Computationally Tractable Riemannian Manifolds for Graph Embeddings

Calin Cruceru 1  Gary Bécigneul 1  Octavian-Eugen Ganea 1 2

Abstract

Representing graphs as sets of node embeddings in certain curved Riemannian manifolds has recently gained momentum in machine learning due to their desirable geometric inductive biases, e.g., hierarchical structures benefit from hyperbolic geometry. However, going beyond embedding spaces of constant sectional curvature, while potentially more representationally powerful, proves to be challenging as one can easily lose the appeal of computationally tractable tools such as geodesic distances or Riemannian gradients. Here, we explore computationally efficient matrix manifolds, showcasing how to learn and optimize graph embeddings in these Riemannian spaces. Empirically, we demonstrate consistent improvements over Euclidean geometry while often outperforming hyperbolic and elliptical embeddings based on various metrics that capture different graph properties. Our results serve as new evidence for the benefits of non-Euclidean embeddings in machine learning pipelines.

1. Introduction

Before representation learning started gravitating around deep representations (Bengio et al., 2009) in the last decade, a line of research that sparked interest in the early 2000s was based on the so called manifold hypothesis (Bengio et al., 2013). According to it, real-world data given in their raw format (e.g., pixels of images) lie on a low-dimensional manifold embedded in the input space. At that time, most manifold learning algorithms were based on locally linear approximations to points on the sought manifold – such as LLE (Roweis & Saul, 2000) and Isomap (Tenenbaum et al., 2000) – and/or spectral methods – such as MDS (Hofmann & Buhmann, 1995) and graph Laplacian eigenmaps (Belkin & Niyogi, 2002).

With no additional qualifiers, graph embedding is a vaguely specified intermediary step used as part of systems solving a wide range of graph analytics problems (Nie et al., 2017; Wang et al., 2017; Wei et al., 2017; Zhou et al., 2017). What they all have in common is the representation of certain

1Not expressible as Cartesian products of other manifolds, be they model spaces, as considered in (Gu et al., 2018), or yet others.
parts of a graph as points in a continuous space. The desired mathematical properties depend on the problem setting. Classically, the Euclidean space has been ubiquitous due to its interpretability and structure: inner product, metric, and, very conveniently for compositional models, linearity.

As a particular instance of that general task, here we embed nodes of graphs with structural information only (i.e., undirected and without node or edge labels), as the one shown in Figure 1, in novel curved spaces, by leveraging the closed-form expressions of the corresponding Riemannian distance between embedding points; the resulting geodesic distances enter a differentiable objective function which “compares” them to the ground-truth metric given through the node-to-node graph distances. We focus on the representation capabilities of the considered matrix manifolds relative to the previously studied spaces by monitoring graph reconstruction metrics. We note that preserving graph structure is essential to downstream tasks such as link prediction (Trouillon et al., 2016) or node classification (Wang et al., 2017).

Our main contributions are (i) the introduction of two families of matrix manifolds for graph embedding purposes: the non-positively curved spaces of symmetric positive definite (SPD) matrices, and the compact, non-negatively curved Grassmann manifolds; (ii) revisiting Stochastic Neighborhood Embedding (SNE) (Hinton & Roweis, 2003) in the context of Riemannian embeddings and showing that it unifies, on the one hand, the loss functions based on the reconstruction likehood of local graph neighborhoods and, on the other hand, the global, all-pairs stress functions used for global metric recovery; (iii) a generalization of the usual ranking-based metric to quantify reconstruction fidelity beyond immediate neighbors; (iv) a comprehensive experimental comparison of the introduced manifolds against the baselines in terms of their graph reconstruction capabilities, focusing on the impact of curvature.

2. Preliminaries & Background

Notation Let \( G = (X, E, w) \) be an undirected graph, with \( X \) the set of nodes, \( E \) the set of edges, and \( w : E \to \mathbb{R}_+ \) the edge-weighting function. Let \( m = |X| \). We denote by \( d_G(x_i, x_j) \) the shortest path distance between nodes \( x_i, x_j \in X \), induced by \( w \). The node embeddings are \( Y = \{y_i\}_{i \in [m]} \subseteq \mathcal{M} \) and the geodesic distance function is \( d_{\mathcal{M}}(y_i, y_j) \), with \( \mathcal{M} \) – the embedding space – a Riemannian manifold. \( \mathcal{N}(x_i) \) denotes the set of neighbors of node \( x_i \).

2.1. Riemannian Geometry

A brief but comprehensive account of the fundamental concepts from Riemannian geometry is included in Appendix A.

Informally, an \( n \)-dimensional manifold \( \mathcal{M} \) is a space that locally resembles \( \mathbb{R}^n \). Each point \( x \in \mathcal{M} \) has attached a tangent space \( T_x \mathcal{M} \) – a vector space that can be thought of as a first-order local approximation of \( \mathcal{M} \) around \( x \). The Riemannian metric \( \langle \cdot, \cdot \rangle_x \) is a collection of inner products on these tangent spaces that vary smoothly with \( x \). It makes possible measuring geodesic distances, angles, and curvatures. The different notions of curvature quantify the ways in which a surface is locally curved around a point. The exponential map is a function \( \exp_x : T_x \mathcal{M} \to \mathcal{M} \) that can be seen as folding or projecting the tangent space onto the manifold. Its inverse is called the logarithm map, \( \log_x(\cdot) \).

2.2. Learning Framework

The embeddings are learned in the framework used in prior work (Nickel & Kiela, 2017; Gu et al., 2018) in which a loss function \( \mathcal{L} \) depending on the embedding points solely via the (Riemannian) distances between them is minimized using stochastic Riemannian optimization (Bonnabel, 2013; Becigneul & Ganea, 2019). In this respect, the following general property is useful (Lee, 2006): for any point \( x \) on a Riemannian manifold \( \mathcal{M} \) and any \( y \) in a neighborhood of \( x \), we have \( \nabla^R_x d^2(x, y) = -2 \log_x(y) \).

Hence, as long as \( \mathcal{L} \) is differentiable with respect to the (squared) distances, it will also be differentiable with respect to the embedding points. The specifics of \( \mathcal{L} \) are deferred to Section 4.

2.3. Model Spaces & Cartesian Products

The model spaces of Riemannian geometry are manifolds with constant sectional curvature \( \kappa \): (i) Euclidean space \( \kappa = 0 \), (ii) hyperbolic space \( \kappa < 0 \), and (iii) elliptical space \( \kappa > 0 \). We summarize the Riemannian geometric properties of the last two in Appendix B. They are used as baselines in our experiments (Section 5).

We also recall that given a set of manifolds \( \{\mathcal{M}_i\}_{i=1}^k \), the product manifold \( \mathcal{M} = \times_{i=1}^k \mathcal{M}_i \) has non-constant sectional curvature and can be used for graph embedding purposes as long as each factor has efficient closed-form formulas for the quantities of interest (Gu et al., 2018).

2.4. Measuring Curvature

Curvature properties are central to our work since they set apart the matrix manifolds discussed in Section 3. Here, we review several analogous concepts defined for graphs. Graphs are different mathematical abstractions but yet similar in many aspects (through, e.g., Laplace operators and heat kernels). Furthermore, we introduce a simple method for quantifying space curvature around a set of embeddings.

\( \nabla^R_x \) denotes the Riemannian gradient at \( x \). See Appendix A.

---

\(^2\)We use \( i \in [m] \) as a short-hand for \( i \in \{1, 2, \ldots, m\} \).

\(^3\)See Appendix A.
Geometric Properties of Graphs. We use the following geometry-inspired graph properties throughout this work (details in Appendix C):

- **Ollivier-Ricci curvature.** Introduced by (Ollivier, 2009) for general metric spaces and specialized in (Lin et al., 2011) to graphs, it is defined for pairs of neighbors \((u, v)\) and is inspired by the continuous Ricci curvature. An analogue to the Riemannian scalar curvature at \(u\) is obtained by averaging its value for all neighbors \(v\). Intuitively, a negative value means the edge/node is part of the backbone, i.e., the graph would get disconnected if it were removed. See Figure 1.

- **\(\delta\)-hyperbolicity** (Gromov, 1987). It quantifies the hyperbolicity of a given metric space: the smaller \(\delta\), the more hyperbolic-like (or negatively-curved) the space. It is based on the following insight: geodesic triangles are “slim” in negatively curved spaces and “thick” in positively curved ones. See Figure 2.

**Sum-of-Angles in Geodesic Triangles**. Note that even in hyperbolic geometry, which has constant negative curvature, placing points close to each other leads to an approximately flat embedding. With that in mind, given three points \(x, y, z \in \mathcal{M}\), a simple quantity that characterizes the actual space curvature between them is the sum of the angles in the geodesic triangle that they form (see Appendix A)

\[
k_{\delta}(x, y, z) = \theta_{x,y} + \theta_{x,z} + \theta_{y,z} - \pi.
\]

\[
\theta_{x_1,x_2} = \cos^{-1}\left(\frac{\langle u_1, u_2 \rangle_{x_3}}{\|u_1\|_{x_3}\|u_2\|_{x_3}}\right), \quad \text{with } u_i = \log_{x_3}(x_i).
\]

In practice, we look at empirical distributions of \(k_{\delta}\) given by triples sampled uniformly from an embedding set \(\{y_i\}_{i=1}^{k}\). Moreover, for presentation purposes, because the sum is between \([0, \pi]\) for hyperbolic triangles and between \([\pi, 3\pi]\) for spherical ones, we translate \(k_{\delta}\) by \(-\pi\) and divide by \(2\pi\). This gives us the ranges \([-0.5, 0]\) and \([0, 1]\), respectively.

3. Matrix Manifolds

We now review the two proposed families of matrix manifolds. We have chosen them such that they cover negative and positive curvature ranges, respectively. Also, essential for graph embedding purposes, they lend themselves to computationally tractable Riemannian optimization. Their properties are summarized in Table 1. In what follows, we insist on those aspects that are relevant for graph embedding.

3.1. Non-positive Curvature: SPD Manifold

**Definition**. The space of \(n \times n\) real symmetric positive-definite matrices,

\[
\mathcal{S}^{++}(n) := \{ A \in \mathcal{S}(n) : \langle x, Ax \rangle > 0 \text{ for all } x \neq 0 \},
\]

is an \(\frac{n(n+1)}{2}\) dimensional differentiable manifold – an embedded submanifold of \(\mathcal{S}(n)\), the space of \(n \times n\) symmetric matrices. Its tangent space can be identified with \(\mathcal{S}(n)\).

**Riemannian Structure**. The most common Riemannian metric endowed to \(\mathcal{S}^{++}(n)\) is \(\langle P,Q \rangle_{A} = \text{Tr} A^{-1} PA^{-1}Q\). Also called the *canonical* metric, it is motivated as being invariant to congruence transformations \(\Gamma_{X}(A) = X^\top AX\), with \(X\) an \(n \times n\) invertible matrix (Pennec et al., 2006). Several geodesic paths are drawn in Figure 3.

**Riemannian Distance**. The induced distance function is equivalent to\(^4\)

\[
d(A,B) = \sqrt{\sum_{i=1}^{n} \log^2 \left( \lambda_i(A^{-1}B) \right)}.
\]

Notice that singular positive semi-definite matrices, which lie on the boundary \(\partial \mathcal{S}^{++}(n)\), are points at infinity. An interpretation of the eigenvalues \(\lambda_i(A^{-1}B)\) can be obtained by recalling that for any \(A,B \in \mathcal{S}^{++}(n)\), there exist an invertible matrix \(X\) and a diagonal matrix \(D\) such that \(X^{-1}AX = Id_n\) and \(X^{-1}BX = D\). Thus, the distance can be seen as measuring how well \(A\) and \(B\) can be simultaneously reduced to the identity matrix (Chossat & Faugeras, 2009). See Appendix D for proofs and details.

**Properties**. The canonical SPD manifold has non-positive sectional curvature everywhere (Bhatia, 2009). It is also a high-rank symmetric space (Lang, 2012). The high-rank property tells us that there are at least planes of the tangent space on which the sectional curvature vanishes. Contrast it with the hyperbolic space which is also a symmetric space but where the only (intrinsic) flats are the geodesics.

\(^4\)We use \(\lambda_i(X)\) to denote the \(i\)th eigenvalue of \(X\) when the order is not important.
Computationally Tractable Riemannian Manifolds for Graph Embeddings

Table 1. Summary of Differential and Riemannian Geometry Tools. Notation: $A, B$ – manifold points; $P, Q$ – tangent space points; $P'$ – ambient space point; $\nabla_A / \nabla_Q$ – Euclidean / Riemannian gradient; $\exp(A) / \log(A)$ – matrix exponential / logarithm. References: SPD (Bhatia, 2009; Bridson & Haefliger, 2013; Jeuris, 2015); Grassmann (Edelman et al., 1998; Zhang et al., 2018).

| Property          | Expr.              | SPD $S^{++}(n)$ | Grassmann $Gr(k, n)$ |
|-------------------|--------------------|-----------------|----------------------|
| Dimension         | $n(n+1)/2$         | $(3)$           | $(4)$                |
| Tangent space     | $T_A M \{ A \in \mathbb{R}^{n \times n} : A = A^\top \}$ | $(5)$           | $(6)$                |
| $T_A M$ projection| $\pi_A(P') \{ P' \in \mathbb{R}^{n \times n} : P = 0 \}$ | $(7)$           | $(8)$                |
| Riem. metric      | $\langle P, Q \rangle_A \text{Tr} A^{-1} PA^{-1} Q$ | $(9)$           | $(10)$               |
| Riem. gradient    | $\nabla_A A \pi_A(\nabla_A^2) A$ | $(11)$          | $(12)$               |
| Geodesic          | $\gamma_{A,P}(t) \exp(t A^{-1} P)$ | $(13)$          | $(14)$               |
| Retraction        | $R_A(P) \text{A} + P + \frac{1}{2} PA^{-1} P$ | $(15)$          | $(16)$               |
| Log-map           | $\log_A(B) \text{log}(A^{-1} B)$ | $(17)$          | $(18)$               |
| Riem. distance    | $d(A, B) \| \log(A^{-1} B) \|_F$ | $(19)$          | $(20)$               |

Moreover, only one degree of freedom can be factored out of the manifold $S^{++}(n)$: it is isometric to $S^{++}(n) \times \mathbb{R}$, with $S^{++}(n) := \{ A \in S^{++}(n) : \det(A) = 1 \}$, an irreducible manifold (Dolcetti & Pertici, 2018a). Therefore, $S^{++}$ achieves a mix of flat and negatively-curved areas that cannot be obtained via other Riemannian Cartesian products.

