SENSE: SENsing Similarity SEeing Structure

Dimensionality reduction (DR) is central to modern ML, allowing high-dimensional data to be mapped into interpretable low-dimensional spaces. Among many DR techniques, a prominent subclass is *neighbor embedding (NE)* methods, which learn embeddings that preserve pairwise similarities. Classical NE algorithms such as t-SNE and UMAP [1] compute $S_{ij}^{d_h} = f(\|x_i - x_j\|),\ x_i \in \mathbb{R}^{d_h}$, and optimize embeddings $Y = \{y_i\} \subset \mathbb{R}^{d_\ell}$ so that low-dimensional similarities $S_{ij}^{d_\ell} = g(\|y_i - y_j\|)$ match $S_{ij}^{d_h}$ by minimizing a divergence mechanism $\mathcal{L}(Y) = \sum_{i,j} D(S_{ij}^{d_h}, S_{ij}^{d_\ell})$. All similarity-based DR methods assume *centralized access* to pairwise distances, which fails in decentralized settings. Different methods degrade differently. Force-layout methods (t-SNE/UMAP) are also fragile since attraction to neighbors and repulsion from others require global negatives. When negatives are sampled only locally, repulsion is biased, yielding cluster drift and distorted layouts [2, 3]. Several approaches have been proposed to address this gap, but they fall short on scalability, privacy, or deployment realism. SMAP [4] secures t-SNE with multi-party computation but requires ~32-50 hrs for 4k points and lacks UMAP support. FedNE [2] uses surrogate distillation with inter-client exchange, is not scalable and is inversion-prone. FedTSNE [3] aligns anchors via MMD but is fragile to adversaries, restricted to multisite, and requires iterative rounds. To overcome these challenges, we propose ***SENSE***, a geometry-aware, privacy-preserving framework for global NE without raw data exchange. SENSE reconstructs global structure using local distance measurements and structured matrix completion, enabling embeddings that preserve both local and global geometry in Euclidean and hyperbolic spaces. This eliminates the need for raw feature sharing, iterative communication, or cryptographic protocols. Privacy is built into the design: when the number of anchors satisfies $K < d_h$, the inverse mapping from anchor distances to original features is provably non-unique, preventing exact recovery. By combining structured matrix completion with anchor-based coordination, SENSE provides an efficient, privacy-preserving alternative to existing NE methods in decentralized environments. It further integrates contrastive learning by deriving cross-client positive and negative pairs from estimated similarities, effectively generalizing negative sampling under structural constraints. Overall, SENSE introduces the following advantages: **1) Communication-efficient and geometry-aware:** Requires a single server-client interaction and supports both Euclidean and hyperbolic spaces for modeling flat and hierarchical data. **2) Deployment flexibility (described in Fig. 1):** Operates under two regimes SENSE-Pointwise for single-point clients (e.g., edge/mobile) and SENSE-Multisite for multi-sample clients (e.g., hospitals, banks). **3) Provable reliability:** Offers theoretical guarantees on privacy preservation, ensuring embedding fidelity, validated across diverse modalities and geometries. These properties make SENSE suitable for privacy-sensitive, structurally diverse domains. Hospitals can jointly visualize patient data without violating HIPAA/GDPR, banks can detect fraud patterns without sharing transactions, and mobile/IoT clients with a single sample can still contribute to global embeddings. Genomic labs can embed single-cell transcriptomes into a shared hyperbolic space that preserves cellular hierarchy and privacy. Crucially, SENSE also supports evolving data scenarios and dynamic client participation: new clients or data points can be integrated by estimating only their partial distances to a subset of existing entities, avoiding full re-computation and preserving global coherence with minimal overhead. This makes SENSE not only privacy-preserving and geometry-aware but also inherently scalable to dynamic and federated settings.

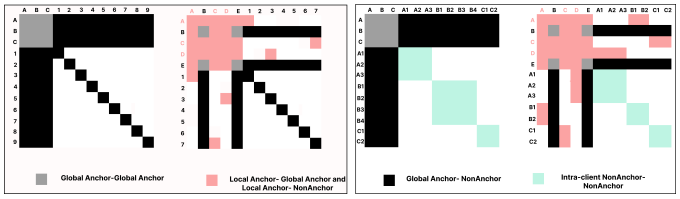

Figure 1: Observed entries in the global distance matrix D under four SENSE configurations: (1) *Pointwise-Full*, (2) *Pointwise-Partial*, (3) *Multisite-Full*, and (4) *Multisite-Partial*. These differ in the visibility of Anchor–NonAnchor (A–NA) and NA–NA blocks, governed by client-level data locality and anchor access. *Multisite* settings permit intra-client NA–NA observations (e.g., A1, A2, ..., C2), while *Pointwise* settings restrict each client to a single NA (e.g., 1, 2, ..., 9). *Full* modes provide all NAs with access to the global anchor set (e.g., A–E), yielding complete A–NA blocks; *Partial* modes expose disjoint anchor subsets per client, resulting in sparse and structured observations.

References

[1] S. Damrich, J. N. Böhm, F. A. Hamprecht, and D. Kobak. From t-sne to umap with contrastive learning, 2023. URL https://arxiv.org/abs/2206.01816.

[2] Z. Li, X. Wang, H.-Y. Chen, H.-W. Shen, and W.-L. Chao. Fedne: Surrogate-assisted federated neighbor embedding for dimensionality reduction, 2024. URL https://arxiv.org/abs/2409.11509.

[3] D. Qiao, X. Ma, and J. Fan. Federated t-sne and umap for distributed data visualization, 2024. URL https://arxiv.org/abs/2412.13495.

[4] J. Xia, T. Chen, L. Zhang, W. Chen, Y. Chen, X. Zhang, C. Xie, and T. Schreck. Smap: A joint dimensionality reduction scheme for secure multi-party visualization. In *2020 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 107–118, 2020. doi: 10.1109/VAST50239.2020.00015.