Consistent Latent Position Estimation and Vertex Classification for Random Dot Product Graphs

Daniel L. Sussman, Minh Tang, and Carey E. Priebe, *Senior Member*, *IEEE*

Abstract—In this work, we show that using the eigen-decomposition of the adjacency matrix, we can consistently estimate latent positions for random dot product graphs provided the latent positions are i.i.d. from some distribution. If class labels are observed for a number of vertices tending to infinity, then we show that the remaining vertices can be classified with error converging to Bayes optimal using the k-nearest-neighbors classification rule. We evaluate the proposed methods on simulated data and a graph derived from Wikipedia.

Index Terms—Random graph, k-nearest-neighbor, latent space model, universal consistency

1 Introduction

This paper is concerned with the semi-supervised setting where a graph, consisting of vertices and edges between vertices, is observed along with class labels for a subset of the graph's vertices. The goal is to build a classifier, using the vertices with known class labels as the training data, to predict the class labels for the remaining vertices. This setting arises in a variety of areas. In social network analysis, we might know of some interesting and uninteresting people and wish to classify the remaining individuals based on their friendships with others. In brain network analysis, we may know the categories for a certain set of neurons and wish to determine the categories for the remaining neurons based on their connectivity patterns. In the analysis of the Internet, we may wish to use the hyper-link graph to classify web pages into categories based on training data.

In this paper, we consider the setting where the graph is distributed according the random dot product graph model [21], [27], an example of a latent space model for random graphs [13]. Latent space models offer a framework in which a graph structure can be parameterized by latent vectors associated with each vertex. Generally, one posits that the adjacency of two vertices is determined by a Bernoulli trial with parameter depending only on the latent positions associated with each vertex; consequently, edges are independent conditioned on the latent positions of the vertices. The value of the latent space approach is that the complexities of the graph structure can be characterized using well-known techniques for vector spaces.

In the case that the latent positions are i.i.d. from some distribution, then the latent space approach is closely related to the theory of exchangeable random graphs [17]. Aldous [2] and Hoover [15] showed that all exchangeable random graphs can be realized by associating each vertex $i \in [n]$ with a latent i.i.d. uniform $[0,1]$ random variable X_i. Conditioned on the $\{X_i\}$, the adjacency of vertices i and j is determined by a Bernoulli trial with parameter $g(X_i, X_j)$, where g is a (symmetric) link function, $g : [0,1]^2 \mapsto [0,1]$. Bickel and Chen [3] provide a concise overview of the details of this framework, and Bickel et al. [4] consider estimation for general exchangeable random graphs using the method of moments.

Taking a more constrained approach, an important example of a latent space model is the stochastic blockmodel [14]. In this model, each latent position can take one of only finitely many distinct values, and two vertices with the same latent position are said to be members of the same block. The block membership of each vertex determines the probabilities of adjacency. Vertices in the same block are said to be stochastically equivalent. This model has been studied extensively, with many efforts focused on unsupervised estimation of vertex block membership. Snijders and Nowicki [24] demonstrated the consistency of an algorithm to estimate block membership in the two block setting. Bickel and Chen [3] showed consistency of a likelihood modularity technique for general stochastic blockmodels in a sparse regime, while Choi et al. [5] demonstrated that maximum likelihood methods are consistent in the case of a growing number of blocks. The value of the stochastic blockmodel is its strong notions of communities and parsimonious structure; however, the assumption of stochastic equivalence may be too strong for many scenarios.

Many latent space approaches seek to generalize the stochastic blockmodel to allow for variation within blocks. For example, the mixed membership model of Airoldi et al. [1] posits that a vertex could have partial membership in multiple blocks. In [12], latent vectors are presumed to be drawn from a mixture of multivariate normal distributions with the adjacency probabilities depending only on the

• *The authors are with The Johns Hopkins University, Whitehead 100, 3400 N. Charles St, Baltimore, MD 21218. E-mail: dsussma3@jhu.edu.*

Manuscript received 19 Aug. 2012; revised 10 Dec. 2012; accepted 24 June 2013; published online 19 July 2013.
Recommended for acceptance by I.S. Dhillon.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2012-08-0641.
Digital Object Identifier no. 10.1109/TPAMI.2013.135.

distance between the latent vectors. For the random dot product model considered here, the probability of adjacency between two vertices is given by the dot product of their latent positions, as described in detail in Section 2. In latent space models, the latent vectors determine the distribution of the random graph; hence, accurate estimates of the latent positions will often lead to accurate subsequent inference.