Alternative Metrics There are several other metrics that one can consider, which can be most efficient log-Euclidean one (see, e.g., Arsigny et al., 2006). However, the induced curvature is zero, so it presents no advantage, unlike Euclidean geometry. Another, more interesting one is the Bures-Wasserstein metric from quantum information theory (Bhatia et al., 2019), which induces a non-negative curvature on $S^{++}(n)$. It is leveraged in (Muzellec & Cuturi, 2018) to embed nodes as elliptical distributions. Finally, a popular alternative to the (square) canonical distance, which we adopt in our experiments, is the asymmetric Stein divergence,

$$S(A, B) := \log \det \left( \frac{A + B}{2} \right) - \frac{1}{2} \log \det(AB).$$

It has been thoroughly studied in (Sra, 2012; Sra & Hosseini, 2015) who prove that $\sqrt{S}$ is indeed a metric, and that $S(A, B)$ shares many properties of the Riemannian distance function (19), such as congruence and inversions invariances, as well as geodesic convexity in each argument. It is particularly appealing for backpropagation-based training due to its computationally efficient gradients (see below).

Computational Aspects We compute gradients via automatic differentiation (Paszke et al., 2017). Notice that if $A = UDUT^\top$ is the eigendecomposition of a symmetric matrix with distinct eigenvalues and $L$ is some loss function that depends on $A$ only via $D$, then (Giles, 2008)

$$\frac{\partial L}{\partial D} = U \frac{\partial L}{\partial D} U^\top.$$

Computing geodesic distances requires the eigenvalues of $A^{-1} B$, though, which may not be symmetric. We overcome that by using the matrix $A^{-1/2} B A^{-1/2}$ instead which is SPD and has the same spectrum. Moreover, for the $2 \times 2$ and $3 \times 3$ cases, we use closed-form eigenvalue formulas to speed up our implementation.\footnote{This could be done in theory for $n \leq 4$ – a consequence of the Abel-Ruffini theorem from algebra. However, for $n = 4$ the formulas are outperformed by numerical eigenvalue algorithms.}

3.2. Non-negative Curvature: Grassmann Manifold

Definition The orthogonal group $O(n)$ is the set of $n \times n$ real orthogonal matrices. It is a special case of the compact Stiefel manifold $V(k, n) := \{ A \in \mathbb{R}^{n \times k} : A^\top A = \text{Id}_k \}$, i.e., the set of $n \times k$ “tall-skinny” matrices with orthonormal columns, for $k \leq n$. The Grassmannian is defined as the space of $k$-dimensional linear subspaces of $\mathbb{R}^n$. It is related to the Stiefel manifold in that every orthonormal $k$-frame in
We refer to (Edelman et al., 1998) for details on how it arises. The closed form formula, shown in Figure 4. Riemannian Distance

The canonical Riemannian metric of \( Gr(k, n) \) with the quotient space \( V(k, n)/O(k) \). In other words, an \( n \times k \) orthonormal matrix \( A \in V(k, n) \) represents the equivalence class \( [A] = \{AQk : Qk \in O(k)\} \cong \text{span}(A) \), which is a single point on \( Gr(k, n) \).

Riemannian Structure The canonical Riemannian metric of \( Gr(k, n) \) is simply the Frobenius inner product (10). We refer to (Edelman et al., 1998) for details on how it arises from its quotient geometry. As before, we include examples of geodesic paths on \( Gr(2, 3) \) in Figure 4.

Riemannian Distance The closed form formula, shown in (20), depends on the set \( \{\theta_i\}_i^k \) of so-called principal angles between two subspaces, defined recursively as

\[
\begin{align*}
\cos \theta_i &= \max_{a_i \in \text{span}(A), b_i \in \text{span}(B)} a_i^\top b_i \\
&= \max \{a_i^\top b_i : a_i \in \text{span}(A), b_i \in \text{span}(B)\} \\
a_i^\top a_i &= 1, \quad b_i^\top b_i = 1, \\
a_i^\top b_i &= 0, \quad b_i^\top a_i = 0, \quad \forall j < i.
\end{align*}
\]

They can be interpreted as the minimal angles between all possible bases of the two subspaces.

Alternative Metrics Several alternatives to the arc-length metric (20) have been proposed, all expressible in terms of the principal angles – see (Edelman et al., 1998, Section 4.3) for an overview. A popular one is the so-called projection norm, \( d_p(A, B) = \|A \text{vec}(A) - B \text{vec}(B)\|_F \). It corresponds to embedding \( Gr(k, n) \) in \( \mathbb{R}^{n \times n} \) but then using the ambient space metric. It is analogous to taking Euclidean distances between points on a sphere, thus ignoring its geometry.

Computational Aspects Computing a geodesic distance requires the SVD decomposition of an \( k \times k \) matrix, which can be significantly smaller than the manifold dimension \( k(n-k) \). For \( k = 2 \), we use closed-form solutions for singular values (see Appendix D). Otherwise, we employ standard numerical algorithms. For the gradients, a result analogous to eq. (22) makes automatic differentiation straight-forward.

Properties The Grassmann manifold \( Gr(k, n) \) is a compact, non-negatively curved manifold. As shown by (Wong, 1968), its sectional curvatures at \( A \in Gr(k, n) \) satisfy

\[
\begin{align*}
K_A(P, Q) &= 1 & k &= 1, \ n > 2 \\
0 &\leq K_A(P, Q) & k &= 1, \ n > k,
\end{align*}
\]

for all \( P, Q \in T_A Gr(k, n) \). Contrast the above with the constant positive curvature of the sphere which can be made arbitrarily large by making its radius \( R \to 0 \).

4. Decoupling Learning and Evaluation

Recall that our goal is to preserve the graph structure given through its node-to-node shortest paths by optimizing an objective which encourages similar (relative) geodesic distances between node embeddings. Prior work broadly uses local or global loss functions that focus on either close neighborhood information or all-pairs interactions, respectively. The methods that fall under the former emphasize correct placement of immediate neighbors, such as the one used in (Nickel & Kiela, 2017) for unweighted graphs

\[
\mathcal{L}_{\text{neigh}} = - \sum_{(i,j) \in E} \log \frac{\exp \left( -d_M(y_i, y_j) \right)}{\sum_{k \in \mathcal{N}(i)} \exp \left( -d_M(y_i, y_k) \right)}.
\]

Those that fall under the latter, on the other hand, compare distances directly via loss functions inspired by generalized MDS (Bronstein et al., 2006), e.g.,

\[
\begin{align*}
\mathcal{L}_{\text{stress}}(Y) &= \sum_{i<j} \left( d_G(x_i, x_j) - d_M(y_i, y_j) \right)^2, \\
\mathcal{L}_{\text{distortion}}(Y) &= \sum_{i<j} \left| \frac{d_G^2(y_i, y_j)}{d_M^2(x_i, x_j)} - 1 \right|.
\end{align*}
\]

Note that (26) focuses mostly on distant nodes, while misrepresenting close ones yields a large loss according to (27) – one of several objective functions used in (Gu et al., 2018). The two types of objectives yield embeddings with different properties. It is thus not surprising that each one of them has been coupled in prior work with a preferred metric quantifying reconstruction fidelity. The likelihood-based one is evaluated via the rank-based mean average precision

\[
mAP = \frac{1}{m} \sum_i \frac{1}{N(i)} \sum_j \frac{\mathcal{N}(i) \cap B(j; i)}{B(j; i)},
\]

with \( B(j; i) = \{y_k \in \mathcal{M} : d_M(y_i, y_k) \leq d_M(y_i, y_j)\} \), while the global, stress-like ones yield best scores when measured by the average distortion of the reference metric

\[
D_{\text{avg}} = \frac{2}{m(m-1)} \sum_{i<j} \frac{|d_M(y_i, y_j) - d_G(x_i, x_j)|}{d_G(x_i, x_j)}.
\]
Our Proposal To decouple learning and evaluation, we propose to optimize another loss function that allows moving in a continuous way on the representation scale ranging from local neighborhoods patching, as encouraged by \(25\), to the global topology matching, as made desirable by \(26\) and \(27\). In the same spirit, we propose a more fine-grained ranking metric that makes the trade-off clearer.

4.1. Riemannian Stochastic Neighbor Embedding

We advocate training embeddings via a version of the celebrated Stochastic Neighbor Embedding (SNE) (Hinton & Roweis, 2003) adapted to the Riemannian setting. As shown next, this is almost trivial while the benefits are significant.

SNE works by attaching to each node a distribution defined over all other nodes and based on the distance to them. This is done for both the input graph distances, yielding the ground truth distribution, and for the embedding distances, yielding the model distribution. That is, with \(j \neq i\), we have

\[
 p_{ij} := p(x_j | x_i) = \frac{1}{Z_{p_i}} \exp \left(-\frac{1}{\lambda} d^2_M(x_i, x_j) \right) \quad (30)
\]

\[
 q_{ij} := q(y_j | y_i) = \frac{1}{Z_{q_i}} \exp \left(-d^2_M(y_i, y_j) \right), \quad (31)
\]

where \(Z_{p_i}\) and \(Z_{q_i}\) are the normalizing constants and \(\lambda\) is the input scale parameter. The original SNE formulation uses \(M = \mathbb{R}^n\). In this case, the probabilities are proportional to an isotropic Gaussian \(\mathcal{N}(y_j | y_i, \lambda)\). As defined above, it is our (natural) generalization to Riemannian manifolds.

The embeddings are then learned by minimizing the sum of Kullback-Leibler (KL) divergences between the two families of distributions, \(p_i := p(\cdot | x_i)\) and \(q_i := q(\cdot | y_i)\),

\[
 \mathcal{L}_{\text{SNE}}(Y) := \sum_{i=1}^{m} \text{D}_{\text{KL}}[p_i || q_i]. \quad (32)
\]

The connection to the local neighborhood regime from \(25\) is stated next.

**Lemma 1** For \(\lambda \to 0\), minimizing \(32\) is equivalent to maximizing the sum of the following per-node terms

\[
 \frac{1}{|N(i)|} \sum_{j \in N(i)} \log \frac{\exp(-d^2_M(y_i, y_j))}{\sum_{k \neq i} \exp(-d^2_M(y_i, y_k))}.
\]

**Proof** The result follows directly from the definition of the KL divergence, \(\text{D}_{\text{KL}}[p_i || q_i] = -\sum_{j \neq i} p_{ij} \log q_{ij} + \text{const.}\) and the limit of the distributions defined in \(30\),

\[
 \lim_{\lambda \to 0} p_{ij} = \frac{1}{|N(i)|} \sum_{k \in N(i)} \delta(x_k - x_j).
\]

The Euclidean intuition is that of a Gaussian becoming “infinitely peaked” around \(x_i\), so its nearest neighbors will have “infinitely more” mass assigned to them than the others. Interestingly, it has been remarked that feeding squared distances to the objective function improves training stability in certain cases because they are continuously differentiable (De Sa et al., 2018). In this regard, Lemma 1 serves as a more principled justification for doing that in \(25\).

Finally, we point out that a connection to an MDS-like loss function is mentioned in (Hinton & Roweis, 2003, Section 6), in the regime \(\lambda \to \infty\), but we have not been able to make sense out of it. That being said, appealing to intuition, we expect that for a large \(\lambda\), the objective \(32\) tends towards placing equal emphasis on the relative distances between all pairs of points, thus behaving similar to eqs. \(26\) and \(27\). The advantage is that the temperature-like parameter \(\lambda\) acts as a knob for controlling the optimization goal.

4.2. F1@k – Generalizing Ranking Fidelity

To the best of our knowledge, none of the metrics proposed in the literature can quantify the ranking fidelity of nodes that are \(k\) hops away from a source node, with \(k > 1\). Recall that the motivation stems, for one, from the limitation of mean average precision to immediate neighbors, and, at the other side of the spectrum, from the sensitivity to absolute values of non-ranking metrics such as the average distortion.

In what follows, we will assume that the input graph \(G\) is unweighted. The definitions can be adapted to graphs with edge weights, but in most of our experiments we have used unweighted graphs, so we limit the treatment as such.

For an input graph \(G\), we denote by \(L_G(u; k)\) the set of nodes that are exactly \(k\) hops away from a source node \(u\) (i.e., on “layer” \(k\)), and by \(B_G(v; u)\) the set of nodes that are closer to node \(u\) than another node \(v\). We can now define the **precision** and the **recall** for the ordered pair \((u, v)\).

