Out-of-sample extension of graph adjacency spectral embedding

Keith Levin [1] **Farbod Roosta-Khorasani** [2 3] **Michael W. Mahoney** [3 4] **Carey E. Priebe** [5]

Abstract

Many popular dimensionality reduction procedures have out-of-sample extensions, which allow a practitioner to apply a learned embedding to observations not seen in the initial training sample. In this work, we consider the problem of obtaining an out-of-sample extension for the adjacency spectral embedding, a procedure for embedding the vertices of a graph into Euclidean space. We present two different approaches to this problem, one based on a least-squares objective and the other based on a maximum-likelihood formulation. We show that if the graph of interest is drawn according to a certain latent position model called a random dot product graph, then both of these out-of-sample extensions estimate the true latent position of the out-of-sample vertex with the same error rate. Further, we prove a central limit theorem for the least-squares-based extension, showing that the estimate is asymptotically normal about the truth in the large-graph limit.

1. Introduction

Given a graph $G = (V, E)$ on n vertices with adjacency matrix $A \in \{0,1\}^{n \times n}$, the problem of graph embedding is to map the vertices of G to some d-dimensional vector space $\mathcal{S}$ in such a way that geometry in $\mathcal{S}$ reflects the topology of G. For example, we may ask that vertices with high conductance in G be assigned to nearby vectors in $\mathcal{S}$. This is a special case of the problem of dimensionality reduction, well-studied in machine learning and related disciplines (van der Maaten et al., 2009). When applied to graph data, each vertex in G is described by an n-dimensional binary vector, namely its corresponding column (or row) in adjacency matrix $A \in \{0,1\}^{n \times n}$, and we wish to associate with each vertex $v \in V$ a lower-dimensional representation, say $x_v \in \mathcal{S}$. The two most commonly-used approaches for graph embeddings are the graph Laplacian embedding and its variants (Belkin & Niyogi, 2003; Coifman & Lafon, 2006) and the adjacency spectral embedding (ASE, Sussman et al., 2012). Both of these embedding procedures produce low-dimensional representations of the vertices in a graph G, and the question of "which embedding is preferable?" is dependent on the downstream task. Indeed, one can show that neither embedding dominates the other for the purposes of vertex classification; see, for example, Section 4.3 of Tang & Priebe (to appear). In addition, the results in Section 4.3 of Tang & Priebe (to appear) suggest that ASE performs better than the Laplacian eigenmaps embedding for graphs that exhibit a core-periphery structure. Such structures are ubiquitous in real networks, such as those arising in social and biological sciences (Jeub et al., 2015; Leskovec et al., 2009).

The ASE and Laplacian embedding differ in that the latter has received far more attention, especially with respect to questions of limit objects (Hein et al., 2005) and out-of-sample extensions (Bengio et al., 2003). The aim of this paper is to establish theoretical foundations for the latter of these two problems in the case of the adjacency spectral embedding.

2. Background and Notation

In the standard out-of-sample (OOS) extension, we are presented with training data $\mathcal{D} = \{z_1, z_2, \ldots, z_n\} \subseteq \mathcal{X}$, where $\mathcal{X}$ is the set of possible observations. The data $\mathcal{D}$ give rise to a symmetric matrix $M = [K(z_i, z_j)] \in \mathbb{R}^{n \times n}$, where $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is a kernel function that measures similarity between elements of $\mathcal{X}$, so that $K(y, z)$ is large if $y, z \in \mathcal{X}$ are similar, and is small otherwise. Suppose that we have computed an embedding of the data $\mathcal{D}$. Let us denote this embedding by $X \in \mathbb{R}^{n \times d}$, so that the embedding of $z_i \in \mathcal{D}$ is given by the i-th row of X. Suppose that we are given an additional observation $z \in \mathcal{X}$, not necessarily included in $\mathcal{D}$, and we wish to embed z under the same scheme as was used to produce X. A naïve approach would be to discard the old embedding X, consider the augmented

[1]Department of Statistics, University of Michigan, USA. [2]School of Mathematics and Physics, University of Queensland, Australia. [3]International Computer Science Institute, Berkeley, USA. [4]Department of Statistics, University of California at Berkeley, USA. [5]Department of Applied Mathematics and Statistics, Johns Hopkins University, USA. Correspondence to: Keith Levin <klevin@umich.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

collection $\mathcal{D} = \mathcal{D} \cup \{z\}$ and construct a new embedding $\tilde{X} \in \mathbb{R}^{(n+1)\times d}$. However, in many applications, it is infeasible to compute this embedding again from scratch, either because of computational constraints or because the similarities $\{K(z_i, z_j) : z_i, z_j \in \mathcal{D}\}$ may no longer be available after X has been computed. Thus, the OOS problem is to embed z using only the available embedding X which was initially learned from $\mathcal{D}$ and the similarities $\{K(z_i, z)\}_{i=1}^n$.

As an example, consider the Laplacian eigenmaps embedding (Belkin & Niyogi, 2003; Belkin et al., 2006). Given a graph $G = (V, E)$ with adjacency matrix $A \in \mathbb{R}^{n\times n}$, the d-dimensional *normalized Laplacian* of G is the matrix $L = D^{-1/2}AD^{-1/2}$, where $D \in \mathbb{R}^{n\times n}$ is the diagonal degree matrix, i.e., $d_{ii} = \sum_j A_{ij}$ is the degree of the vertex i (Luxburg, 2007; Vishnoi, 2013). The d-dimensional *normalized Laplacian eigenmaps embedding* of G is given by the rows of the matrix $U_L \in \mathbb{R}^{n\times d}$, whose columns are the d orthonormal eigenvectors corresponding to the top d eigenvalues of L, excepting the trivial eigenvalue 1. We note that some authors (see, for example, Chung, 1997) use $I - D^{-1/2}AD^{-1/2}$ to be the normalized graph Laplacian, but since this matrix has the same eigenspace as our L, results concerning the eigenvectors of either of these matrices are equivalent. Suppose that a vertex v is added to graph G, to form graph $\tilde{G}$ with adjacency matrix

$$\tilde{A} = \begin{bmatrix} A & \vec{a} \\ \vec{a}^T & 0 \end{bmatrix}, \tag{1}$$

where $\vec{a} \in \{0, 1\}^n$. A naïve approach to embedding $\tilde{G}$ would be to compute the top eigenvectors of the graph Laplacian of $\tilde{G}$ as before. However, the OOS extension problem requires that we only use the information available in U_L and $\vec{a}$ to compute an embedding of the new vertex v.