In this approach, given a latent space model for a graph, we first estimate the latent positions and then use the estimated latent positions to perform subsequent analysis. In this work, we show that under the random dot product graph model, the latent positions can be estimated consistently as the number of vertices in the graph tends to infinity using an eigen-decomposition of the adjacency matrix (see Theorem 4.1). This approach is similar to those in [23], [26], where it is shown that an eigen-decomposition of the normalized Laplacian and the adjacency matrix, respectively, leads to consistent block membership estimation in the stochastic blockmodel. Indeed, the use of spectral methods in graph theory has a long history [7]; specifically, the work of Fiedler [10] demonstrated the importance of the eigenvector corresponding to the second smallest eigenvalue of the graph Laplacian for the min-cut problem. McSherry [20] also considered clustering vertices using an eigen-decomposition of the adjacency matrix.

As discussed above, most previous work in this area considers the unsupervised setting where no class labels are known and the goal is to cluster the vertices. In this work, we consider the vertex classification task in the semi-supervised setting where some vertices have known class labels. Once the latent positions have been estimated, we show that using a k-nearest-neighbor classifier trained on the estimated latent positions yields a universally consistent classifier, i.e., a classifier with error rate that converges to the best possible (Bayes optimal) error rate as the number of vertices with known class labels tends to infinity, regardless of the distribution of the latent positions (see Section 2 and [9]).

The remainder of the paper is structured as follows: Section 2 provides additional background on classical pattern recognition, latent space models, and the random dot product graph. In Section 3, we introduce the basic framework for random dot product graphs and our proposed latent position estimator. In Section 4, we argue that the estimator is consistent, and in Section 5, we show that the k-nearest-neighbors algorithm yields consistent vertex classification. In Section 6, we consider some immediate ways the results presented herein can be extended and discuss some possible implications. Finally, Section 7 provides illustrative examples of applications of this work through simulations and a graph derived from Wikipedia articles and hyper-links.

2 Background

The classical statistical pattern recognition setting involves

$$(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} F_{X,Y},$$

where the $X_i : \Omega \mapsto \mathbb{R}^d$ are observed feature vectors and the $Y_i : \Omega \mapsto \{0, 1\}$ are observed class labels for some probability space Ω. If the joint distribution $F_{X,Y}$ is known, then it is well known that the Bayes classifier given by $h^*(x) = \operatorname{argmax}_{y \in \{0,1\}} \mathbb{P}[Y = y | X = x]$ has optimal performance in the sense of minimum probability of error. This probability of error, which we denote $L^* = \mathbb{P}[h^*(X) \neq Y]$, is called the Bayes optimal probability of error.

Given a training set, $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, the goal is to learn a classifier $h(\cdot; \mathcal{D}) : \mathbb{R}^d \to \{0, 1\}$ such that the probability of error $\mathbb{P}[h(X; \mathcal{D}) \neq Y | \mathcal{D}]$ converges in probability to Bayes optimal as $n \to \infty$ for all distributions $F_{X,Y}$ — this is known as universal consistency [9]. Here we consider the case wherein the feature vectors $X, X_1, \ldots, X_n$ are unobserved, and we observe instead a latent position graph $G(X, X_1, \ldots, X_n)$ on $n + 1$ vertices. In Section 5, we show that a universally consistent classification rule (specifically, k-nearest neighbors) remains universally consistent for this extension of the pattern recognition setup to the random dot product graph model.

The adjacency matrix for a random graph is the symmetric hollow random matrix $\mathbf{A} : \Omega \mapsto \{0, 1\}^{n \times n}$ such that $\mathbf{A}_{ij} = 1$ if and only if the vertices i and j are adjacent. In the latent space approach, one posits that each vertex $i \in [n]$ is associated with a latent position X_i. We assume that the latent positions $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$ for some distribution F on $\mathcal{X} \subset \mathbb{R}^d$. In addition, there is a link function $g : \mathcal{X} \times \mathcal{X} \mapsto [0, 1]$ such that conditioned on the $\{X_i\}$, the upper triangular entries of $\mathbf{A}$ are independent with $\mathbf{A}_{ij} \sim \text{Bern}(g(X_i, X_j))$. Altogether we have that

$$\begin{aligned} X_1, X_2, \ldots\ X_n &\overset{iid}{\sim} F, \quad \text{and} \\ A_{ij} | X_i, X_j &\overset{ind.}{\sim} \text{Bern}(\langle X_i, X_j \rangle), \quad \forall i < j. \end{aligned} \tag{1}$$

Note that since the distribution of the latent vectors and the link function do not depend on n, we are considering only the dense regime in the sense that the number of edges is $\Omega_{\mathbb{P}}(n^2)$.