**Definition 2** For an embedding \(f : G \to M\), the precision and recall of a node \(v\) in the shortest-path tree rooted at \(u\), with \(u \neq v\), are given by

\[
 P(v; u) := \frac{|B_G(v; u) \cap B_M(f(v); f(u))|}{|B_M(f(v); f(u))|}, \quad (33)
\]

\[
 R(v; u) := \frac{|B_G(v; u) \cap B_M(f(v); f(u))|}{|B_G(v; u)|}. \quad (34)
\]

They follow the conventional definitions. For instance, the numerator is the number of true positives: the nodes that appear before \(v\) in the shortest-path tree rooted at \(u\) and, at the same time, are embedded closer to \(u\) than \(v\) is. Moreover, notice that our definition of precision recovers the one used in \(28\) when restricting to layer-1 nodes (i.e., neighbors).

The definition of the F1 score of \((u, v)\), denoted by \(F_1(v; u)\), follows naturally as the harmonic mean of precision and recall. Then, the F1@k metric is obtained by averaging the F1 scores of all nodes that are on layer \(k \geq 1\), across all
5. Experiments

We restrict our experiments to evaluating the graph reconstruc-
tion capabilities of the proposed matrix manifolds relative to
the constant curvature baseline spaces. An analysis via
properties of nearest-neighbor graphs constructed from
random samples is included in Appendix F. Our code is
accessible at http://github.com/dalab/matrix-manifolds.

Training Details We compute and save all-pairs shortest-
paths in all input graphs. Then, we optimize a set of embed-
dings for each combination of optimization setting and loss
function, including both the newly proposed Riemannian
SNE, for several values of $\lambda$, and the ones used in prior work
(Section 4). This is described in more detail in Appendix F.

Evaluation We report on F1@1, the area under the F1@k
curve, and the average distortion. Given our transductive
inference setting (i.e., lower loss is better), we report the best
performing numbers across the aforementioned repetitions.

Synthetic Graphs We begin by showcasing the F1@k
metric that we advocate for several generated graphs in Fig-
ure 5. On the $10 \times 10 \times 10$ grid and the 500-nodes cycle

all manifolds perform well. This is because every Riemann-
ian manifold generalizes Euclidean space and Euclidean
geometry suffices for grids and cycles (e.g., a cycle looks
locally as a line). The more discriminative ones are the two
other graphs – a full balanced tree (branching factor $r = 4$
and depth $h = 5$) and a cycle of 10 trees ($r = 3$ and $h = 4$).
The best performing embeddings involve a hyperbolic com-
ponent while the SPD ones come between those and the
non-negatively curved ones, which are indistinguishable.
The results confirm our expectations: (more) negative cur-
vature is useful when embedding trees. Finally, notice that
the high-temperature SNE regime encourages the recovery
of the global structure more than the local neighborhoods.

Real-world Graphs We compare SPD and hyperbolic
spaces on several real datasets (Table 2). Details about them

Table 2. The “$S^+$” vs. “$\mathbb{H}$” results on 4 datasets and 2 dimensions.
Best and second-best (or slightly better) results are highlighted.

| Graph     | Dim | Manifold | F1@1 | AUC   | Avg. Dist. |
|-----------|-----|----------|------|-------|------------|
| facebook  | 3   | Eucl     | 70.28| 95.27 | 0.193      |
|           |     | Hyp      | 71.08| 95.46 | 0.173      |
|           |     | SPD      | 71.09| 95.26 | 0.170      |
|           |     | Stein    | 75.91| 95.59 | 0.114      |
|           | 6   | Eucl     | 79.60| 96.41 | 0.090      |
|           |     | Hyp      | 81.83| 96.53 | 0.089      |
|           |     | SPD      | 79.52| 96.37 | 0.090      |
|           |     | Stein    | 83.95| 96.74 | 0.061      |
| web-edu   | 3   | Eucl     | 29.18| 87.14 | 0.245      |
|           |     | Hyp      | 55.60| 92.10 | 0.245      |
|           |     | SPD      | 29.02| 88.54 | 0.246      |
|           |     | Stein    | 48.28| 90.87 | 0.084      |
|           | 6   | Eucl     | 49.31| 91.19 | 0.143      |
|           |     | Hyp      | 66.23| 95.78 | 0.143      |
|           |     | SPD      | 42.16| 91.90 | 0.142      |
|           |     | Stein    | 62.81| 96.51 | 0.043      |
| bio-disease | 3 | Eucl     | 83.78| 91.21 | 0.145      |
|           |     | Hyp      | 86.21| 95.72 | 0.137      |
|           |     | SPD      | 83.99| 91.32 | 0.140      |
|           |     | Stein    | 86.70| 94.54 | 0.105      |
|           | 6   | Eucl     | 93.48| 95.84 | 0.073      |
|           |     | Hyp      | 96.50| 98.42 | 0.071      |
|           |     | SPD      | 93.83| 95.93 | 0.072      |
|           |     | Stein    | 94.86| 97.64 | 0.066      |
| power     | 3   | Eucl     | 49.34| 87.84 | 0.119      |
|           |     | Hyp      | 60.18| 91.28 | 0.068      |
|           |     | SPD      | 52.48| 90.17 | 0.121      |
|           |     | Stein    | 54.06| 90.16 | 0.076      |
|           | 6   | Eucl     | 63.62| 92.09 | 0.061      |
|           |     | Hyp      | 75.02| 94.34 | 0.060      |
|           |     | SPD      | 67.69| 91.76 | 0.062      |
|           |     | Stein    | 70.70| 93.32 | 0.049      |

Figure 5. F1@k curves (left y-axis) and PDFs of node pairs per hop
counts (right y-axis) for several synthetic graphs. The objective
was SNE at high temperature $\lambda$.

shortest-path trees. That is, with $K = \sum_{u \in G} | L_G(u;k)|$,

$$F_1(k) := \frac{1}{K} \sum_{u \in G} \sum_{v \in L_G(u;k)} F_1(v; u).$$

This draws a curve $\{(k, F_1(k))\}_{k \in [d(G)]}$, where $d(G)$
denotes the diameter of the graph. In our results, we sometimes
summarize performance via the area under $F_1(k)$.
Computationally Tractable Riemannian Manifolds for Graph Embeddings

Figure 6. The graphs embedded in Tables 2 and 3 (‘facebook’ is shown in Figure 1). More such drawings are included in Appendix G.

Figure 7. Distributions of (normalized) sum-of-angles in geodesic triangles formed by the learned embeddings that yield the best F1@1 metrics (up) and the best average distortion metrics (down), for all datasets from Table 2, for n = 3. 10000 triples are sampled.

and an analysis of their geometric properties are attached in Appendix G. We plot the ones shown here in Figures 1 and 6. Extended results are included in Appendix H.

Discussion First of all, we see that the (partial) negative curvature of the SPD and hyperbolic manifolds is beneficial: they outperform the flat Euclidean embeddings in almost all scenarios. This can be explained by the complex-network structure of the input graphs (Krioukov et al., 2010). Second, we see that especially when using the Stein divergence, the SPD embeddings achieve significant improvements on the average distortion metric and are competitive and sometimes better w.r.t. the ranking metrics. We attribute this to a better-behaved optimization task thanks to its geodesic convexity and stable gradients (see Section 3.1).

How Do the Embeddings Curve? It is a priori unclear to what extent the curvature of the embedding space is lever-aged. To shed light on that, we employ our technique based on sum-of-angles in geodesic triangles (see Section 2.4). We recognize in Figure 7 something remarkable: the better performing embeddings (as per Table 2) yield more negatively-curved triangles. Notice, for instance, the collapsed box plot corresponding to the “web-edu” hyperbolic embedding (a), i.e., almost all triangles have sum-of-angles close to 0. This is explained by its obvious tree-like structure (Figure 6a). Similarly, the SPD-Stein embedding of “facebook” outperforms the hyperbolic one in terms of F1@1 and that reflects in the slightly more stretched box plot (b). Moreover, the pattern applies to the best average-distortion embeddings, where the SPD-Stein embeddings are the only ones leveraging negative curvature and, hence, perform better – the only exception is the “power” graph (c), for which indeed Table 2 confirms that the hyperbolic embeddings are slightly better.

Compact Embeddings We embed several graphs with traits associated with positive curvature in Grassmann manifolds and compare them to spherical embeddings. Table 3

Table 3. The “Gr vs. S” results on 2 datasets and 3 dimensions. The dataset “cat-cortex” (Scannell et al., 1995) is a dissimilarity matrix, lacking graph information, so F1@k cannot be computed.

| Graph      | Dim | Manifold       | F1@1 | AUC  | Avg. Dist. |
|------------|-----|----------------|------|------|------------|
| road-minnesota | 2   | Sphere         | 82.19| 94.02| 0.085      |
|            |     | Gr(1, 3)       | 78.91| 91.04| 0.085      |
|            | 3   | Sphere         | 89.55| 95.89| 0.059      |
|            |     | Gr(1, 4)       | 90.02| 95.88| 0.058      |
|            | 4   | Sphere         | 93.65| 96.66| 0.049      |
|            |     | Gr(1, 5)       | 93.89| 96.67| 0.049      |
|            |     | Gr(2, 4)       | 94.01| 96.66| 0.049      |
| cat-cortex | 2   | Sphere         | -    | -    | 0.255      |
|            |     | Gr(1, 3)       | -    | -    | 0.234      |
|            | 3   | Sphere         | -    | -    | 0.195      |
|            |     | Gr(1, 4)       | -    | -    | 0.168      |
|            | 4   | Sphere         | -    | -    | 0.156      |
|            |     | Gr(1, 5)       | -    | -    | 0.139      |
|            |     | Gr(2, 4)       | -    | -    | 0.129      |
shows that the former yields non-negligibly lower average distortion on the “cat-cortex” dissimilarity dataset and that the two are on-par on the “road-minnesota” graph (displayed in Figure 6c – notice its particular structure, characterized by cycles and low node degrees). More such results are included in Appendix H. As a general pattern, we find learning compact embeddings to be optimization-unfriendly.

6. Conclusion

We proposed embedding graph nodes into matrix spaces of non-constant sectional curvature, such as the SPD and Grassmann manifolds. Leveraging their powerful representational capabilities, we showed that they can consistently and significantly improve over Euclidean embeddings as well as often outperform hyperbolic and elliptical ones on the graph reconstruction task. This suggests that their geometry can accommodate certain graphs with better precision and less distortion than other embedding spaces. We also advocate the Riemannian SNE objective for learning embeddings and explained how, in a sense, it unifies previously used loss functions. Finally, we defined the F1@k metric as an extension of mAP for quantifying ranking fidelity.

Acknowledgements

We would like to thank Andreas Bloch for suggesting the curvature quantification approach based on the sum of angles in geodesic triangles. We are grateful to Prof. Thomas Hofmann for making this collaboration possible.

Gary Bécigneul is funded by the Max Planck ETH Center for Learning Systems.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. Optimization algorithms on matrix manifolds. Princeton University Press, 2009.

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. Log-euclidean metrics for fast and simple calculus on diffusion tensors. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 56(2):411–421, 2006.

Becigneul, G. and Ganea, O.-E. Riemannian adaptive optimization methods. In International Conference on Learning Representations, 2019.

Belkin, M. and Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In Advances in neural information processing systems, pp. 585–591, 2002.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8): 1798–1828, 2013.

Bengio, Y. et al. Learning deep architectures for ai. Foundations and trends® in Machine Learning, 2(1):1–127, 2009.

Bhatia, R. Positive definite matrices, volume 24. Princeton university press, 2009.

Bhatia, R., Jain, T., and Lim, Y. On the bures–wasserstein distance between positive definite matrices. Expositiones Mathematicae, 37(2):165–191, 2019.

Bonnabel, S. Stochastic gradient descent on riemannian manifolds. IEEE Transactions on Automatic Control, 58 (9):2217–2229, 2013.

Bridson, M. R. and Haefliger, A. Metric spaces of non-positive curvature, volume 319. Springer Science & Business Media, 2013.

Bronstein, A. M., Bronstein, M. M., and Kimmel, R. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. Proceedings of the National Academy of Sciences, 103(5):1168–1172, 2006.

Bruna, J., Zaremba, W., Szlam, A., and Lecun, Y. Spectral networks and locally connected networks on graphs. In International Conference on Learning Representations (ICLR2014), CBLS, April 2014, 2014.

Cardoso, J. R. and Leite, F. S. Exponentials of skew-symmetric matrices and logarithms of orthogonal matrices. Journal of computational and applied mathematics, 233(11):2867–2875, 2010.

Carmo, M. P. d. Riemannian geometry. Birkhäuser, 1992.

Cho, A., Shin, J., Hwang, S., Kim, C., Shim, H., Kim, H., Kim, H., and Lee, I. Wormnet v3: a network-assisted hypothesis-generating server for caenorhabditis elegans. Nucleic acids research, 42(W1):W76–W82, 2014.

Chossat, P. and Faugeras, O. Hyperbolic planforms in relation to visual edges and textures perception. PLoS computational biology, 5(12):e1000625, 2009.

Cohen, N., Coudert, D., and Lancin, A. On computing the gromov hyperbolicity. Journal of Experimental Algorithmics (JEA), 20:1–6, 2015.

De Nooy, W., Mrvar, A., and Batagelj, V. Exploratory social network analysis with Pajek: Revised and expanded edition for updated software, volume 46. Cambridge University Press, 2018.
Computationally Tractable Riemannian Manifolds for Graph Embeddings

De Sa, C., Gu, A., Ré, C., and Sala, F. Representation tradeoffs for hyperbolic embeddings. *Proceedings of machine learning research*, 80:4460, 2018.

Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pp. 3844–3852, 2016.

Dolcetti, A. and Pertici, D. Differential properties of spaces of symmetric real matrices. *arXiv preprint arXiv:1807.01113*, 2018a.

Dolcetti, A. and Pertici, D. Skew symmetric logarithms and geodesics on on o(n, r). *Advances in Geometry*, 18(4):495–507, 2018b.

Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Fournier, H., Ismail, A., and Vigneron, A. Computing the gromov hyperbolicity of a discrete metric space. *Information Processing Letters*, 115(6-8):576–579, 2015.