Bengio et al. (2003) presented out-of-sample extensions for multidimensional scaling (MDS, Torgerson, 1952; Borg & Groenen, 2005), spectral clustering (Weiss, 1999; Ng et al., 2002), Laplacian eigenmaps (Belkin & Niyogi, 2003) and ISOMAP (Tenenbaum et al., 2000). These OOS extensions were based on a least-squares formulation of the embedding problem, arising from the fact that the in-sample embeddings are given by functions of the eigenvalues and eigenfunctions. Trosset & Priebe (2008) considered a different OOS extension for MDS. Rather than following the approach of Bengio et al. (2003), Trosset & Priebe (2008) cast the MDS OOS extension as a simple modification of the in-sample MDS optimization problem.

Let $\{(\lambda_t, v_t)\}_{t=1}^n$ be the eigen-pairs of the matrix M, constructed from some suitably-chosen similarity function, K, defined on pairs of observations in $\mathcal{D} \times \mathcal{D}$. In general, OOS extensions for eigenvector-based embeddings can be derived as in Bengio et al. (2003) as the solution of a least-squares problem

$$\min_{f(x)\in\mathbb{R}^d} \sum_{i=1}^n \left(K(x, x_i) - \frac{1}{n} \sum_{t=1}^d \lambda_t f_t(x_i) f_t(x) \right)^2,$$

where $\{x_i\}_{i=1}^n$ are the in-sample observations, and $f_t(x_i) = [v_t]_i$ is i^{th} component of v_t. Belkin et al. (2006) presented a slightly different approach that incorporates regularization in both the intrinsic geometry of the data distribution and the geometry of the similarity function K. Their approach applies to Laplacian eigenmaps as well as to regularized least squares and SVM. The authors also introduced a Laplacian SVM, in which a Laplacian penalty term is added to the standard SVM objective function. Belkin et al. (2006) showed that all of these embeddings have OOS extensions that arise as the solution of a generalized eigenvalue problem. We refer the interested reader to Levin et al. (2015) for a practical application of this OOS extension. More recent approaches to OOS extension have avoided altogether the need to solve a least squares or eigenvalue problem by, instead, training a neural net to learn the embedding directly (see, for example, Quispe et al., 2016; Jansen et al., 2017).

The only existing work to date on the ASE OOS extension of which we are aware appears in Tang et al. (2013a). The authors considered the OOS extension for ASE applied to *latent position graphs* (see, for example Hoff et al., 2002), in which each vertex is associated with an element of a vector space and edge probabilities are given by a suitably-chosen inner product. The authors introduced a least-squares OOS extension for embeddings of latent position graphs and proved a theorem, analogous to our Theorem 1, for the error of this extension about the true latent position. Theorem 1 simplifies the proof of the result due to Tang et al. (2013a) for the case of random dot product graphs (see Definition 1 below).

Of crucial importance in assessing OOS extensions, but largely missing from the existing literature, is an investigation of how the OOS estimate compares with the in-sample embedding. That is, for an out-of-sample observation $z \in \mathcal{X}$, how well does its OOS embedding $\hat{X}_z \in \mathbb{R}^d$, approximate the embedding that would be obtained by considering the full sample $\mathcal{D} = \mathcal{D} \cup \{z\}$? In this paper, we address this question in the context of the adjacency spectral embedding. In particular, we show in our main results, Theorems 1 and 2, that two different approaches to the ASE OOS extension recover the in-sample embedding at a rate that is, in a certain sense, optimal (see the discussion at the end of Section 4). We conjecture that analogous rate results can be obtained for other OOS extensions such as those presented in Bengio et al. (2003).

2.1. Notation

We pause briefly to establish notational conventions for this paper. For a matrix $B \in \mathbb{R}^{n_1 \times n_2}$, we let $\sigma_i(B)$ denote the i-th singular value of B, so that $\sigma_1(B) \geq \sigma_2(B) \geq \cdots \geq \sigma_k(B) \geq 0$, where $k = \min\{n_1, n_2\}$. For positive integer n, we let $[n] = \{1, 2, \ldots, n\}$. Throughout this paper, n will index the number of vertices in a hollow graph G, the observed data, and we let $c > 0$ denote a positive constant, not depending on n, whose value may change from line to line. For an event E, we let E^c denote its complement. We will say that event E_n, indexed so as to depend on n, occurs *with high probability*, and write E_n w.h.p. , if for some constant $\epsilon > 0$, it holds for all suitably large n that $\Pr[E_n^c] \leq n^{-(1+\epsilon)}$. In this paper, we will show $\Pr[E^c] \leq cn^{-2}$ any time we wish to show that event E occurs with high probability. All our results involve showing that some event E_n occurs w.h.p., and we note that in all such cases, the Borel-Cantelli Lemma implies that with probability 1, the event E_n^c occurs for at most finitely many n. That is, all our finite-sample results can be easily altered to yield corresponding asymptotic results, as well. For a function $f : \mathbb{Z}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and a sequence of random variables $\{Z_n\}$, we will write $Z_n = O(f(n))$ if there exists a constant C and a number n_0 such that $Z_n \leq Cf(n)$ for all $n \geq n_0$, and write $Z_n = O(f(n))$ a.s. if the event $Z_n \leq Cf(n)$ occurs a.s.a.a. For a vector $x \in \mathbb{R}^d$, we use the unadorned norm $\|x\|$ to denote the Euclidean norm of x. For a matrix $M \in \mathbb{R}^{n \times d}$, we use the unadorned norm $\|M\|$ to denote the operator norm

$$\|M\| = \max_{x \in \mathbb{R}^d : \|x\| = 1} \|Mx\|$$

and we use $\|\cdot\|_{2\to\infty}$ to denote the matrix operator norm

$$\|M\|_{2\to\infty} = \max_{x : \|x\|=1} \|Mx\|_\infty = \max_{i \in [n]} \|M_{i,\cdot}\|,$$

which can be proven via the Cauchy-Schwarz inequality (Horn & Johnson, 2013). This latter operator norm will be especially useful for us, in that a bound on $\|M\|_{2\to\infty}$ gives a uniform bound on the rows of matrix M.

2.2. Roadmap

The remainder of this paper is structured as follows. In Section 3, we present two OOS extensions of the ASE. In Section 4, we prove convergence of these two OOS extensions when applied to random dot product graphs. In Section 5, we explore the empirical performance of the two extensions presented in Section 3, and we conclude with a brief discussion in Section 6.

