Literature Review (CSCI2952G)

An Overview of Graph Neural Networks, Protein-Protein Interaction, and Manifold Graph Embedding

Charu Narayanan, Daniel Posmik, Minh Le Tran

October 7, 2025

1. Introduction

Blah Blah

2. Using GNNs for Protein-Protein Interaction

Jha et al. (2022)

3. GNN Architectures

4. Manifold Graph Embedding

Manifold estimation is motivated by the manifold hypothesis (Fefferman et al., 2013) which intuitively states that our realized sample (i.e. our data) stems from a lower-dimensional data generating process that is subsequently corrupted by high-dimensional noise (Fefferman et al. (2013), Meng and Eloyan (2021)). We shall refer to the process of reconstructing and estimating this low-dimensional manifold as manifold estimation or manifold embedding (two terms we will use interchangeably). Intuitively, manifold estimation is a generalization of non-linear dimensionality reduction in which we may explicitly preserve geometric properties of our data.

For high-dimensional noisy graph-structured data, manifold estimation is particularly well-suited to well-known links between discrete graphs and the curvature of non-Euclidian embedding spaces (Bronstein et al., 2017). While no choice of embedding space is perfect due to distortion trade-offs, it can be shown that certain "characteristic graphs" correspond to principle curvatures, e.g., tree-like graphs are optimally embedded in hyperbolic spaces while chain graphs are best represented in spherical space (Weber and Nickel (2018), Weber (2020)).

Manifold estimation is a non-trivial learning problem that requires the careful consideration of trade-offs (e.g., distortion and relational context) and parameter estimation. The parameters that need to be estimated are the class of manifold $\mathcal{M}$, choosing mixed effects structure (e.g., clique-level fixed effects), the intrinsic dimension of the manifold (p), and the projection map from the data ($\mathcal{D}$) to the manifold, i.e.

$$\pi : \mathcal{D} \to \mathcal{M}_{\kappa}^{p} \quad where \quad \mathcal{D} = G = \{V, E\}$$

Here, κ is a measure of curvature and our data $\mathcal{D}$ can be expressed as a graph $G = \{V, E\}$. Lubold et al. (2022) present a data-driven, replicable alternative to ex-ante choosing manifold class ($\mathcal{M}$), dimension (p), and curvature (κ). The technique of Lubold et al. (2022) demonstrate how the connection likelihood has an inverse relationship with the distance of the projected points on the latent manifold. Meng and Eloyan (2021) present a regularized principled manifold estimation technique.

Under correct specification of $\pi(\cdot)$ and the manifold parameters, the key is that for any node $X_{ij} \in G = \{V_X, E_X\}$, its projection on the manifold $\pi^{p}_{\mathcal{M}}(X_{i,j})$ of class $\mathcal{M}$ and intrinsic dimension $\mathcal{M}$ gives us valuable information about node importance (e.g., Yamada (2025) use Forman Curvature, Xu (2020) offers a broader discussion). By the above embedding results, highly connected nodes will lie in "flatter" space while more isolated nodes (e.g., the terminal nodes of a tree-graph) will lie in highly curved space. Thus, the inverse of the absolute gradient at the projection $\pi^{p}_{\mathcal{M}}(X_{i,j})$, say $\hat{w}_{ij} := |\nabla\, \pi^{p}_{\mathcal{M}}(X_{i,j})|^{-1}$, will give us insight into how important this node is in the graph. Our goal is to use these inverse weights w_{ij} as an additional feature in protein-protein interaction prediction. These weights w_{ij} lend themselves particularly well to high-dimensional, unsupervised environments since they respect geometric (e.g., curvature, embedding space), topological, and relational contexts of the data. Alternative embedding techniques, e.g., well-established spectral matrix embedding methods, may not readily extend to heterogeneous geometry and high-dimensional settings (for relevant discussions, see Baptista et al. (2023), Rubin-Delanchy (2021)).

While we have described manifold estimation in theory, we have yet to cover practical considerations. Due to time constraints, we want flexible, out-of-the-box embedding algorithms that can handle high-dimensional, heterogeneous graphs. Giovanni et al. (2022) propose "soft manifolds" as a curvature-aware, flexible solution to the heterogeneity problem. In Giovanni et al. (2022) Proposition 4.1, we are given a numerical embedding procedure based on Riemannian stochastic gradient descent (R-SGD) (Bonnabel, 2013). Marinoni et al. (2023) offer a similar implementation. Jyothish and Jannesari (2025) offer insight into how manifold embeddings work within a graph transformer framework and provide algorithmic details.

Appendix A.

Blah Blah

References

Anthony Baptista, Rubén J. Sánchez-García, Anaïs Baudot, and Ginestra Bianconi. Zoo guide to network embedding, 2023. URL https://arxiv.org/abs/2305.03474.

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

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, July 2017. ISSN 1558-0792. doi: 10.1109/msp.2017.2693418. URL http://dx.doi.org/10.1109/MSP.2017.2693418.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis, 2013. URL https://arxiv.org/abs/1310.0425.

Francesco Di Giovanni, Giulia Luise, and Michael Bronstein. Heterogeneous manifolds for curvature-aware graph embedding, 2022. URL https://arxiv.org/abs/2202.01185.

Kanchan Jha, Sriparna Saha, and Hiteshi Singh. Prediction of protein–protein interaction using graph neural networks. *Scientific Reports*, 12(1):8360, 2022. doi: 10.1038/s41598-022-12201-9. URL https://doi.org/10.1038/s41598-022-12201-9.

Ankit Jyothish and Ali Jannesari. Leveraging manifold embeddings for enhanced graph transformer representations and learning, 2025. URL https://arxiv.org/abs/2507.07335.

Shane Lubold, Arun G. Chandrasekhar, and Tyler H. McCormick. Identifying the latent space geometry of network models through analysis of curvature, 2022. URL https://arxiv.org/abs/2012.10559.

Andrea Marinoni, Pietro Lio', Alessandro Barp, Christian Jutten, and Mark Girolami. Improving embedding of graphs with missing data by soft manifolds, 2023. URL https://arxiv.org/abs/2311.17598.

Kun Meng and Ani Eloyan. Principal manifold estimation via model complexity selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):369–394, March 2021. ISSN 1467-9868. doi: 10.1111/rssb.12416. URL http://dx.doi.org/10.1111/rssb.12416.

Patrick Rubin-Delanchy. Manifold structure in graph embeddings, 2021. URL https://arxiv.org/abs/2006.05168.

Melanie Weber. Neighborhood growth determines geometric priors for relational representation learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 2020.

Melanie Weber and Maximilian Nickel. Curvature and representation learning: Identifying embedding spaces for relational data. In *Advances in Neural Information Processing Systems*, volume 31. NeurIPS, 2018.

Mengjia Xu. Understanding graph embedding methods and their applications, 2020. URL https://arxiv.org/abs/2012.08019.

Taiki Yamada. Vertex evaluation of multiplex graphs using forman curvature, 2025. URL https://arxiv.org/abs/2504.17286.