Ganea, O., Becigneul, G., and Hofmann, T. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pp. 1646–1655, 2018.

Gleich, D., Zhukov, L., and Berkhin, P. Fast parallel pagerank: A linear system approach. *Yahoo! Research Technical Report YRL-2004-038*, 13:22, 2004.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

Gromov, M. Hyperbolic groups. In *Essays in group theory*, pp. 75–263. Springer, 1987.

Grove, A. and Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016.

Gu, A., Sala, F., Gunel, B., and Ré, C. Learning mixed-curvature representations in product spaces. 2018.

Henaff, M., Bruna, J., and LeCun, Y. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

Hinton, G. E. and Roweis, S. T. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pp. 857–864, 2003.

Hofmann, T. and Buhmann, J. Multidimensional scaling and data clustering. In *Advances in neural information processing systems*, pp. 459–466, 1995.

Jeuris, B. Riemannian optimization for averaging positive definite matrices. 2015.

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguná, M. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

Lang, S. *Fundamentals of differential geometry*, volume 191. Springer Science & Business Media, 2012.

Lee, J. M. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006.

Leskovec, J. and Krevl, A. SNAP Datasets: Stanford large network dataset collection. *http://snap.stanford.edu/data*, June 2014.

Leskovec, J. and Mcauley, J. J. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pp. 539–547, 2012.

Levoy, M., Gerth, J., Curless, B., and Pull, K. The stanford 3d scanning repository. *URL http://www-graphics.stanford.edu/data/3dscanrep*, 5, 2005.

Lin, Y., Lu, L., and Yau, S.-T. Ricci curvature of graphs. *Tohoku Mathematical Journal, Second Series*, 63(4):605–627, 2011.

Muzellec, B. and Cuturi, M. Generalizing point embeddings using the wasserstein space of elliptical distributions. In *Advances in Neural Information Processing Systems*, pp. 10237–10248, 2018.

Ni, C.-C., Lin, Y.-Y., Gao, J., Gu, X. D., and Saucan, E. Ricci curvature of the internet topology. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 2758–2766. IEEE, 2015.

Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, pp. 6338–6347, 2017.

Nickel, M. and Kiela, D. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pp. 3779–3788, 2018.

Nie, F., Zhu, W., and Li, X. Unsupervised large graph embedding. In *Thirty-first AAAI conference on artificial intelligence*, 2017.
Ollivier, Y. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256:810–864, 2009.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Pennec, X. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127, 2006.

Pennec, X., Fillard, P., and Ayache, N. A riemannian framework for tensor computing. *International Journal of computer vision*, 66(1):41–66, 2006.

Rose, K. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998.

Rossi, R. A. and Ahmed, N. K. The network data repository with interactive graph analytics and visualization. In AAAI, 2015. URL http://networkrepository.com.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Sarkar, R. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pp. 355–366. Springer, 2011.

Scannell, J. W., Blakemore, C., and Young, M. P. Analysis of connectivity in the cat cerebral cortex. *Journal of Neuroscience*, 15(2):1463–1483, 1995.

Skopek, O., Ganea, O.-E., and Bcigneul, G. Mixed-curvature variational autoencoders. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1g6xeSKDS.

Sra, S. A new metric on the manifold of kernel matrices with application to matrix geometric means. In *Advances in neural information processing systems*, pp. 144–152, 2012.

Sra, S. and Hosseini, R. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Tifrea, A., Becigneul, G., and Ganea, O.-E. Poincare glove: Hyperbolic word embeddings. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Ske5r3AqK7.

Tron, R., Afsari, B., and Vidal, R. Average consensus on riemannian manifolds with bounded curvature. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pp. 7855–7862. IEEE, 2011.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. Complex embeddings for simple link prediction. International Conference on Machine Learning (ICML), 2016.

Wang, X., Cui, P., Wang, J., Pei, J., Zhu, W., and Yang, S. Community preserving network embedding. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Watts, D. J. and Strogatz, S. H. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440, 1998.

Wei, X., Xu, L., Cao, B., and Yu, P. S. Cross view link prediction by learning noise-resilient representation consensus. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1611–1619, 2017.

Wilson, R. C., Hancock, E. R., Pekalska, E., and Duin, R. P. Spherical and hyperbolic embeddings of data. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2255–2269, 2014.

Wong, Y.-C. Sectional curvatures of grassmann manifolds. *Proceedings of the National Academy of Sciences of the United States of America*, 60(1):75, 1968.

Zhang, J., Zhu, G., Heath Jr, R. W., and Huang, K. Grassmannian learning: Embedding geometry awareness in shallow and deep learning. *arXiv preprint arXiv:1808.02229*, 2018.

Zhou, C., Liu, Y., Liu, X., Liu, Z., and Gao, J. Scalable graph embedding for asymmetric proximity. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
A. Overview of Differential & Riemannian Geometry

In this section, we introduce the foundational concepts from differential geometry, a discipline that has arisen from the study of differentiable functions on curves and surfaces, thus generalizing calculus. Then, we go a step further, into the more specific Riemannian geometry which enables the abstract definitions of lengths, angles, and curvatures. We base this section on (Carmo, 1992; Absil et al., 2009) and point the reader to them for a more thorough treatment of the subject.

Differentiable Manifolds Informally, an n-dimensional manifold is a set \( M \) which locally resembles \( n \)-dimensional Euclidean space. To be formal, one first introduces the notions of charts and atlases. A bijection \( \varphi \) from a subset \( U \subset M \) onto an open subset of \( \mathbb{R}^n \) is called an \( n \)-dimensional chart of the set \( M \), denoted by \((U, \varphi)\). It enables the study of points \( x \in M \) via their coordinates \( \varphi(x) \in \mathbb{R}^n \). A differentiable atlas \( A \) of \( M \) is a collection of charts \((U_\alpha, \varphi_\alpha)\) of the set \( M \) such that

(i) \( \bigcup_\alpha U_\alpha = M \)

(ii) For any pair \( \alpha, \beta \) with \( U_\alpha \cap U_\beta \neq \emptyset \), the sets \( \varphi_\alpha(U_\alpha \cap U_\beta) \) and \( \varphi_\beta(U_\alpha \cap U_\beta) \) are open sets in \( \mathbb{R}^n \) and the change of coordinates \( \varphi_\beta \circ \varphi_\alpha^{-1} \) is smooth.

Two atlases \( A_1 \) and \( A_2 \) are equivalent if they generate the same maximal atlas. The maximal atlas \( A^+ \) is the set of all charts \((U, \varphi)\) such that \( A \cup \{(U, \varphi)\} \) is also an atlas. It is also called a differentiable structure on \( M \). With that, an \( n \)-dimensional differentiable manifold is a couple \((M, A^+)\), with \( M \) a set and \( A^+ \) a maximal atlas of \( M \) into \( \mathbb{R}^n \). In more formal treatments, \( A^+ \) is also constrained to induce a well-behaved topology on \( M \).

Embedded Submanifolds and Quotient Manifolds

How is a differentiable structure for a set of interest usually constructed? From the definition above it is clear that it is something one endows the set with. That being said, in most useful cases it is not explicitly chosen or constructed. Instead, a rather recursive approach is taken: manifolds are obtained by considering either subsets or quotients (see last subsection paragraph) of other manifolds, thus inheriting a “natural” differentiable structure. Where does this recursion end? It mainly ends when one reaches a vector space, which is trivially a manifold via the global chart \( \varphi: \mathbb{R}^{n \times k} \to \mathbb{R}^n \), with \( X \mapsto \text{vec}(X) \). That is the case for the matrix manifolds considered in Section 3 too.

What makes the aforementioned construction approach (almost) assumption-free is the following essential property: if \( M \) is a manifold and \( N \) is a subset of the set \( M \) (respectively, a quotient \( M/\sim \)), then there is at most one differentiable structure that agrees with the subset topology (respectively, with the quotient projection).

7 The resulting manifolds are called embedded submanifolds and quotient manifolds, respectively. Sufficient conditions for their existence (hence, uniqueness) are known too and do apply for our manifolds. For instance, the submersion theorem says that for a smooth function \( F: M_1 \to M_2 \) with

\[
\dim(M_1) = d_1 > d_2 = \dim(M_2)
\]

and a point \( y \in M_2 \) such that \( F \) has full rank\(^8\) for all \( x \in F^{-1}(y) \), its preimage \( F^{-1}(y) \) is a closed embedded submanifold of \( M_1 \) and \( \dim(F^{-1}(y)) = d_1 - d_2 \).

The quotient of a set \( M \) by an equivalence relation \( \sim \) is defined as

\[
M/\sim := \{ \{x\} : x \in M \},
\]

with \( [x] := \{y \in M : y \sim x\} \) – the equivalence class of \( x \). The function \( \pi: M \to M/\sim \), given by \( x \mapsto [x] \), is the canonical projection. The simplest example of a quotient manifold is the real projective space, \( \mathbb{RP}^n \). It is the set of lines through the origin in \( \mathbb{R}^{n+1} \). With the notation \( \mathbb{RP}^n = \mathbb{R}^{n+1} \setminus \{0\} \), the real projective space can be identified with the quotient \( \mathbb{R}^n/\sim \) given by the equivalence relation

\[
x \sim y \iff \exists t \in \mathbb{R}_*: y = tx.
\]

Tangent Spaces To do even basic calculus on a manifold, one has to properly define the derivatives of manifold curves, \( \gamma: \mathbb{R} \to M \), as well as the directional derivatives of smooth real-valued functions defined on the manifold, \( f: M \to \mathbb{R} \).

The usual definitions,

\[
\gamma'(t) := \lim_{\tau \to 0} \frac{\gamma(t + \tau) - \gamma(t)}{\tau}
\]

\[
D f(x)[\eta] := \lim_{t \to 0} \frac{f(x + t \eta) - f(x)}{t},
\]

are invalid as such because addition does not make sense on general manifolds. However, notice that \( f \circ \gamma: t \mapsto f(\gamma(t)) \) is differentiable in the usual sense. With that in mind, let \( T_x(M) \) denote the set of smooth real-valued functions defined on a neighborhood of \( x \). Then, the mapping \( \dot{\gamma}(0) \) from \( T_x(M) \) to \( \mathbb{R} \) defined by

\[
\dot{\gamma}(0) f := \left. \frac{df(\gamma(t))}{dt} \right|_{t=0}
\]

is called the tangent vector to the curve \( \gamma \) at \( t = 0 \). The equivalence class

\[
[\dot{\gamma}(0)] := \{ \gamma_1: \mathbb{R} \to M : \dot{\gamma}_1(0) f = \dot{\gamma}(0) f, \forall f \in T_x(M) \}
\]

\(^7\)We say almost assumption-free because we still assume that this agreement is desirable.

\(^8\)That is, the Jacobian \( \frac{\partial \gamma(x)}{\partial x} \in \mathbb{R}^{d_2 \times d_1} \) has rank \( d_2 \) irrespective of the chosen charts.
is a tangent vector at \( x \in \mathcal{M} \). The set of such equivalence classes forms the tangent space \( T_x \mathcal{M} \). It is immediate from the linearity of the differentiation operator from (41) that \( T_x \mathcal{M} \) inherits a linear structure, thus forming a vector space.

The abstract definition from above recovers the classical definition of directional derivative from (39) in the following sense: if \( \mathcal{M} \) is an embedded submanifold of a vector space \( \mathcal{E} \) and \( \mathcal{J} \) is the extension of \( f \) in a neighborhood of \( \gamma(0) \in \mathcal{E} \), then

\[
\dot{\gamma}(0) f = D\mathcal{J}(\gamma(0)) [\dot{\gamma}(0)],
\]

so there is a natural identification of the mapping \( \dot{\gamma}(0) \) with the vector \( \dot{\gamma}(0) \).

For quotient manifolds \( \mathcal{M}/\sim \), the tangent space splits into two complementary linear subspaces called the vertical space \( \mathcal{V}_x \) and the horizontal space \( \mathcal{H}_x \). Intuitively, the vectors in the former point in tangent directions which, if we were to follow for an infinitesimal step, we would get another element of \( [x] \). Thus, only the horizontal tangent vectors make us move on the quotient manifold.

**Riemannian Metrics** They are inner products \( \langle \cdot, \cdot \rangle_x \), sometimes denoted by \( g_x(\cdot, \cdot) \), attached to each tangent space \( T_x \mathcal{M} \). They give a notion of length via \( \| \xi_x \|_x := \sqrt{\langle \xi_x, \xi_x \rangle_x} \), for all \( \xi_x \in T_x \mathcal{M} \). A Riemannian metric is an additional structure added to a differentiable manifold \( (\mathcal{M}, A^+, g_x) \), yielding the Riemannian manifold \( (\mathcal{M}, A^+, g_x) \). However, as it was the case for the differentiable structure, for Riemannian submanifolds and Riemannian quotient manifolds it is inherited from the “parent” manifold in a natural way – see (Absil et al., 2009) for several examples.

The Riemannian metric enables measuring the length of a curve \( \gamma : [a, b] \to \mathcal{M} \),

\[
L(\gamma) = \int_a^b \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle_{\gamma(t)}} dt,
\]

which, in turn, yields the Riemannian distance function,

\[
d_\mathcal{M} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}, \quad d_\mathcal{M}(x, y) = \inf_\gamma L(\gamma),
\]

that is, the shortest path between two points on the manifold. The infimum is taken over all curves \( \gamma : [a, b] \to \mathcal{M} \) with \( \gamma(a) = x \) and \( \gamma(b) = y \). Note that in general it is only defined locally because \( \mathcal{M} \) might have several connected components or it might not be geodesically complete (see next paragraphs). Deriving a closed-form expression of it is paramount for graph embedding purposes.