3. Out-of-sample Embedding for ASE

Given a graph G encoded by adjacency matrix $A \in \{0,1\}^{n \times n}$, the adjacency spectral embedding (ASE) produces a d-dimensional embedding of the vertices of G, given by the rows of the n-by-d matrix

$$\hat{X} = U_A S_A^{1/2}, \tag{2}$$

where $U_A \in \mathbb{R}^{n \times d}$ is a matrix with orthonormal columns given by the d eigenvectors corresponding to the top d eigenvalues of A, which we collect in the diagonal matrix $S_A \in \mathbb{R}^{d \times d}$. We note that in general, one would be better-suited to consider the matrix $[A^T A]^{1/2}$, so that all eigenvalues are guaranteed to be nonnegative, but we will see that in the random dot product graph, the model that is the focus of this paper, the top d eigenvalues of A are positive with high probability (see, for example, either Lemma 1 in Athreya et al. (2016) or Observation 2 in Levin et al. (2017), or refer to the technical report, (Levin et al., 2018)).

The random dot product graph (RDPG, Young & Scheinerman, 2007) is an edge-independent random graph model in which the graph structure arises from the geometry of a set of *latent positions*, i.e., vectors associated to the vertices of the graph. As such, the adjacency spectral embedding is particularly well-suited to this model.

Definition 1. (Random Dot Product Graph) *Let F be a distribution on $\mathbb{R}^d$ such that $x^T y \in [0,1]$ whenever $x, y \in \operatorname{supp} F$, and let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. from F. Collect these n random points in the rows of a matrix $X \in \mathbb{R}^{n \times d}$. Suppose that (symmetric) adjacency matrix $A \in \{0,1\}^{n \times n}$ is distributed in such a way that*

$$\Pr[A|X] = \prod_{1 \leq i < j \leq n} (X_i^T X_j)^{A_{ij}} (1 - X_i^T X_j)^{1 - A_{ij}}. \tag{3}$$

When this is the case, we write $(A, X) \sim \mathrm{RDPG}(F, n)$. If G is the random graph corresponding to adjacency matrix A, we say that G is a random dot product graph *with* latent positions $X_1, X_2, \ldots, X_n$*, where X_i is the latent position corresponding to the i-th vertex.*

A number of results exist showing that the adjacency spectral embedding yields consistent estimates of the latent positions in a random dot product graph (Sussman et al., 2012; Tang et al., 2013b) and recovers community structure in the stochastic block model (Lyzinski et al., 2014). We note an inherent nonidentifiability in the random dot product graph, arising from the fact that for any orthogonal matrix $W \in \mathbb{R}^{d \times d}$, the latent positions $X \in \mathbb{R}^{n \times d}$ and $XW \in \mathbb{R}^{d \times d}$ give rise to the same distribution over graphs, since $XX^T = (XW)(XW)^T = \mathbb{E}[A \mid X]$. Owing to this nonidentifiability, we can only hope to recover the latent positions in X up to some orthogonal rotation. The reader may notice that the RDPG as defined has the limitation that it can only capture graphs with positive semi-definite expected value. This limitation can be overcome by extending the RDPG to the *generalized* RDPG (Rubin-Delanchy

et al., 2017). The results stated in the present work can, for the most part, be extended to this generalized model, but we restrict ourselves here to the RDPG as it appears in Definition 1 for the sake of simplicity.

Suppose that, given adjacency matrix A, we compute embedding

$$\hat{X} = [\hat{X}_1 \hat{X}_2 \dots \hat{X}_n]^T,$$

where $\hat{X}_i \in \mathbb{R}^d$ denotes the embedding of the i-th vertex. Now suppose we add a vertex v with latent position $\bar{w} \in \mathbb{R}^d$ to the original graph G, obtaining an augmented graph $\tilde{G} = ([n] \cup \{v\}, E \cup E_v)$, where E_v denotes the set of edges between v and the vertices of G. One would like to embed vertex v according to the same distribution as the original n vertices and obtain an estimate of $\bar{w}$. Let the binary vector $\vec{a} \in \{0,1\}^n$ encode the edges E_v incident upon vertex v, with entries $a_i = (\vec{a})_i \sim \text{Bernoulli}(X_i^T \bar{w})$. The augmented graph $\tilde{G}$ then has the adjacency matrix as in (1). As discussed earlier, the natural approach to embedding vertex v is to simply re-embed the whole matrix $\tilde{G}$ by computing the ASE of $\tilde{A}$. Suppose that we wish to avoid such a computation, for example due to resource constraints. The problem then becomes one of embedding the new vertex v based solely on the information present in $\hat{X}$ and $\vec{a}$. Two natural approaches to such an OOS extension suggest themselves.

3.1. Linear Least Squares OOS Extension

A natural approach to OOS embedding, pursued by, for example, Bengio et al. (2003), is to embed vertex v as the least-squares solution to $\hat{X}w = \vec{a}$. That is, we embed the vertex v as the vector $\hat{w}_{\text{LS}}$ solving

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \left(a_i - \hat{X}_i^T w \right)^2, \tag{4}$$

where a_i denotes the i-th component of the binary vector $\vec{a}$ encoding the edges between v and the original n vertices. We will denote the solution to the least-squares optimization in Equation (4) by $\hat{w}_{\text{LS}}$, and term this the *linear least squares out-of-sample* (LLS OOS) embedding.

3.2. Maximum Likelihood OOS Extension

A more principled approach to OOS extension, but perhaps more involved computationally, is to consider the following maximum-likelihood formulation. The entries of the vector $\vec{a}$ are distributed independently as $a_i \sim \text{Bernoulli}(X_i^T \bar{w})$, where $\bar{w}$ denotes the true latent position of OOS vertex v. Since we do not have access to the latent positions $\{X_i\}_{i=1}^n$, we use instead their estimates $\{\hat{X}_i\}_{i=1}^n$. This yields the following objective:

$$\max_{w \in \mathbb{R}^d} \sum_{i=1}^{n} a_i \log \hat{X}_i^T w + (1 - a_i) \log \left(1 - \hat{X}_i^T w \right). \tag{5}$$

Unfortunately, this optimization problem may fail to achieve its optimum inside the support of F. Indeed, it may not even have a finite solution. Thus, we will instead settle for solving the following constrained modification of Equation (5),

$$\max_{w \in \widehat{\mathcal{T}}_\epsilon} \sum_{i=1}^{n} a_i \log \hat{X}_i^T w + (1 - a_i) \log \left(1 - \hat{X}_i^T w \right), \tag{6}$$

where $\widehat{\mathcal{T}}_\epsilon = \{w \in \mathbb{R}^d : \epsilon \le \hat{X}_i^T w \le 1 - \epsilon, i \in [n]\}$, and $\epsilon > 0$ is a small constant. We note that this is based only on the edges incident on the OOS vertex rather than on the full data $\tilde{A}$, and uses the spectral estimates $\{\hat{X}_i\}_{i=1}^n$ rather than the true latent positions $\{X_i\}_{i=1}^n$. Despite both of these facts, we will term the extension given by Equation (6) as the *maximum-likelihood out-of-sample* (ML OOS) extension, and we will let $\hat{w}_{\text{ML}}$ denote its solution.

