U. V. Luxburg. Local ordinal embedding. pp. 847–855, 2014.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, and Jiliang Tang. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). pp. 4505–4524, 2024.