The Riemannian gradient of a smooth function \( f : \mathcal{M} \to \mathbb{R} \) at \( x \), denoted by \( \nabla^R f(x) \) is defined as the unique element of \( T_x \mathcal{M} \) that satisfies

\[
\langle \nabla^R f(x), \xi \rangle_x = D f(x)[\xi], \quad \forall \xi \in T_x \mathcal{M}.
\]

**Retractions** Up until this point, we have not introduced the concept of “moving in the direction of a tangent vector”, although we have intuitively used it. This is achieved via retractions. At a point \( x \in \mathcal{M} \), the retraction \( R_x \) is a map from \( T_x \mathcal{M} \) to \( \mathcal{M} \) satisfying local rigidity conditions: \( R_x(0) = x \) and \( D R_x(0) = \text{Id}_{\mathcal{T}x} \). For embedded submanifolds, the two-step approach consisting of (i) taking a step along \( \xi \) in the ambient space, and (ii) projecting back onto the manifold, defines a valid retraction. In quotient manifolds, the retractions of the base space that move an entire equivalence class to another equivalence class induce retractions on the quotient space.

**Riemannian Connections** Let us first briefly introduce and motivate the need for an additional structure attached to differentiable manifolds, called affine connections. They are functions

\[
\nabla : \mathbb{X} \times \mathbb{X} \to \mathbb{X}, \quad (\xi, \zeta) \to \nabla_{\xi}\zeta
\]

where \( \mathbb{X} \) is the set of vector fields on \( \mathcal{M} \), i.e., functions assigning to each point \( x \in \mathcal{M} \) a tangent vector \( \xi_x \in T_x \mathcal{M} \). They satisfy several properties that represent the generalization of the directional derivative of a vector field in Euclidean space. Affine connections are needed, for instance, to generalize second-order optimization algorithms, such as Newton’s method, to functions defined on manifolds.

The Riemannian connection, also known as the Levi-Civita connection, is the unique affine connection that, besides the properties referred to above, satisfies two others, one of which depends on the Riemannian metric – see (Carmo, 1992) for details. It is the affine connection implicitly assumed when working with Riemannian manifolds.

**Geodesics, Exponential Map, Logarithm Map, Parallel Transport** They are all concepts from Riemannian geometry, defined in terms of the Riemannian connection. A geodesic is a curve with zero acceleration,

\[
\nabla_{\dot{\gamma}(t)} \dot{\gamma}(t) = 0.
\]

Geodesics are the generalization of straight lines from Euclidean space. They are locally distance-minimizing and parameterized by arc-length. Thus, for every \( \xi \in T_x \mathcal{M} \), there exists a unique geodesic \( \gamma(t; x, \xi) \) such that \( \gamma(0) = x \) and \( \dot{\gamma}(0) = \xi \).

The exponential map is the function

\[
\exp_x : \tilde{U} \subseteq T_x \mathcal{M} \to \mathcal{M}, \quad \xi \mapsto \exp_x(\xi) := \gamma(1; x, \xi),
\]

where \( \tilde{U} \) is a neighborhood of 0. The manifold is said to be geodesically complete if the exponential map is defined on the entire tangent space, i.e., \( \tilde{U} = T_x \mathcal{M} \). It can be shown that \( \exp_x(\cdot) \) is a retraction and satisfies the following useful
The exponential map defines a diffeomorphism between a neighborhood of 0 ∈ T_xM onto a neighborhood of x ∈ M. If we follow a geodesic γ(t; x, ξ) from t = 0 to infinity, it can happen that it is minimizing only up to t_0 < ∞. If that is the case, the point y = γ(t_0; x, ξ) is called a cut point. The set of all such points, gathered across all geodesics starting at x, is called the cut locus of M at x, C(x) ⊂ M. It can be proven that the cut locus has finite measure (Pennec, 2006). The maximal domain where the exponential map is a diffeomorphism is given by its preimage on M \ C(x).

Hence, the inverse is called the logarithm map,\[ \log_x : M \setminus C(x) \to T_x M. \] (51)

The parallel transport of a vector ξ ∈ T_xM along a curve γ : I → M is the unique vector field ξ that points along the curve to tangent vectors, satisfying\[ \nabla_{\dot{\gamma}(t)} \xi(\gamma(t)) = 0 \quad \text{and} \quad \xi(x) = \xi_x. \] (52)

Curvature The Riemann curvature tensor is a tensor field that assigns a tensor to each point of a Riemannian manifold. Each such tensor measures the extent to which the manifold is not locally isometric to Euclidean space. It is defined in terms of the Levi-Civita connection. For each pair of tangent vectors u, v ∈ T_xM, R_x(u, v) is a linear transformation on the tangent space. The vector w = R_x(u, v) w quantifies the failure of parallel transport to bring w back to its original position when following a quadrilateral determined by -tu, -tv, tu, tv, with t → 0. This failure is caused by curvature and it is also known as the infinitesimal non-holonomy of the manifold.

The sectional curvature is defined for a fixed point x ∈ M and two tangent vectors u, v ∈ T_xM as\[ K_x(u, v) = \frac{\langle R_x(u, v)v, u \rangle_x}{\langle u, u \rangle_x \langle v, v \rangle_x} - \langle u, v \rangle_x^2. \] (53)

It measures how far apart two geodesics emanating from x diverge. If it is positive, the two geodesics will eventually converge. It is the most common curvature characterization that we use to compare, from a theoretical perspective, the manifolds discussed in this work.

The Ricci tensor Ric(u, v) is defined as the trace of the linear map T_xM → T_xM given by u → R(u, v)w. It is fully determined by specifying the scalars Ric(u, u) for all unit vectors u ∈ T_xM, which is known simply as the Ricci curvature. It is equal to the average sectional curvature across all planes containing u times (n − 1). Intuitively, it measures how the volume of a geodesic cone in direction u compares to that of an Euclidean cone.

Finally, the scalar curvature (or Ricci scalar) is the most coarse-grained notion of curvature at a point on a Riemannian manifold. It is the trace of the Ricci tensor, or, equivalently, n(n − 1) times the average of all sectional curvatures. Note that a space of non-constant sectional curvature can have constant Ricci scalar. This is true, in particular, for homogeneous spaces.

B. Differential Geometry Tools for Hyperbolic and Elliptical Spaces

We include in Table 4 the differential geometry tools for the hyperbolic and elliptical spaces. We use the hyperboloid model for the former and the hyperspherical model for the latter, depicted in Figure 8. Some prior work prefers working with the Poincaré ball model and/or the stereographic projection. They have both advantages and disadvantages.

For instance, our choice yields simple formulas for certain quantities of interest, such as exponential and logarithm maps. They are also more numerically stable. In fact, it is claimed in (Nickel & Kiela, 2018) that numerical stability, together with its impact on optimization, is the only explanation for the Lorentz model outperforming the prior experiments in the (theoretically equivalent) Poincaré ball (Nickel & Kiela, 2017).

On the other hand, the just-mentioned alternative models have the strong advantage that the corresponding metric tensors are conformal. This means that they are proportional to the Riemannian metric of Euclidean space,

\[ g_{H} = \left( \frac{2}{1 - \|x\|_2} \right)^2 g_E \quad \text{and} \quad g_{S} = \left( \frac{2}{1 + \|x\|_2} \right)^2 g_E. \] (54)

Notice the syntactic similarity between them. Furthermore,
the effect of the denominators in the \textit{conformal factors} reinforce the intuition we have about the two spaces: distances around far away points are \textit{increasingly larger} in the hyperbolic space and \textit{increasingly smaller} in the elliptical space.

Let us point out that both hyperbolic and elliptical spaces are \textit{isotropic}. Informally, isotropy means “uniformity in all directions.” Note that this is a stronger property than the homogeneity of the matrix manifolds discussed in Section 3 (see also Appendix D) which means that the space “looks the same around each point.”

### C. Geometric Properties of Graphs

Graphs and manifolds, while different mathematical abstractions, share many similar properties through Laplace operators, heat kernels, and random walks. Another example is the deep connection between trees and the hyperbolic plane: any tree can be embedded in $\mathbb{H}(2)$ with arbitrarily small distortion (Sarkar, 2011). On a similar note, complex networks arise naturally from hyperbolic geometry (Krioukov et al., 2010). With these insights in mind, in this section we review some continuous geometric properties that have been adapted to arbitrary weighted graphs. See also (Ni et al., 2015).

#### Gromov Hyperbolicity

Also known as $\delta$-hyperbolicity (Gromov, 1987), it quantifies with a single number the hyperbolicity of a given metric space: the smaller $\delta$ is, the more hyperbolic-like or negatively-curved the space is. The definition that makes it easier to picture it is via the \textit{slim triangles property}: a metric space \footnote{Recall that a Riemannian manifold with its induced distance function is a metric space only if it is connected.} $(M, d_M)$ is $\delta$-hyperbolic if all geodesic triangles are $\delta$-slim. Three points $x, y, w \in M$ form a $\delta$-slim triangle if any point on the geodesic segment between any two of them is within distance $\delta$ from the other two geodesics (i.e., “sides” of the geodesic triangle).

For discrete metric spaces such as graphs, an equivalent definition using the so-called “4-points condition” can be used to devise algorithms that look at quadruples of points. Both exact and approximating algorithms exist that run fast enough to analyze graphs with tens of thousands of nodes within minutes (Fourier et al., 2015; Cohen et al., 2015). Finally, in practice we look at histograms of $\delta$ instead of the worst-case value.

##### Ollivier-Ricci Curvature

(Ollivier, 2009) generalized the \textit{Ricci curvature} to metric spaces $(M, d_M)$ equipped with a family of probability measures $\{\mu_x \cdot\}_{x \in M}$. It is defined in a way that mimics the interpretation of Ricci curvature on Riemannian manifolds: it is the average distance between two small balls taken relative to the distance between their centers. The difference is that now the former is given by the Wasserstein distance (i.e., Earth mover’s distance) between the corresponding probability measures,

\[
\text{Ric}_1(x, y) := 1 - \frac{W(m_x, m_y)}{d_M(x, y)},
\]

| Property               | Expr.                        | Hyperbolic $\mathbb{H}(n)$ | Elliptical $\mathbb{E}(n)$ |
|------------------------|------------------------------|-----------------------------|-----------------------------|
| Representation         | $\mathbb{H}(n) / \mathbb{S}(n)$ | $\{ x \in \mathbb{R}^{n+1} : \langle x, x \rangle_L = -1, x_0 > 0 \}$ | $\{ x \in \mathbb{R}^{n+1} : \|x\|_2 = 1 \}$ |
| Tangent space          | $T_x M$ | $\{ u \in \mathbb{R}^{n+1} : \langle u, x \rangle_L = 0 \}$ | $\{ u \in \mathbb{R}^{n+1} : x^\top u = 0 \}$ |
| Tsp. projection        | $\pi_x(u)$ | $u + \langle u, x \rangle_L x$ | $\langle u, v \rangle_L$ |
| Riem. metric           | $\langle u, v \rangle_x$ | $\langle u, v \rangle_L$ | $\langle u, v \rangle_L$ |
| Riem. gradient         | $\nabla^R_x$ | $\text{diag}(-1, 1, \ldots, 1) \pi_x(\nabla^E_x)$ | $\pi_x(\nabla^E_x)$ |
| Geodesics              | $\gamma_{x,y}(t)$ | $x \cosh(t) + y \sinh(t)$ | $x \cos(t) + y \cos(t)$ |
| Retraction             | $R_x(u)$ | Not used | Not used |
| Log-map                | $\log_x(y)$ | $\frac{\cosh^{-1}(\alpha)}{\sqrt{\alpha^2 - 1}} (y - \alpha x)$ | $\log_x(y) = \frac{\cosh^{-1}(\|w\|_2)}{\|w\|_2} u'$ with $u' = \pi_x(y - x)$ |
| Riem. distance         | $d(x, y)$ | $\cosh^{-1}(-\langle x, y \rangle_L)$ | $\cosh^{-1}(\langle x, y \rangle)$ |
| Parallel transport     | $T_{x,y}(u)$ | $u + \frac{y - \alpha x}{\alpha + 1} (x + y)$ with $\alpha = -\langle x, y \rangle_L$ | $\pi_b(u)$ |
| Characterizations      |                        | Constant negative curvature | Constant positive curvature |
|                        |                        | Isotropic                    | Isotropic                    |
with \( W(\mu_1, \mu_2) := \inf_{\xi} \int \int d(x, y) d\xi(x, y) \) and \( \xi(x, y) \) - a join distribution with marginals \( \mu_1 \) and \( \mu_2 \). This definition was then specialized in (Lin et al., 2011) for graphs by making \( m_x \) assign a probability mass of \( \alpha \in [0, 1) \) to itself (node \( x \)) and spread the remaining \( 1 - \alpha \) uniformly across its neighbors. They refer to \( \text{Ric}_\alpha(x, y) := \text{Ric}_1(x, y) \) as the \( \alpha \)-Ricci curvature and define instead

\[
\text{Ric}_\alpha(x, y) := \lim_{\alpha \to 1} \frac{\text{Ric}_1(x, y)}{1 - \alpha}
\]

(56)

to be the Ricci curvature of \( (x, y) \) in the graph. In practice, we approximate the limit via a large \( \alpha \), e.g., \( \alpha = 0.999 \).

Notice that in contrast to \( \delta \)-hyperbolicity, the Ricci curvature characterizes the space only locally. It yields the curvatures one would expect in several cases: negative curvatures for trees (except for the edges connecting the leaves) and positive for complete graphs and hypercubes; but it does not capture the curvature of a cycle with more than 5 nodes because it locally looks like a straight line.