4. Main Results

Our main results show that both the linear least-squares and maximum-likelihood OOS extensions in Equations (4) and (6) recover the true latent position $\bar{w}$ of v. Further, both OOS extensions converge to $\bar{w}$ at the same asymptotic rate (i.e., up to a constant) as we would have obtained, had we computed the ASE of $\tilde{A}$ in (1) directly. This rate is given by Lemma 2.5 from Lyzinski et al. (2014), which we state here in a slightly adapted form. The lemma states, in essence, that the ASE recovers the latent positions with error of order $n^{-1/2} \log n$, uniformly over the n vertices. We remind the reader that $\|M\|_{2\to\infty}$ denotes the 2-to-∞ operator norm, $\|M\|_{2\to\infty} = \max_{x:\|x\|=1} \|Mx\|_\infty$.

Lemma 1 (Adapted from Lyzinski et al. (2014), Lemma 2.5)**.** *Let $X = [X_1, X_2, \dots, X_n]^T \in \mathbb{R}^{n \times d}$ be the matrix of latent positions of an RDPG, and let $\hat{X} \in \mathbb{R}^{n \times d}$ denote the matrix of estimated latent positions yielded by ASE as in (2). Then with probability at least $1 - cn^{-2}$, there exists orthogonal matrix $W \in \mathbb{R}^{d \times d}$ such that*

$$\|\hat{X} - XW\|_{2\to\infty} \le \frac{c \log n}{n^{1/2}}.$$

That is, it holds with high probability that for all $i \in [n]$,

$$\|\hat{X}_i - W^T X_i\| \le \frac{c \log n}{n^{1/2}}.$$

In what follows, we let $A \in \{0,1\}^{n \times n}$ denote the random adjacency matrix of an RDPG G, and let $X_1, X_2, \dots, X_n \in \mathbb{R}^d$ denote its latent positions, collected in matrix $X = [X_1, X_2, \dots, X_n]^T \in \mathbb{R}^{n \times d}$. That is, $(A, X) \sim \text{RDPG}(F, n)$. We use $\hat{X} = [\hat{X}_1, \hat{X}_2, \dots, \hat{X}_n]^T \in \mathbb{R}^{n \times d}$ to denote the matrix whose rows are the estimated latent positions, obtained via ASE as in (2). We let $\bar{w}$ denote the true latent position of the OOS vertex v.

Theorem 1. *With notation as above, let $\hat{w}_{\text{LS}}$ denote the least-squares estimate of $\bar{w}$, i.e., the solution to (4). Then there exists an orthogonal matrix $W \in \mathbb{R}^{d\times d}$ such that*

$$\|W\hat{w}_{\text{LS}} - \bar{w}\| \leq cn^{-1/2}\log n \text{ w.h.p.}$$

Proof. The proof of this result relies upon a classic result for solutions of perturbed linear systems to establish that with high probability, $\|W\hat{w}_{\text{LS}} - w_{\text{LS}}\| \leq cn^{-1/2}\log n$, where $W \in \mathbb{R}^{d\times d}$ is the orthogonal matrix guaranteed by Lemma 1 and w_{LS} is the LS estimate based on the true latent positions $\{X_i\}$ rather than on the estimates $\{\hat{X}_i\}$. A basic Hoeffding inequality to show that with high probability, $\|w_{\text{LS}} - \bar{w}\| \leq cn^{-1/2}\log n$, where again $W \in \mathbb{R}^{d\times d}$ is the orthogonal matrix in Lemma 1. A triangle inequality applied to $\|W\hat{w}_{\text{LS}} - \bar{w}\|$ combined with a union bound over the two high-probability events just described yields the result. A detailed version of this proof can be found in the technical report (Levin et al., 2018). □

As mentioned in Section 3, we would like to consider a maximum-likelihood OOS extension based on the likelihood $\hat{\ell}(w) = \sum_{i=1}^{n} a_i \log \hat{X}_i^T w + (1-a_i)\log(1-\hat{X}_i^T w)$. Toward this end, we would ideally like to use the solution to the optimization problem

$$\arg\max_{w\in\mathbb{R}^d} \hat{\ell}(w),$$

but to ensure a sensible solution, we instead consider

$$\hat{w}_{\text{ML}} = \arg\max_{w\in\widehat{\mathcal{T}}_\epsilon} \hat{\ell}(w), \tag{7}$$

where we remind the reader that $\widehat{\mathcal{T}}_\epsilon = \{w \in \mathbb{R}^d : \epsilon \leq \hat{X}_i^T w \leq 1-\epsilon, i = 1,2,\ldots,n\}$. Theorem 2 shows that $\hat{w}_{\text{ML}}$ recovers the true latent position of the OOS vertex, up to rotation, with error decaying at the same rate as that obtained in Theorem 1 for the LS OOS extension.

Theorem 2. *With notation as above, let $\hat{w}_{\text{ML}}$ be the estimate defined in Equation (7), and let $\epsilon > 0$ be such that $x, y \in \operatorname{supp} F$ implies $\epsilon < x^T y < 1-\epsilon$. Denote the true latent position of the OOS vertex v by $\bar{w} \in \operatorname{supp} F$. Then for all n suitably large, there exists an orthogonal matrix $W \in \mathbb{R}^{d\times d}$ such that with high probability,*

$$\|W\hat{w}_{\text{ML}} - \bar{w}\| \leq cn^{-1/2}\log n \text{ w.h.p.},$$

and this matrix W is the same one guaranteed by Lemma 1.