For random dot product graphs, the link function is given by the dot product, $g(x, y) = \langle x, y \rangle$ and the space $\mathcal{X}$ is restricted so that the dot products are guaranteed to be in $[0, 1]$ (see Section 3). Note that the latent space approach replaces the latent uniform random variables of the exchangeable graph theory with random variables in a more general space $\mathcal{X}$. On one hand, these random graphs still have exchangeable vertices and so could be represented in the i.i.d. uniform framework. On the other hand, d-dimensional latent vectors allow for additional structure and advance interpretation of the latent positions.

In fact, the following result provides a characterization of *finite-dimensional* exchangeable graphs as random dot product graphs. First, we say a link function g is rank $d < \infty$ and positive semidefinite if g can be written as $g(x, y) = \sum_{i=1}^d \psi_i(x)\psi_i(y)$ for some linearly independent functions $\psi_j : [0, 1] \mapsto [-1, 1]$. Using this definition and the inverse probability transform, one can easily show the following:

Proposition 2.1. *An exchangeable random graph has rank $d < \infty$ and positive semidefinite link function if and only if the random graph is distributed according to a random dot product graph with i.i.d. latent vectors in $\mathbb{R}^d$.*

Put in another way, random dot products graphs are exactly the finite-dimensional exchangeable random graphs with positive semidefinite link function, and hence, they

represent a key area for exploration when studying exchangeable random graphs.

In Section 4, we show that if we observe a graph from the random dot product graph model, then we can consistently estimate the original latent positions up to a nonidentifiable orthogonal transformation. Our work relies on techniques developed in [23] and [26] to estimate latent vectors. In particular, Rohe et al. [23] prove that the eigenvectors of the normalized Laplacian can be orthogonally transformed to closely approximate the eigenvectors of the population Laplacian. Their results do not use a specific model but rather rely on assumptions for the Laplacian. Sussman et al. [26] show that for the directed stochastic blockmodel, the eigenvectors/singular vectors of the adjacency matrix can be orthogonally transformed to approximate the eigenvectors/singular vectors of the population adjacency matrix. Fishkind et al. [11] extend these results to the case when the number of blocks in the stochastic blockmodel are unknown. Marchette et al. [19] also use techniques closely related to those presented here to investigate the semi-supervised vertex nomination task.

Finally, another line of work is exemplified by Oliveira [22]. This work shows that, under the independent edge assumption, the adjacency matrix and the normalized Laplacian concentrate around the respective population matrices in the sense of the induced L^2 norm. This work uses techniques from random matrix theory. Other work, such as Chung et al. [6], investigates the spectra of the adjacency and Laplacian matrices for random graphs under a different type of random graph model.

3 Framework

Let $\mathcal{M}_n(A)$ and $\mathcal{M}_{nm}(A)$ denote the set of $n \times n$ and $n \times m$ matrices with values in A for some set A. Additionally, for $\mathbf{M} \in \mathcal{M}_n(\mathbb{R})$, let $\lambda_i(\mathbf{M})$ denote the eigenvalue of $\mathbf{M}$ with the ith largest magnitude. All vectors are column vectors.

Let $\mathcal{X}$ be a subset of the unit ball $\mathcal{B}(0,1) \subset \mathbb{R}^d$ such that $\langle x_1, x_2 \rangle \in [0,1]$, for all $x_1, x_2 \in \mathcal{X}$, where $\langle \cdot, \cdot \rangle$ denotes the standard euclidean inner product. Let F be a probability measure on $\mathcal{X}$ and let $X, X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$. Define $\mathbf{X} := [X_1, X_2, \ldots, X_n]^\top : \Omega \mapsto \mathcal{M}_{n,d}(\mathbb{R})$ and $\mathbf{P} := \mathbf{X}\mathbf{X}^\top : \Omega \mapsto \mathcal{M}_n(\mathbb{R})$.