**Sectional Curvatures** A discrete analogue of sectional curvature for graphs is obtained as the deviation from the parallelogram law in Euclidean geometry (Gu et al., 2018). It uses the same intuition as before: in non-positively curved spaces triangles are “slimmer” while in non-negatively curved ones they are “thicker” (see Figures 2 and 9).

For any Riemannian manifold \( M \), let \( x, y, z \in M \) form a geodesic triangle, and let \( m \) be the midpoint of the geodesic between \( y \) and \( z \). Then, the following quantity

\[
k_M(x, y, z) := d_M(x, m)^2 + d_M(y, z)^2 - 2d_M(x, y)d_M(x, z)
\]

(57)

has the same sign as the sectional curvatures in \( M \) and it is identically zero if \( M \) is flat. For a graph \( G = (V, E) \), an analogue is defined for each node \( m \) and two neighbors \( y \) and \( z \), as

\[
k_G(m; y, z) = \frac{1}{2d_G(m, x)} k_G(x, y, z).
\]

(58)

With that, the sectional curvature of the graph \( G \) at \( m \) in “directions” \( y \) and \( z \) is defined as the average of (58) across all \( x \in G \),

\[
k_G(m; y, z) = \frac{1}{|V| - 1} \sum_{x \neq m} k_G(m; y, z; x).
\]

(59)

It is also shown in (Gu et al., 2018, AppxC.2) that this definition recovers the expected signs for trees (negative or zero), cycles (positive or zero) and lines (zero).

**D. Matrix Manifolds – Details**

In this section, we include several results that are useful for better understanding and working with the matrix manifolds introduced in Section 3. They have been left out from the main text due to space constraints. Furthermore, we describe the orthogonal group \( O(n) \) which is used in our random manifold graphs analysis from Appendix E. We do not use it for graph embedding purposes in this work because its corresponding distance function does not immediately lend itself to simple backpropagation-based training (in general). This is left for future work.

**Homogeneity** It is a common property of the matrix manifolds used in this work. Formally, it means that the isometry group of \( M \) acts transitively: for any \( A, B \in M \) there is an isometry that maps \( A \) to \( B \). In non-technical terms this says that \( M \) “looks locally the same” around each point. A consequence of homogeneity is in order to prove curvature properties, it suffices to do so at a single point, e.g., the identity matrix for \( S^{++}(n) \).

**D.1. SPD Manifold**

The following theorem, proved here for completeness, states that \( S^{++}(n) \) is a differentiable manifold.

**Theorem 3** The set \( S^{++}(n) \) of symmetric positive-definite matrices is an \( \frac{n(n+1)}{2} \)-dimensional differentiable manifold.

**Proof** The set \( S(n) \) is an \( \frac{n(n+1)}{2} \)-dimensional vector space. Any finite dimensional vector space is a differentiable manifold: fix a basis and use it as a global chart mapping points to the Euclidean space with the same dimension.

The set \( S^{++}(n) \) is an open subset of \( S(n) \). This follows from \( A, v \mapsto v^T A v \) being a continuous function. The fact that open subsets of (differentiable) manifolds are (differentiable) manifolds concludes the proof.

The following result from linear algebra makes it easier to compute geodesic distances between SPD matrices.

**Lemma 4** Let \( A, B \in S^{++}(n) \). Then \( AB \) and \( A^{1/2}BA^{1/2} \) have the same eigenvalues.

**Proof** Let \( A = A^{1/2}A^{1/2} \), where \( A^{1/2} \) is the unique square root of \( A \). Note that \( A^{1/2} \) is itself symmetric and positive-definite because every analytic function \( f(A) \) is equivalent to \( U \text{diag}(\{ f(\lambda_i) \}_j))U^T \), where \( U \text{diag}(\{ \lambda_i \}_j))U^T \) is the eigenvalue decomposition of \( A \), and \( \lambda_i > 0 \) for all \( i \) from positive-definiteness. Then, we have

\[
AB = A^{1/2}A^{1/2}B = A^{1/2}(A^{1/2}BA^{1/2})A^{-1/2},
\]
Therefore, $AB$ and $A^{1/2}BA^{1/2}$ are similar matrices so they have the same eigenvalues. 

This is useful from a computational perspective because $A^{-1/2}BA^{-1/2}$ is an SPD matrix while $A^{-1}B$ may not even be symmetric. It also makes it easier to see the following equivalence, for any matrix $A \in S^{++}(n)$:

$$
\|\log(A)\|_F = \left\|U \text{diag} \left(\log(\lambda_i(A))\right)U^T\right\|_F = \left\|\text{diag} \left(\log(\lambda_i(A))\right)\right\|_F = \sqrt{\sum_{i=1}^{n} \log^2(\lambda_i(A))},
$$

(60)

where the first step decomposes the principle logarithm of $A$ and the second one uses the change of basis invariance of the Frobenius norm.

An even better behaved matrix that has the same spectrum and avoids matrix square roots altogether is $LBL^\top$, where $LL^\top$ is the Cholesky decomposition of $A$. Note that for the Stein divergence (21), the log-determinants can be computed in terms of $L$ as $\log \det(A) = 2 \sum_{i=1}^{n} \log(L_{ii})$.

An additional challenge with eigenvalue computations is that most linear algebra libraries are optimized for large matrices while our use-case involves millions of very small matrices.\(^{10}\) To overcome that, for $2 \times 2$ and $3 \times 3$ matrices we use custom formulas that can be derived explicitly:

- For $A \in S^{++}(2)$, we have

$$
\lambda_k(A) = \frac{t}{2} \pm \sqrt{\left(\frac{t}{2}\right)^2 - d},
$$

(61)

with $t = \text{Tr} A$ and $d = \det(A)$.

- For $A \in S^{++}(3)$, we express it as an affine transformation of another matrix, i.e., $A = pB + q \text{Id}_n$. Then we have $\lambda_k(A) = p\lambda_k(B) + q$. Concretely, if we let

$$
q = \frac{\text{Tr} A}{3} \quad \text{and} \quad p = \sqrt{\frac{\text{Tr} (A - q \text{Id}_n)^2}{6}},
$$

then the eigenvalues of $B$ are

$$
\lambda_k(B) = 2 \cos \left(\frac{1}{3} \arccos \left(\frac{\det(B)}{2}\right) + \frac{2k\pi}{3}\right).
$$

(62)

### D.2. Grassmann Manifold

The following singular value formulas are useful in computing geodesic distances (20) between points on $Gr(2, n)$.

**Proposition 5** The singular values of $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ are

$$
\sigma_1 = \sqrt{\frac{S_1 + S_2}{2}} \quad \text{and} \quad \sigma_2 = \sqrt{\frac{S_1 - S_2}{2}},
$$

with

$$
S_1 = a^2 + b^2 + c^2 + d^2
$$

and

$$
S_2 = \sqrt{(a^2 + b^2 - c^2 - d^2)^2 + 4(ac + bd)^2}.
$$

### D.3. Orthogonal Group

It is defined as the set of $n \times n$ real orthogonal matrices,

$$
O(n) := \{A \in \mathbb{R}^{n \times n} : A^\top A = A A^\top = \text{Id}_n\}. \quad (63)
$$

We are interested in the geometry of $O(n)$ rather than its group properties. Formally, this means that in what follows we describe its so-called principal homogeneous space, the special Stiefel manifold $V(n, n)$, rather than the group $O(n)$.

**Theorem 6** $O(n)$ is an $\frac{n(n-1)}{2}$-dimensional differentiable manifold.

**Proof** Consider the map $F : \mathbb{R}^{n \times n} \to S(n)$, $F(A) = A^\top A - \text{Id}_n$. It is clear that $O(n) = F^{-1}(0)$. The differential at $A$, $DF(A)[B] = A^\top B + B^\top A$, is onto the space of symmetric matrices as a function of the direction $B \in \mathbb{R}^{n \times n}$, as shown by

$$
DF(A)\begin{bmatrix} 1 \\ 2 \end{bmatrix} A P = P \quad \text{for all} \quad P \in S(n).
$$

Therefore, $F(\cdot)$ is full rank and by the submersion theorem (Appendix A; Absil et al. (2009)) the orthogonal group is a differentiable (sub-)manifold. Its dimension is

$$
d = \dim(\mathbb{R}^{n \times n}) - \dim(S(n)) = \frac{n(n-1)}{2}.
$$

As a Riemannian manifold, $O(n)$ inherits the metric structure from the ambient space and thus the Riemannian metric is simply the Frobenius inner product,

$$
\langle P, Q \rangle_A = \text{Tr} P^\top Q.
$$

(64)

for $A \in O(n)$ and $P, Q \in T_A O(n)$. To derive its tangent space, we use the following general result.

**Proposition 7** (Absil et al. (2009)) If $M$ is an embedded submanifold of a vector space $E$, defined as a level set of a constant-rank function $F : E \to \mathbb{R}^n$, we have

$$
T_x M = \ker(D F(x)).
$$
Differentiating $A^TA = I_d$ yields $\dot{A}^T A + A^T \dot{A} = 0$, so the tangent space at $A$ is

$$T_A O(n) = \{ P \in \mathbb{R}^{n \times n} : A^T P = X \in S_{skew}(n) \} = A S_{skew}(n),$$

(65)

where $S_{skew}(n)$ is the space of $n \times n$ skew-symmetric matrices.

Many Riemannian quantities of interest are syntactically very similar to those for the canonical SPD manifold (see Table 1). The connection is made precise by the following property.

**Lemma 8 (Dolcetti & Pertici (2018b))** The metric (64) is the opposite of the affine-invariant metric (9) that the SPD manifold is endowed with.

**Proof** Let $A \in O(n)$ and $P_1, P_2 \in T_A O(n)$. Then,

$$\text{Tr} A^{-1} P_1 A^{-1} P_2 =$$

$$= \text{Tr} A^T X_1 A^T X_2 = \text{Tr} X_1 X_2 \quad (P_1 = AX_1)$$

$$= - \text{Tr} X_1^T X_2 = - \text{Tr} (AX_1)^T AX_2 \quad (X_i \in S_{skew}(n))$$

$$= - \text{Tr} P_1^T P_2.$$

An important characteristic of $O(n)$ is its compact Lie group structure which implies, via the Hopf-Rinow theorem, that its components (see below) are geodesically complete: all geodesics $t \mapsto A \exp(t A^T P)$ are defined on the whole real line. Note, though, that they are minimizing only up to some $t_0 \in \mathbb{R}$. This is another consequence of its compactness.

Moreover, $O(n)$ has two connected components so the expression for the geodesic between two matrices $A, B \in O(n)$,

$$\gamma_{A,B}(t) = A(A^T B)^t,$$

(66)

only makes sense if they belong to the same component. The same is true for the logarithm map,

$$\log_A (B) = A \log(A^T B).$$

(67)

The two components contain the orthogonal matrices with determinant 1 and $-1$, respectively. The former is the so-called special orthogonal group,

$$SO(n) := \{ A \in O(n) : \det(A) = 1 \}.$$

(68)

Restricting to one of them guarantees that the logarithm map is defined on the whole manifold except for the cut locus (see Appendix A). This is true in general for compact connected Lie groups endowed with bi-invariant exponential maps, but we can bring it down to a clearer matrix property by looking at its expression,

$$d(A, B) = \|\log(A^T B)\|_F,$$

(69)

Notice that $A^T B$ is in $SO(n)$ whenever $A$ and $B$ belong to the same component,

$$\det(A) = \det(B) = d \in \{-1, 1\} \implies \det(A^T B) = 1.$$

(70)

Then, the claim is true due the surjectivity of the matrix exponential, as follows.

**Proposition 9 (Cardoso & Leite (2010))** For any matrix $A \in SO(n)$, there exists a matrix $X \in S_{skew}(n)$ such that $\exp(X) = A$, or, equivalently, $\log(A) = X$. Moreover, if $A$ has no negative eigenvalues, there is a unique such matrix with $\text{Im} \lambda_i(X) \in (-\pi, \pi)$, called its principal logarithm.

This property makes it easier to see that the cut locus at a point $A \in SO(n)$ consists of those matrices $B \in SO(n)$ such that $A^T B$ has eigenvalues equal to $-1$. They can be thought of as analogous to the antipodal points on the sphere. It also implies that the distance function (69), expanded as

$$d(A, B) = \|\log(A^T B)\|_F = \sqrt{\sum_{i=1}^n \text{Arg}(\lambda_i(A^T B))^2},$$

(71)

is well-defined for points in the same connected components. The $\text{Arg}()$ operator denotes the complex argument. Notice again the similarity to the canonical SPD distance (19). However, the nicely behaved symmetric eigenvalue decomposition cannot be used anymore and the various approximations to the matrix logarithm are too slow for our graph embedding purposes. That is why we limit our graph reconstruction experiments with compact matrix manifolds to the Grassmann manifold. Moreover, the small dimensional cases where we could compute the complex argument “manually” are isometric to other manifolds that we experiment with: $O(2) \cong S(1)$ and $O(3) \cong Gr(1, 4)$.

---

11Such eigenvalues must come in pairs since $\det(A^T B) = 1$. 

Computationally Tractable Riemannian Manifolds for Graph Embeddings
E. Manifold Random Graphs

The idea of sampling points on the manifolds of interest and constructing nearest-neighbor graphs, used in prior influential works such as (Krioukov et al., 2010), is motivated, in our case, by the otherwise black-box nature of matrix embeddings. However, in Section 5 we do not stop at reporting reconstruction metrics, but look into manifold properties around the embeddings. But even with that, the discretization of the embedding spaces is a natural first step in developing an intuition for them.