Proof. By a standard argument from convex optimization, alongside the definition of $\widehat{\mathcal{T}}_\epsilon$, one can thow that for suitably large n,

$$\|W\hat{w}_{\text{ML}} - \bar{w}\| \leq \frac{c\|\nabla\hat{\ell}(W^T\bar{w})\|}{n} \text{ w.h.p.}$$

By the triangle inequality one can then show that

$$\|\nabla\hat{\ell}(W^T\bar{w})\| \leq c\sqrt{n}\log n \text{ w.h.p.}$$

A detailed proof can be found in (Levin et al., 2018). □

Remark 1. Given our in-sample embedding $\hat{X}$ and the vector of edge indicators $\vec{a}$, we can think of the OOS extension as an estimate of $\bar{w}$, the latent position of the OOS vertex v. Lemma 1 implies that if we took the naïve approach of applying ASE to the adjacency matrix $\tilde{A}$ in (1), our estimate would have error of order at most $O(n^{-1/2}\log n)$. Theorems 1 and 2 imply that the OOS estimate obtains the *same asymptotic estimation error*, without recomputing the embedding of $\tilde{A}$.

In addition to the bounds in Theorems 1 and 2, we can show that the least-squares OOS extension satisfies a stronger property, namely the following central limit theorem.

Theorem 3. *Let $(A, X) \sim \text{RDPG}(F, n)$ be a d-dimensional RDPG. Let $\bar{w} \in \operatorname{supp} F$ and $\hat{w}_{\text{LS}} \in \mathbb{R}^d$ be, respectively, the latent position and the least-squares embedding from (4) of an OOS vertex v. There exists a sequence of orthogonal $d \times d$ matrices $\{V_n\}_{n=1}^{\infty}$ such that*

$$\sqrt{n}(V_n^T\hat{w}_{\text{LS}} - \bar{w}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{\bar{w}}),$$

where $\Sigma_{\bar{w}} \in \mathbb{R}^{d\times d}$ is given by

$$\Sigma_{\bar{w}} = \Delta^{-1}\mathbb{E}\left[X_1^T\bar{w}(1-X_1^T\bar{w})X_1X_1^T\right]\Delta^{-1}, \tag{8}$$

and $\Delta = \mathbb{E}X_1X_1^T$.

Proof. This theorem follows from an adaptation of Theorem 1 in (Levin et al., 2017) A detailed proof can be found in (Levin et al., 2018). □

If the OOS vertex is distributed according to F, we have the following corollary by integrating $\bar{w}$ with respect to F.

Corollary 1. *Let $(A, X) \sim \text{RDPG}(F, n)$ be a d-dimensional RDPG, and let $\bar{w}$ be distributed according to F, independent of (A, X). Then there exists a sequence of orthogonal $d \times d$ matrices $\{V_n\}_{n=1}^{\infty}$ such that*

$$\sqrt{n}(V_n^T\hat{w}_{\text{LS}} - \bar{w}) \xrightarrow{\mathcal{L}} \int \mathcal{N}(0, \Sigma_w)dF(w),$$

where Σ_w is defined as in Equation (8) above.

We conjecture that a CLT analogous to Theorem 3 holds for the ML OOS extension.

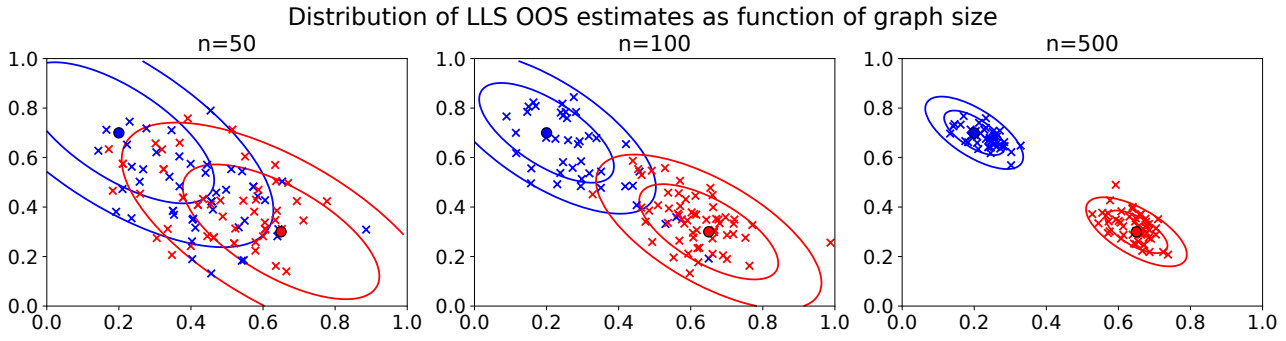

Figure 1. Empirical distribution of the LLS OOS estimate for 100 independent trials for number of vertices $n = 50$ (left), $n = 100$ (middle) and $n = 500$ (right). Each plot shows the positions of 100 independent OOS embeddings, indicated by crosses, and colored according to cluster membership. Contours indicate two generalized standard deviations of the multivariate normal (i.e., 68% and 95% of the probability mass) about the true latent positions, which are indicated by solid circles. We note that even with merely 100 vertices, the normal approximation is already quite reasonable.

5. Experiments

In this section, we briefly explore our results through simulations. We leave a more thorough experimental examination of our results, particularly as they apply to real-world data, for future work. We first give a brief exploration of how quickly the asymptotic distribution in Theorem 3 becomes a good approximation. Toward this end, let us consider a simple mixture of point masses, $F = F_{\lambda,x_1,x_2} = \lambda\delta_{x_1} + (1-\lambda)\delta_{x_2}$, where $x_1, x_2 \in \mathbb{R}^2$ and $\lambda \in (0,1)$. This corresponds to a two-block stochastic block model (Holland et al., 1983), in which the block probability matrix is given by

$$\begin{bmatrix} x_1^T x_1 & x_1^T x_2 \\ x_1^T x_2 & x_2^T x_2 \end{bmatrix}.$$

Corollary 1 implies that if all latent positions (including the OOS vertex) are drawn according to F, then the OOS estimate should be distributed as a mixture of normals centered at x_1 and x_2, with respective mixing coefficients λ and $1-\lambda$.

To assess how well the asymptotic distribution predicted by Theorem 3 and Corollary 1 holds, we generate RDPGs with latent positions drawn i.i.d. from distribution $F = F_{\lambda,x_1,x_2}$ defined above, with

$$\lambda = 0.4,\ x_1 = (0.2, 0.7)^T,\ \text{and}\ x_2 = (0.65, 0.3)^T.$$

For each trial, we draw $n+1$ independent latent positions from F, and generate a binary adjacency matrix from these latent positions. We let the $(n+1)$-th vertex be the OOS vertex. Retaining the subgraph induced by the first n vertices, we obtain an estimate $\hat{X} \in \mathbb{R}^{n\times 2}$ via ASE, from which we obtain an estimate for the OOS vertex via the LS OOS extension as defined in (4). We remind the reader that for each RDPG draw, we initially recover the latent positions only up to a rotation. Thus, for each trial, we compute a Procrustes alignment (Gower & Dijksterhuis, 2004) of the in-sample estimates $\hat{X}$ to their true latent positions. This yields a rotation matrix R, which we apply to the OOS estimate. Thus, the OOS estimates are sensibly comparable across trials. Figure 1 shows the empirical distribution of the OOS embeddings of 100 independent RDPG draws, for $n = 50$ (left), $n = 100$ (center) and $n = 500$ (right) in-sample vertices. Each cross is the location of the OOS estimate for a single draw from the RDPG with latent position distribution F, colored according to true latent position. OOS estimates with true latent position x_1 are plotted as blue crosses, while OOS estimates with true latent position x_2 are plotted as red crosses. The true latent positions x_1 and x_2 are plotted as solid circles, colored accordingly. The plot includes contours for the two normals centered at x_1 and x_2 predicted by Theorem 3 and Corollary 1, with the ellipses indicating the isoclines corresponding to one and two (generalized) standard deviations.

Examining Figure 1, we see that even with only 100 vertices, the mixture of normal distributions predicted by Theorem 3 holds quite well, with the exception of a few gross outliers from the blue cluster. With $n = 500$ vertices, the approximation is particularly good. Indeed, the $n = 500$ case appears to be slightly under-dispersed, possibly due to the Procrustes alignment. It is natural to wonder whether a similarly good fit is exhibited by the ML-based OOS extension. We conjectured at the end of Section 4 that a CLT similar to that in Theorem 3 would also hold for the ML-based OOS extension as defined in Equation (7). Figure 2 shows the empirical distribution of 100 independent OOS estimates, under the same experimental setup as Figure 1, but using the ML OOS extension rather than the linear least-squares extension. The plot supports our conjecture that the ML-based OOS estimates are also approximately normally distributed

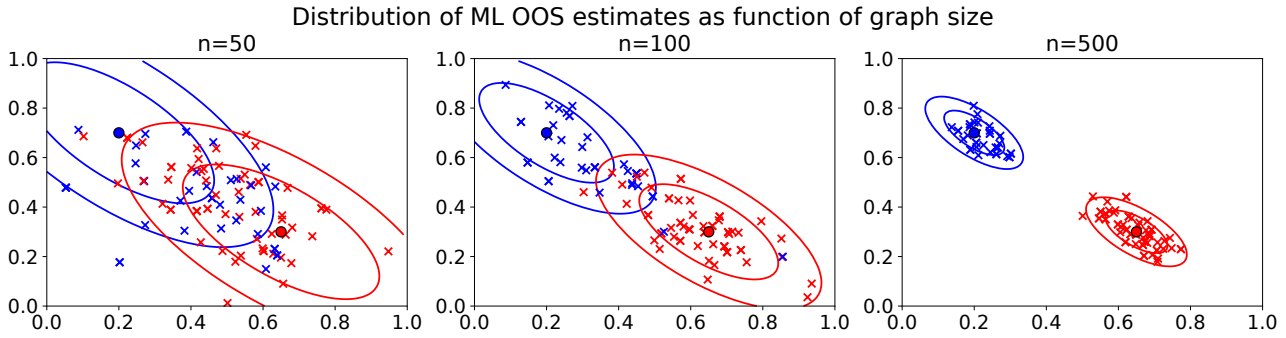

Figure 2. Empirical distribution of the ML OOS estimate for 100 independent trials for number of vertices $n = 50$ (left), $n = 100$ (middle) and $n = 500$ (right). Each plot shows the positions of 100 independent OOS embeddings, indicated by crosses, and colored according to cluster membership. Contours indicate two generalized standard deviations of the multivariate normal (i.e., 68% and 95% of the probability mass) about the true latent positions, which are indicated by solid circles. Once again, even with merely 100 vertices, the normal approximation is already quite reasonable, supporting our conjecture that the ML OOS estimates also distributed as a mixture of normals according to the latent position distribution F.

about the true latent positions.

Figure 1 suggests that we may be confident in applying the large-sample approximation suggested by Theorem 3 and Corollary 1. Applying this approximation allows us to investigate the trade-offs between computational cost and classification accuracy, to which we now turn our attention. The mixture distribution F_{λ,x_1,x_2} above suggests a task in which, given an adjacency matrix A, we wish to classify the vertices according to which of two clusters or communities they belong. That is, we will view two vertices as belonging to the same community if their latent positions are the same (Holland et al., 1983, i.e., the latent positions specify an SBM,). More generally, one may view the task of recovering vertex block memberships in a stochastic block model as a clustering problem. Lyzinski et al. (2014) showed that applying ASE to such a graph, followed by k-means clustering of the estimated latent positions, correctly recovers community memberships of all the vertices (i.e., correctly assigns all vertices to their true latent positions) with high probability.

For concreteness, let us consider a still simpler mixture model, $F = F_{\lambda,p,q} = \lambda\delta_p + (1-\lambda)\delta_q$, where $0 < p < q < 1$, and draw an RDPG $(\tilde{A}, X) \sim \mathrm{RDPG}(F, n+m)$, taking the first n vertices to be in-sample, with induced adjacency matrix $A \in \mathbb{R}^{n\times n}$. That is, we draw the full matrix

$$\tilde{A} = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix},$$

where $C \in \mathbb{R}^{m\times m}$ is the adjacency matrix of the subgraph induced by the m OOS vertices and $B \in \mathbb{R}^{n\times m}$ encodes the edges between the in-sample vertices and the OOS vertices. The latent positions p and q encode a community structure in the graph $\tilde{A}$, and, as alluded to above, a common task in network statistics is to recover this community structure.