The discretization is achieved as follows: several samples (1000 in our case) are generated randomly on the target manifold and a graph is constructed by linking two nodes together (corresponding to two sample points) when the geodesic distance between them is less than some threshold. The details of the random generation is discussed in each of the following two subsections. As mentioned in the introductory paragraph, it is the same approach employed in (Krioukov et al., 2010) who sample from the hyperbolic plane and obtain graphs that resemble complex networks. In our experiments, we go beyond their technique and study the properties of the generated graphs when varying the distance threshold.

E.1. Compact Matrix Manifolds vs. Sphere

We begin the analysis of random manifold graphs with the compact matrix manifolds compared against the elliptical geometry. The sampling is performed uniformly on each of the compact spaces.

The results are shown in Figure 10. The node degrees and the graph sectional curvatures (computed via the “deviation from parallelogram law”; see Appendix C) are on the first two rows. The markers are the median values and the shaded area corresponds to the inter-quartile range (IQR). The last row shows normalized sum-of-angles histograms. All of them are repeated for three consecutive dimensions, organized column-wise. The distance thresholds used to link nodes in the graph range from $d_{\text{max}}/10$ to $d_{\text{max}}$, where $d_{\text{max}}$ is the maximum distance between any two sample points, for each instance except for the Euclidean baseline which uses the maximum distance across all manifolds (i.e., $\approx 4.4$ corresponding to $SO(3)$).

Degree Distributions. From the degree distributions, we first notice that all compact manifolds lead to graphs with higher degrees for the same distance threshold than the ones based on uniform samples from Euclidean balls. This is not surprising given that points tend to be closer due to positive curvature. Furthermore, the distributions are concentrated around the median – the IQRs are hardly visible. To get more manifold-specific, we see that Grassmannians lead to full cliques faster than the spherical geometry, but the curves get closer and closer as the dimension is increased. This is the same behavior that in Euclidean space is known as the “curse of dimensionality”, i.e., a small threshold change takes us from a disconnected graph to a fully-connected one. For the compact manifolds, though, this can be noticed already at a very small dimension: the degree distributions tend towards a step function. The only difference is the point where that happens, which depends on the maximum distance on each manifold. For instance, that is much earlier on the real projective space $\mathbb{RP}(n-1) \cong Gr(1, n)$ than the special orthogonal group $SO(n)$.

Graph Sectional Curvatures. They confirm the similarity between the compact manifolds that we have just hinted at: for each of the three dimensions, the curves look almost identical up to a frequency change (i.e., “stretching them”). We say “almost” because we can still see certain differences between them (explained by the different geometries); for instance, in Figure 10f, the graph curvatures corresponding to Grassmann manifolds are slightly larger at distance threshold $\approx 1$. This frequency change seems to be intimately related to the injectivity radii of the manifolds (see, e.g., Tron et al., 2011). We also see that distributions are mostly positive, matching their continuous analogues. A-priori, it is unclear if manifold discretization will preserve them. Finally, the convergence point is $k_{\text{cl}}(m; x, y) = 1/8$ – the constant sectional curvature of a complete graph (see eqs. (57) and (58)).

Triangles Thickness. The normalized sum-of-angles plots do not depend on the generated graphs: the geodesic triangles are randomly selected from the manifold-sampled points. As a sanity check, we first point out that they are all positive.\textsuperscript{12} We observe that the Grassmann samples yield empirical distributions that look bi-modal. At the same time, the elliptical ones result in normalized sum-of-angles that resemble Poisson distributions, with the dimension playing the role of the parameter $\lambda$. We could not justify these contrasting behaviors, but they show that the spaces curve differently. What we can justify is the perfect overlap of the distributions corresponding to $SO(3)$ and $Gr(1, 4)$ in Figure 10h: the two manifolds are isometric. The seemingly different degree distributions from Figure 10b should, in fact, be identical after a rescaling. In other words, they have different volumes but curve in the same way.

E.2. SPD vs. Hyperbolic vs. Euclidean

In this subsection, using the same framework as for compact manifolds, we compare the non-positively curved manifold of positive-definite matrices to hyperbolic and Euclidean

\textsuperscript{12}To make it clear, note that in contrast to the discrete sectional curvatures of random nearest-neighbor graphs, this is a property of the manifold.
Figure 10. Analysis of graphs sampled from the compact manifolds. The results are grouped by manifold dimension (one column each). The first row (a)-(c) shows the node distributions as the distance threshold used to link two nodes is varied. Samples from an Euclidean ball of radius 4.4 are included for comparison. The shaded range, hardly noticeable for compact manifolds, is the IQR. The second row (d)-(f), with the same x-axis as the first one, shows the distributions of graph sectional curvatures obtained via the deviation from parallelogram law (see Appendix C). The third row (g)-(i) shows histograms of normalized angle sums obtained from manifold-sampled triangles. It does not depend on the graphs analyzed in the previous plots but only on the manifold samples.

A Word on Uniform Sampling Ideally, to follow (Krioukov et al., 2010) and to recover their results, we would have to sample uniformly from geodesic balls of some fixed radius. However, this is non-trivial for arbitrary Riemannian manifolds and, in particular, to the best of our knowledge, for the SPD manifold. One would have to sample from the corresponding Riemannian measure while enforcing the maximum distance constraint given by the geodesic ball. More precisely, using the following formula for the measure $d\mu_g$ in terms of the Riemannian metric $g(x)$, expressed in a normal coordinates system $x$ (see, e.g. Pennec, 2006)

$$d\mu_g = \sqrt{\text{det} g(x)} \, d^n x,$$

one can sample uniformly with, for instance, a rejection sampling algorithm, as long as the right hand side can be computed and the parametrization allows enforcing the support constraint. This is possible for hyperbolic and elliptical spaces through their polar parametrizations. In spite of our efforts, we have not managed to devise a similar procedure for the SPD manifold.

Instead, we have chosen a non-uniform sampling approach, applied consistently, which first generates uniform tangent vectors from a ball in the tangent space at some point and then maps them onto the manifold via the exponential map. Note that the manifold homogeneity guarantees that we get the same results irrespective of the chosen base point, so in our experiments we sample around the identity matrix.
Figure 11. Visualization of the bias incurred when sampling through the exponential map. For the 3-dimensional Poincaré ball (left), uniform samples (from a geodesic ball of radius 5) are more concentrated towards the boundary. The bias is even more apparent for the 2-dimensional sphere (right), where sampling through the exponential map at the south pole, in a ball of radius \( \pi \), yields too many samples around the north pole. That is because the curvature is not taken into account.

Figure 12. Analysis of graphs sampled from geodesic balls in \( \mathbb{R}^n \), \( \mathbb{H}(n) \), and \( S^+(n) \). The plots follow the same organization as in Figure 10. The points are obtained by sampling tangent vectors from a ball of radius 5, uniformly, and mapping them onto the manifold through the exponential map. A few points are missing from the second set of plots because the largest connected components in the corresponding graphs have too few nodes (less than 100).
The difference from the uniform distribution is visualized in Figure 11 for $H(3)$ and $S(2)$.

With that, we proceed to discussing the results from Figure 12. They are organized as in the previous subsection. One difference, besides the sampling procedure, is that we now vary the distance threshold used to link nodes in the graphs from $R/10$ to $1.5R$, where $R$ is the radius of the geodesic balls (i.e., half of the maximum distance between any two points). It is identically $5$ in all scenarios.

**Degree Distributions.** The first obvious characteristic is that the Euclidean space now bounds from above the two non-flat manifolds (that is, their corresponding median curves). Contrast this with the compact manifolds (Figure 10). It tells us that points are in general farther away than the others (relative to the Euclidean space), which is expected, given their (partly) negative curvature. Moreover, recall that the SPD manifold is a higher-rank symmetric space, which means that there are tangent space subspaces with dimension greater than 1 on which the sectional curvature is 0. In light of this property, its interpolating behavior (of Euclidean and Hyperbolic curves), which is already apparent from the evolution of the degree distribution, is not surprising. Note that in 10 dimensions, the graphs constructed from the hyperbolic space are poorly connected even for the higher end of the threshold range. A similar trend characterizes the Euclidean and SPD spaces, but at a lower rate. In the limit $n \to \infty$, essentially all the mass is at the boundary of the geodesic ball for all of them, but how quickly this happens is what distinguishes them.

**Graph Sectional Curvatures.** Here, the difference between the three spaces is hardly noticeable for $n = 3$, in Figure 12d, but it becomes clearer in higher dimensions. We see that as long as the graphs are below a certain median degree, the estimated discrete curvatures are mostly negative, as one would expect. Furthermore, the order of the empirical distributions seems to match the intuition built in the previous paragraph: the hyperbolic space is “more negatively curved” than the SPD manifold.

**Triangles Thickness.** The last row in Figure 12 serves as additional evidence that the SPD manifold is not “as negatively curved” as the hyperbolic space. Without any further comments, we need only emphasize that, as the dimension increases, almost all hyperbolic triangles are “ideal triangles”, i.e., the largest possible triangles in hyperbolic geometry, with the sum-of-angles equal to 0. The analogous histograms for the SPD manifold shift and peak slightly to the left but at a much slower pace.
F. Experimental Setup

The training procedure is the same across all the experiments from Section 5. The structure of the input graphs is not altered. We compute all-pairs shortest-paths in each one of the input graphs and serialize them to disk. They can then be reused throughout our experiments. Since we are not interested in representing the distances exactly (in absolute value) but only relatively, we max-scale them. This has the (empirically observed) advantage of making learning less sensitive to the scaling factors (see below).

The Task  We emphasize that we consider the graph reconstruction task in Section 5. Hence, all results correspond to embedding the input graphs containing all information: no edges or nodes are left out. In other words, we do not evaluate generalization. That is why for each embedding space and fixed dimension, we report the best performing embedding across all runs and loss functions. This is consistent with our goal of decoupling learning objectives and evaluation metrics (see Section 4). We have chosen to restrict our focus as such in order to have fewer factors in our ablation studies. As future work, it is natural to extend our work to downstream tasks and generalization-based scenarios, and study the properties of the introduced matrix manifolds in those settings.

Training  We train for a maximum of 3000 epochs with batches of 512 nodes – that is, 130816 pairwise distances. We use the burn-in strategy from (Nickel & Kiela, 2017; Ganea et al., 2018): training with a 10 times smaller learning rate for the first 10 epochs. Moreover, if the per-epoch loss does not improve for more than 50 epochs, we decrease the learning rate by a factor of 10; the training is ended earlier if this makes the learning rate smaller than $10^{-5}$. This saves time without affecting performance.

Optimization  We repeat the training, as described so far, for three optimization settings. The first one uses RADAM (Becigneul & Ganea, 2019) to learn the embeddings, which we have seen to be the most consistent across our early experiments. In the other two, we train the embeddings using RSGD and RADAM, respectively, but we also train a scaling factor of pairwise distances (with the same optimizer). This is inspired by (Gu et al., 2018). The idea is that scaling the distance function is equivalent to representing the points on a more or less curved sphere or hyperboloid. In the spirit of Riemannian SNE (Section 4.1), this can also be seen as controlling how peaked around the MAP configuration the resulting distribution should be. We have chosen to optimize without scaling in the first setting because it seemed that even for simple, synthetic examples, jointly learning the scaling factor is challenging.

We have also experimented with an optimization inspired by deterministic annealing (Rose, 1998): starting with a high “temperature” and progressively cooling it down. Since we did not see significant improvements, we did not systematically employ this approach.

Computing Infrastructure  We used 4 NVIDIA GeForce GTX 1080 Ti GPUs for the data-parallelizable parts.
G. Input Graphs – Properties & Visualizations

First of all, we include a visualization of each one of the graph used throughout our graph reconstruction experiments in Figure 14. In the rest of this section, we look at some of their geometric properties, summarized in Figure 13. We refer the reader to Appendix C for background on the quantities discussed next.

We start with node degree distributions. Judging by the positions of the means (red diamonds) relative to those of the medians (orange lines), we conclude that most degree distributions are long-tailed, a feature of complex networks. This is particularly severe for “web-edu”, where the mean degree is larger than the 90th percentile. Looking at Figure 6a, it becomes clear what makes it so: there are a few nodes (i.e., web pages) to which almost all the others are connected.

Next, we turn to the estimated sectional curvatures. They indicate a preference for the negative half of the range, reinforcing the complex network resemblance. We see that many of the distributions resemble those of the random graphs sampled from hyperbolic and SPD manifolds (Figures 12e and 12f).

Finally, with one exception, the graphs are close to 0 hyperbolicity, which is an additional indicator of a preference for negative curvature; but how much negative curvature is beneficial remains, a-priori, unclear. The exception is “road-minnesota” which, as its name implies, is a road network and, unsurprisingly, has a different geometry than the others. Its \( \delta \)-hyperbolicity values are mostly large, an indicator of thick triangles and, hence, positive curvature, as discussed in Appendix C.

| Graph         | \( \delta \)-mean | \( \delta \)-max |
|---------------|-------------------|-----------------|
| bio-disome    | 0.20              | 2.50            |
| bio-wormnet   | 0.24              | 2.00            |
| california    | 0.31              | 2.50            |
| csphd         | 0.51              | 6.50            |
| facebook      | 0.10              | 1.50            |
| grqc          | 0.38              | 3.00            |
| power         | 0.96              | 10.00           |
| road-minnesota| 3.13              | 25.00           |
| web-edu       | 0.02              | 1.00            |

Figure 13. The node degrees, (graph) sectional curvatures, and \( \delta \)-hyperbolicities (including the mean of the empirical distribution sampled as shown in (Cohen et al., 2015)) of the input graphs used throughout our experiments with non-positively curved manifolds. Shortened dataset names are used to save space.
(a) “bio-diseasome” (|V| = 516, |E| = 1188) – disorder-gene associations (Goh et al., 2007)

(b) “bio-wormnet” (|V| = 2274, |E| = 78328) – worms gene network (Cho et al., 2014)

(c) “csphd” (|V| = 1025, |E| = 1043) – scientific collaboration network (De Nooy et al., 2018)

(d) “grqc” (|V| = 4158, |E| = 13422) – General Relativity collaboration network (Leskovec & Krevl, 2014)

(e) “power” (|V| = 4941, |E| = 6594) – grid distribution network (Watts & Strogatz, 1998)

(f) “road-minnesota” (|V| = 2640, |E| = 3302) – road network in Minnesota, US (Rossi & Ahmed, 2015)

(g) “web-edu” (|V| = 3031, |E| = 6474) – web network from the .edu domain (Gleich et al., 2004)

(h) “california” (|V| = 5925, |E| = 15770) – (Rossi & Ahmed, 2015)

Figure 14. Visualizations of embedded graphs. The edge color depicts the corresponding Ollivier-Ricci curvature, as described in Appendix C. Similarly, each node is colored according to the average curvature of its adjacent edges. Thus, red nodes behave more like backbone nodes, while blue nodes are either leaves or are part of a clique. In each plot, the size of each node is proportional to its degree.
H. Extended Graph Reconstruction Results

$\mathcal{S}^+ \times \mathbb{H}$ All graph reconstruction results comparing the SPD and hyperbolic manifolds are shown in Tables 5 to 7.