Let $\bar{w}^{(1)}, \bar{w}^{(2)}, \ldots, \bar{w}^{(m)} \in \{p, q\}$ denote the true latent positions of the m OOS vertices, with respective least-squares OOS estimates $\hat{w}^{(1)}_{\mathrm{LS}}, \hat{w}^{(2)}_{\mathrm{LS}}, \ldots, \hat{w}^{(m)}_{\mathrm{LS}}$, each obtained from the in-sample ASE $\hat{X} \in \mathbb{R}^n$ of A. We note that one could devise a different OOS embedding procedure that makes use of the subgraph C induced by these m OOS vertices, but we leave the development of such a method to future work. Corollary 1 implies that each $\hat{w}^{(t)}_{\mathrm{LS}}$ for $t \in [m]$ is marginally (approximately) distributed as

$$\hat{w}^{(t)}_{\mathrm{LS}} \sim \lambda\mathcal{N}(p, (n+1)^{-1}\sigma_p^2) + (1-\lambda)\mathcal{N}(q, (n+1)^{-1}\sigma_q^2),$$

where

$$\begin{aligned} \sigma_p^2 &= \Delta^{-2}\left(\lambda p^2(1-p^2)p^2 + (1-\lambda)pq(1-pq)q^2\right), \\ \sigma_q^2 &= \Delta^{-2}\left(\lambda pq(1-pq)p^2 + (1-\lambda)q^2(1-q^2)q^2\right), \\ \text{and } \Delta &= \lambda p^2 + (1-\lambda)q^2. \end{aligned}$$

Classifying the t-th OOS vertex based on $\hat{w}^{(t)}_{\mathrm{LS}}$ via likelihood ratio thus has (approximate) probability of error

$$\begin{aligned} \eta_{n,p,q} = \lambda\Big(1 - \Phi\left(\frac{\sqrt{n+1}(x_{n+1,p,q} - p)}{\sigma_p}\right) \\ + (1-\lambda)\Phi\left(\frac{\sqrt{n+1}(x_{n+1,p,q} - q)}{\sigma_q}\right), \end{aligned}$$

where Φ denotes the cdf of the standard normal and $x_{n,p,q}$ is the value of x solving

$$\begin{aligned} &\lambda\sigma_p^{-1}\exp\{n(x-p)^2/(2\sigma_p^2)\} \\ &\quad = (1-\lambda)\sigma_q^{-1}\exp\{n(x-q)^2/(2\sigma_q^2)\}, \end{aligned}$$

and hence our overall error rate when classifying the m OOS vertices will grow as $m\eta_{n+1,p,q}$.

As discussed previously, the OOS extension allows us to avoid the expense of computing the ASE of the full matrix

$$\tilde{A} = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}.$$

The LLS OOS extension is computationally inexpensive, requiring only the computation of the matrix-vector product $S_A^{-1/2}U_A^T\vec{a}$, with a time complexity $O(d^2n)$ (assuming one does not precompute the product $S_A^{-1/2}U_A^T$). The eigenvalue computation required for embedding $\tilde{A}$ is far more expensive than the LLS OOS extension. Nonetheless, if one were intent on reducing the OOS classification error $\eta_{n+1,p,q}$, one might consider paying the computational expense of embedding $\tilde{A}$ to obtain estimates $\tilde{w}^{(1)}, \tilde{w}^{(2)}, \ldots, \tilde{w}^{(m)}$ of the m OOS vertices. That is, we obtain estimates for the m OOS vertices by making them in-sample vertices, at the expense of solving an eigenproblem on the $(m+n)$-by-$(m+n)$ adjacency matrix. Of course, the entire motivation of our approach is that the in-sample matrix A may not be available. Nonetheless, a comparison against this baseline, in which all data is used to compute our embeddings, is instructive.

Theorem 1 in Athreya et al. (2016) implies that the $\tilde{w}^{(t)}$ estimates based on embedding the full matrix $\tilde{A}$ are (approximately) marginally distributed as

$$\tilde{w}^{(t)} \sim \lambda\mathcal{N}(p, (n+m)^{-1}\sigma_p^2) + (1-\lambda)\mathcal{N}(q, (n+m)^{-1}\sigma_q^2),$$

with classification error

$$\eta_{n+m,p,q} = \lambda\Phi\left(\frac{p - x_{n+m,p,q}}{\sigma_p}\right) + (1-\lambda)\Phi\left(\frac{x_{n+m,p,q} - q}{\sigma_q}\right),$$

where $x_{n+m,p,q}$ is the value of x solving

$$\lambda\sigma_p^{-1}\exp\{(m+n)(x-p)^2/(2\sigma_p^2)\} = (1-\lambda)\sigma_q^{-1}\exp\{(m+n)(x-q)^2/(2\sigma_q^2)\},$$

and it can be checked that $\eta_{n+m,q,p} < \eta_{n,q,p}$ when $m > 1$. Thus, at the cost of computing the ASE of $\tilde{A}$, we may obtain a better estimate. How much does this additional computation improve classification the OOS vertices? Figure 3 explores this question.

Figure 3 compares the error rates of the in-sample and OOS estimates as a function of m and n in the model just described, with $\lambda = 0.4, p = 0.6$ and $q = 0.61$. The plot depicts the ratio of the (approximate) in-sample classification error $\eta_{(n+m),p,q}$ to the (approximate) OOS classification error $\eta_{(n+1),p,q}$, as a function of the number of OOS vertices m, for differently-sized in-sample graphs, $n = 100, 1000,$ and 10000. We see that over several magnitudes of graph size, the in-sample embedding does not improve appreciably over the OOS embedding except when multiple hundreds of OOS vertices are available. When hundreds or thousands of OOS vertices are available simultaneously, we see in the right-hand side of Figure 3 that the in-sample embedding classification error may improve upon the OOS classification error by a large multiplicative factor. Whether or not this improvement is worth the additional computational expense will, depend upon the available resources and desired accuracy, but this suggests that the additional expense associated with performing a second ASE computation is only worthwhile in the event that hundreds or thousands of OOS vertices are available simultaneously. This surfeit of OOS vertices is rather divorced from the typical setting of OOS extension problems, where one typically wishes to embed at most a few previously unseen observations.

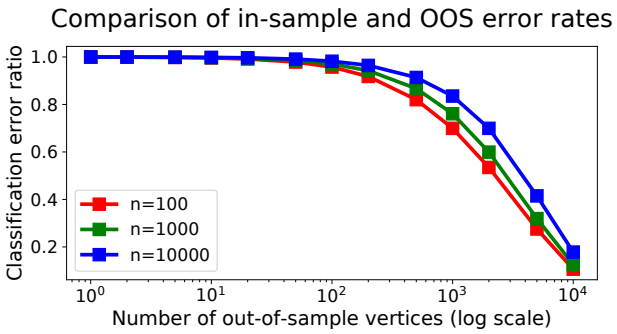

Figure 3. Ratio of the OOS classification error to the in-sample classification error as a function of the number of OOS vertices m, for $n = 100$ vertices, $n = 1000$ vertices and $n = 10000$ vertices. We see that for $m \leq 100$, the expensive in-sample embedding does not improve appreciably on the OOS classification error. However, when many hundreds or thousands of OOS vertices are available simultaneously (i.e., $m \geq 100$), we see that the in-sample embedding may improve upon the OOS estimate by a significant multiplicative factor.

6. Discussion and Conclusion

We have presented a theoretical investigation of two OOS extensions of the ASE, one based on a linear least squares estimate and the other based on a plug-in maximum-likelihood estimate. We have also proven a central limit theorem for the LLS-based extension, and simulation suggests that this CLT is a good approximation even with just a few hundred vertices. We conjecture that a similar CLT holds for the ML-based OOS extension, a conjecture supported by similar simulation data. Finally, we have given a brief illustration of how this OOS extension and the approximation it introduces might be weighed against the computational expense of recomputing a full graph embedding by examining how vertex classification error depends on the size of the set of OOS vertices. We leave a more thorough exploration of this trade-off for future work.

Acknowledgements

The authors would like to thank the reviewers for their helpful comments, which markedly improved the quality of this paper. Keith Levin was partially supported by NSF grant DMS-1646108. Farbod Roosta-Khorasani was partially supported by the Australian Research Council through a Discovery Early Career Researcher Award (DE180100923). Carey E. Priebe was supported by the DARPA D3M program through contract FA8750-17-2-0112. Farbod Roosta-Khorasani and Michael Mahoney also gratefully acknowledge support from DARPA D3M.

References

Athreya, A., Lyzinski, V., Marchette, D. J., Priebe, C. E., Sussman, D. L., and Tang, M. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78:1–18, 2016.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Belkin, M., Niyogi, P., and Sindhwani, V. Manifold Regularization: A Geometric Framework for Learning from Examples. *Journal of Machine Learning Research*, 7: 2399–2434, 2006.

Bengio, Y., Paiement, J., Vincent, P., Delalleau, O., Roux, N. L., and Ouimet, M. Out-of-sample extensions for LLE, ISOMAP, MDS, eigenmaps, and spectral clustering. In *NIPS*, 2003.

Borg, I. and Groenen, P. J. F. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Chung, F. *Spectral Graph Theory*. Number 92 in Conference Board of the Mathematical Sciences Regional Conference Series in Mathematics. American Mathematical Society, 1997.

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21:5–30, 2006.

Gower, J. C. and Dijksterhuis, G. B. *Procrustes Problems*. Number 30 in Oxford Statistical Science Series. Oxford University Press, 2004.

Hein, M., Audibert, J.-Y., and von Luxburg, U. From graphs to manifolds – weak and strong pointwise consistency of graph Laplacians. In *Proceedings of the 18th Annual Conference on Learning Theory*, pp. 470–485, 2005.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

Holland, P. W., Laskey, K., and Leinhardt, S. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

Horn, R. A. and Johnson, C. R. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013.

Jansen, A., Sell, G., and Lyzinski, V. Scalable out-of-sample extension of graph embeddings using deep neural networks. *Pattern Recognition Letters*, 94(15):1–6, 2017.

Jeub, L. G. S., Balachandran, P., Porter, M. A., Mucha, P. J., and Mahoney, M. W. Think locally, act locally: The detection of small, medium-sized, and large communities in large networks. *Physical Review E*, 91(012821), 2015.

Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

Levin, K., Jansen, A., and Van Durme, B. Segmental acoustic indexing for zero resource keyword search. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

Levin, K., Athreya, A., Tang, M., Lyzinski, V., and Priebe, C. E. A central limit theorem for an omnibus embedding of random dot product graphs. *Retrieved from arXiv*, 2017. URL https://arxiv.org/abs/1705.09355.

Levin, K., Roosta-Khorasani, F., Mahoney, M. W., and Priebe, C. E. Out-of-sample extension of graph adjacency spectral embedding. *Retrieved from arXiv*, 2018. URL https://arxiv.org/abs/1802.06307.

Luxburg, U. V. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Lyzinski, V., Sussman, D. L., Tang, M., Athreya, A., and Priebe, C. E. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8(2):2905–2922, 2014.

Ng, A. Y., Jordan, M. I., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In Dietterich, T. G., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems 14*, pp. 849–856. MIT Press, 2002.

Quispe, A. M., Petitjean, C., and Heutte, L. Extreme learning machine for out-of-sample extension in laplacian eigenmaps. *Pattern Recognition Letters*, 74:68–73, 2016.

Rubin-Delanchy, P., Priebe, C. E., and Tang, M. The generalised random dot product graph. *Retrieved from arXiv*, 2017. URL https://arxiv.org/abs/1709.05506.

Sussman, D. L., Tang, M., Fishkind, D. E., and Priebe, C. E. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

Tang, M. and Priebe, C. E. Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *Annals of Statistics*, to appear. URL https://arxiv.org/abs/1607.08601.

Tang, M., Park, Y., and Priebe, C. E. Out-of-sample extension of latent position graphs. *Retrieved from arXiv*, 2013a. URL https://arxiv.org/abs/1305.4893.

Tang, M., Sussman, D. L., and Priebe, C. E. Universally consistent vertex classification for latent position graphs. *The Annals of Statistics*, 31:1406–1430, 2013b.

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

Torgerson, W. S. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

Trosset, M. W. and Priebe, C. E. The out-of-sample problem for classical multidimensional scaling. *Computational statistics and data analysis*, 52(10):4635–4642, 2008.

van der Maaten, L. J. P., Postma, E. O., and van den Herik, H. J. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(1-41):66–71, 2009.

Vishnoi, N. K. Lx = b. *Foundations and Trends in Theoretical Computer Science*, 8(1–2):1–141, 2013.

Weiss, Y. Segmentation using eigenvectors: a unifying view. In *Proc. IEEE International Conference on Computer Vision*, pp. 975–982, 1999.

Young, S. and Scheinerman, E. Random dot product graph models for social networks. In *Proceedings of the 5th International Conference on Algorithms and Models for the Web-graph*, pp. 138–149, 2007.