$Gr \times S$ All graph reconstruction results comparing the Grassmann and spherical manifolds are shown in Table 9.

Cartesian Products Graph reconstruction results comparing several product manifolds are shown in Table 8.

Table 5. All graph reconstruction results for “$\mathcal{S}^+$ vs. $\mathbb{H}$”. Compared to Table 2 from the main text, this table includes, on one hand, more graphs and, at the same time, the performance results on 10-dimensional manifolds for most datasets. Notice, in particular, that the results for two much larger graphs are included (Leskovec & Krevl, 2014): “cit-dblp” and “condmat”. They are both citation networks with about 12000 and 23000 nodes, respectively. The columns are the same as in Table 2.

| Graph  | Dim | Manifold | F1@1 | AUC  | Avg. Dist. |
|--------|-----|----------|-------|------|------------|
| bio-diseasome | 3 | Euc | 83.78 | 91.21 | 0.145 |
|        |     | Hyp | 86.21 | 95.72 | 0.137 |
|        |     | SPD | 83.99 | 91.32 | 0.140 |
|        |     | Stein | 86.70 | 94.54 | 0.105 |
|        | 6 | Euc | 93.48 | 95.84 | 0.073 |
|        |     | Hyp | 96.50 | 98.42 | 0.071 |
|        |     | SPD | 93.83 | 95.93 | 0.072 |
|        |     | Stein | 94.86 | 97.64 | 0.066 |
| condmat | 3 | Euc | 16.01 | 74.62 | 0.303 |
|        |     | Hyp | 21.67 | 84.14 | 0.186 |
|        |     | SPD | 16.63 | 80.66 | 0.303 |
|        | 6 | Euc | 31.53 | 81.01 | 0.242 |
|        |     | Hyp | 47.53 | 87.58 | 0.180 |
|        |     | SPD | 32.31 | 83.04 | 0.200 |
|        | 10 | Euc | 52.12 | 89.36 | 0.123 |
|        |     | Hyp | 55.55 | 92.46 | 0.124 |
|        |     | SPD | 52.34 | 89.71 | 0.127 |
|        |     | Stein | 54.45 | 91.61 | 0.098 |
| csphd   | 3 | Euc | 52.12 | 89.36 | 0.123 |
|        |     | Hyp | 55.55 | 92.46 | 0.124 |
|        |     | SPD | 52.34 | 89.71 | 0.127 |
|        | 6 | Euc | 60.63 | 93.91 | 0.065 |
|        |     | Hyp | 64.89 | 94.73 | 0.065 |
|        |     | SPD | 60.59 | 94.12 | 0.066 |
|        |     | Stein | 62.73 | 94.95 | 0.062 |
|        | 10 | Euc | 66.66 | 96.06 | 0.050 |
|        |     | Hyp | 74.76 | 96.38 | 0.045 |
|        |     | SPD | 67.46 | 96.34 | 0.048 |
|        |     | Stein | 70.53 | 96.22 | 0.050 |
| facebook | 3 | Euc | 70.28 | 95.27 | 0.193 |
|        |     | Hyp | 71.08 | 95.46 | 0.173 |
|        |     | SPD | 71.09 | 95.26 | 0.170 |
|        | 6 | Euc | 79.60 | 96.41 | 0.090 |
|        |     | Hyp | 81.83 | 96.53 | 0.089 |
|        |     | SPD | 79.52 | 96.37 | 0.090 |
|        |     | Stein | 83.95 | 96.74 | 0.061 |
|        | 10 | Euc | 85.03 | 97.99 | 0.054 |
|        |     | Hyp | 86.93 | 97.28 | 0.053 |
|        |     | SPD | 85.25 | 97.98 | 0.049 |
|        |     | Stein | 89.25 | 97.82 | 0.044 |
| grqc    | 3 | Euc | 49.61 | 79.99 | 0.212 |
|        |     | Hyp | 66.54 | 87.34 | 0.108 |
|        |     | SPD | 50.41 | 80.48 | 0.208 |
|        |     | Stein | 57.26 | 85.20 | 0.115 |
|        | 6 | Euc | 71.71 | 86.89 | 0.125 |
|        |     | Hyp | 82.43 | 91.53 | 0.091 |
|        |     | SPD | 72.06 | 88.60 | 0.125 |
|        |     | Stein | 78.00 | 90.20 | 0.094 |
|        | 10 | Euc | 83.97 | 91.28 | 0.090 |
|        |     | Hyp | 89.33 | 94.19 | 0.077 |
|        |     | SPD | 84.74 | 93.48 | 0.081 |
|        |     | Stein | 87.62 | 93.19 | 0.081 |
|        | 3 | Euc | 49.34 | 87.84 | 0.119 |
|        |     | Hyp | 60.18 | 91.28 | 0.068 |
|        |     | SPD | 52.48 | 90.17 | 0.121 |
|        |     | Stein | 54.06 | 90.16 | 0.076 |
|        | 6 | Euc | 63.62 | 92.09 | 0.061 |
|        |     | Hyp | 75.02 | 94.34 | 0.060 |
|        |     | SPD | 67.69 | 91.76 | 0.062 |
|        |     | Stein | 70.70 | 93.32 | 0.049 |
|        | 10 | Euc | 74.14 | 94.35 | 0.042 |
|        |     | Hyp | 84.77 | 96.25 | 0.038 |
|        |     | SPD | 79.36 | 95.59 | 0.033 |
|        |     | Stein | 78.13 | 94.67 | 0.040 |
### Table 7. Continuation of Table 6.

| Graph   | Dim | Manifold | F1@1 | AUC  | Avg. Dist. |
|---------|-----|----------|------|------|------------|
| road-minnesota | 3   | Euc      | 90.35| 95.94| 0.058      |
|         |     | Hyp      | 90.02| 95.82| 0.058      |
|         |     | SPD      | 89.71| 95.94| 0.058      |
|         |     | Stein    | 91.96| 96.02| 0.060      |
|         | 6   | Euc      | 96.68| 97.21| 0.045      |
|         |     | Hyp      | 96.59| 97.17| 0.046      |
|         |     | SPD      | 96.81| 97.20| 0.045      |
|         |     | Stein    | 96.21| 97.20| 0.046      |
| web-edu | 3   | Euc      | 29.18| 87.14| 0.245      |
|         |     | Hyp      | 55.60| 92.10| 0.245      |
|         |     | SPD      | 29.02| 88.54| 0.246      |
|         |     | Stein    | 48.28| 90.87| 0.084      |
|         | 6   | Euc      | 49.31| 91.19| 0.143      |
|         |     | Hyp      | 66.23| 95.78| 0.143      |
|         |     | SPD      | 42.16| 91.90| 0.142      |
|         |     | Stein    | 62.81| 96.51| 0.043      |
|         | 10  | Euc      | 42.47| 93.31| 0.082      |
|         |     | Hyp      | 98.43| 98.18| 0.073      |
|         |     | SPD      | 88.30| 96.86| 0.045      |
|         |     | Stein    | 91.02| 98.24| 0.037      |

Table 8. Graph reconstruction results for six Cartesian products of Riemannian manifolds. The three datasets were chosen as some of the most challenging based on the previous results. The SPD embeddings are trained using the Stein divergence. All six product manifolds have 12 free parameters.

| Graph   | Manifold | F1@1 | AUC  | Avg. Dist. |
|---------|----------|------|------|------------|
| california | $\mathbb{H}^4(3)$ | 55.15| 91.35| 0.096      |
|         | $\mathbb{H}^2(6)$ | 56.93| 91.55| 0.096      |
|         | $\mathbb{S}^2 \times \mathbb{S}(6)$ | 55.12| 91.15| 0.096      |
|         | $\mathbb{S}^2 \times \mathbb{S}(6)$ | 49.49| 90.77| 0.087      |
|         | $\mathbb{S}^2 \times \mathbb{S}(6)$ | 50.78| 90.82| 0.086      |
|         | $\mathbb{S}^2 \times \mathbb{S}(6)$ | 47.21| 90.53| 0.089      |
| gqge     | $\mathbb{H}^4(3)$ | 92.03| 95.06| 0.081      |
|         | $\mathbb{H}^2(6)$ | 91.14| 94.97| 0.080      |
|         | $\mathbb{S}^2 \times \mathbb{S}(6)$ | 91.82| 94.44| 0.081      |
|         | $\mathbb{S}^2 \times \mathbb{S}(6)$ | 89.39| 94.25| 0.080      |
|         | $\mathbb{S}^2 \times \mathbb{S}(6)$ | 89.29| 94.20| 0.076      |
|         | $\mathbb{S}^2 \times \mathbb{S}(6)$ | 89.54| 93.79| 0.081      |
| web-edu  | $\mathbb{H}^4(3)$ | 99.26| 98.54| 0.070      |
|         | $\mathbb{H}^2(6)$ | 98.62| 99.24| 0.071      |
|         | $\mathbb{S}^2 \times \mathbb{S}(6)$ | 99.14| 98.39| 0.071      |
|         | $\mathbb{S}^2 \times \mathbb{S}(6)$ | 71.90| 96.78| 0.027      |
|         | $\mathbb{S}^2 \times \mathbb{S}(6)$ | 78.47| 97.12| 0.027      |
|         | $\mathbb{S}^2 \times \mathbb{S}(6)$ | 69.76| 96.24| 0.073      |

### Table 9. All graph reconstruction results for “Gr vs. S”. It includes two 3D models from the Stanford 3D Scanning Repository (Levoy et al., 2005): the notorious “Stanford bunny” and a “drill shaft” (the mesh of a drill bit).

| Graph   | Dim | Manifold | F1@1 | AUC  | Avg. Dist. |
|---------|-----|----------|------|------|------------|
| sphere-mesh | 2   | Sphere | 99.99| 98.31| 0.051      |
|         |     | $Gr(1, 3)$ | 97.20| 90.66| 0.148      |
|         | 3   | Sphere | 100.00| 98.76| 0.042      |
|         |     | $Gr(1, 4)$ | 100.00| 98.38| 0.060      |
|         | 4   | Sphere | 100.00| 98.84| 0.041      |
|         |     | $Gr(1, 5)$ | 100.00| 98.69| 0.060      |
|         |     | $Gr(2, 4)$ | 100.00| 98.94| 0.040      |
| bunny   | 2   | Sphere | 88.12| 89.61| 0.146      |
|         |     | $Gr(1, 3)$ | 85.06| 85.95| 0.146      |
|         | 3   | Sphere | 94.96| 96.86| 0.062      |
|         |     | $Gr(1, 4)$ | 95.41| 97.22| 0.062      |
|         | 4   | Sphere | 95.91| 97.53| 0.055      |
|         |     | $Gr(1, 5)$ | 96.03| 97.63| 0.057      |
|         |     | $Gr(2, 4)$ | 95.86| 97.62| 0.058      |
| drill-shaft | 2   | Sphere | 84.67| 96.27| 0.073      |
|         |     | $Gr(1, 3)$ | 83.40| 96.34| 0.074      |
|         | 3   | Sphere | 89.14| 97.85| 0.052      |
|         |     | $Gr(1, 4)$ | 88.89| 97.81| 0.052      |
|         | 4   | Sphere | 92.45| 98.51| 0.043      |
|         |     | $Gr(1, 5)$ | 92.44| 98.51| 0.043      |
|         |     | $Gr(2, 4)$ | 92.77| 98.54| 0.043      |
| road-minnesota | 2   | Sphere | 82.19| 94.02| 0.085      |
|         |     | $Gr(1, 3)$ | 78.91| 94.02| 0.085      |
|         | 3   | Sphere | 89.55| 95.89| 0.059      |
|         |     | $Gr(1, 4)$ | 90.02| 95.88| 0.058      |
|         | 4   | Sphere | 93.65| 96.66| 0.049      |
|         |     | $Gr(1, 5)$ | 93.89| 96.67| 0.049      |
|         |     | $Gr(2, 4)$ | 94.01| 96.66| 0.049      |
| cat-cortex | 2   | Sphere | - | - | 0.255      |
|         |     | $Gr(1, 3)$ | - | - | 0.234      |
|         | 3   | Sphere | - | - | 0.195      |
|         |     | $Gr(1, 4)$ | - | - | 0.168      |
|         | 4   | Sphere | - | - | 0.156      |
|         |     | $Gr(1, 5)$ | - | - | 0.139      |
|         |     | $Gr(2, 4)$ | - | - | 0.129      |