We assume that the (second moment) matrix $\mathbb{E}[X_1 X_1^\top] \in \mathcal{M}_d(\mathbb{R})$ is rank d and has distinct eigenvalues $\{\lambda_i(\mathbb{E}[XX^\top])\}$. In particular, we suppose there exists $\delta > 0$ such that

$$\begin{aligned} 2\delta &< \min_{i \neq j} |\lambda_i(\mathbb{E}[XX^\top]) - \lambda_j(\mathbb{E}[XX^\top])| \quad \text{and} \\ 2\delta &< \lambda_d(\mathbb{E}[XX^\top]). \end{aligned} \tag{2}$$

Remark 3.1. The distinct eigenvalue assumption is not critical to the results that follow but is assumed for ease of presentation. The theorems hold in the general case with minor changes.

Additionally, we assume that the dimension d of the latent positions is known.

Let $\mathbf{A}$ be a random symmetric hollow matrix such that the entries $\{\mathbf{A}_{ij}\}_{i<j}$ are independent Bernoulli random variables with $\mathbb{P}[\mathbf{A}_{ij} = 1] = \mathbf{P}_{ij}$ for all $i, j \in [n]$, $i < j$. We will refer to $\mathbf{A}$ as the adjacency matrix that corresponds to a graph with vertex set $\{1, \ldots, n\}$. Let $\widetilde{\mathbf{U}}_\mathbf{A}\widetilde{\mathbf{S}}_\mathbf{A}\widetilde{\mathbf{U}}_\mathbf{A}^\top$ be the eigen-decomposition of $|\mathbf{A}|$, where $|\mathbf{A}| = (\mathbf{A}\mathbf{A}^\top)^{1/2}$ with $\widetilde{\mathbf{S}}_\mathbf{A}$ having positive decreasing diagonal entries. Let $\mathbf{U}_\mathbf{A} \in \mathcal{M}_{n,d}(\mathbb{R})$ be given by the first d columns of $\widetilde{\mathbf{U}}_\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$, and let $\mathbf{S}_\mathbf{A} \in \mathcal{M}_d(\mathbb{R})$ be given by the first d rows and columns of $\widetilde{S}_\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$. Let $\mathbf{U}_\mathbf{P}$ and $\mathbf{S}_\mathbf{P}$ be defined similarly.

4 Estimation of Latent Positions

The key result of this section is the following theorem which shows that, using the eigen-decomposition of $|\mathbf{A}|$, we can accurately estimate the true latent positions up to an orthogonal transformation.

Theorem 4.1. *Suppose $\mathbf{A}$ and $\mathbf{X}$ are as in Section 3. Suppose δ is as in (2). With probability greater than $1 - \frac{2(d^2+1)}{n^2}$, there exists an orthogonal matrix $\mathbf{W} \in \mathcal{M}_d(\mathbb{R})$ such that*

$$\left\|\mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}^{1/2}\mathbf{W} - \mathbf{X}\right\|_F \leq 2d\sqrt{\frac{3\log n}{\delta^3}}. \tag{3}$$

Let $\mathbf{W}$ be as above and define $\widehat{\mathbf{X}} = \mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}^{1/2}\mathbf{W}$ with row i denoted by $\widehat{X}_i$. Then, for each $i \in [n]$ and all $\gamma < 1$,

$$\mathbb{P}[\|\widehat{X}_i - X_i\|^2 > n^{-\gamma}] = O(n^{\gamma-1}\log n). \tag{4}$$

We now proceed to prove this result. First, the following result, proven in [26], provides a useful Frobenius bound for the difference between $\mathbf{A}^2$ and $\mathbf{P}^2$.

Proposition 4.2 [26]. *For $\mathbf{A}$ and $\mathbf{P}$ as above, it holds with probability greater than $1 - \frac{2}{n^2}$ that*

$$\|\mathbf{A}^2 - \mathbf{P}^2\|_F \leq \sqrt{3n^3 \log n}. \tag{5}$$

The proof of this proposition is omitted and uses the same Hoeffding bound as is used to prove (8) below.

Proposition 4.3. *For $i \leq d$, it holds with probability greater than $1 - \frac{2d^2}{n^2}$ that*

$$|\lambda_i(\mathbf{P}) - n\lambda_i(\mathbb{E}[XX^\top])| \leq 2d^2\sqrt{n\log n}, \tag{6}$$

and for $i > d$, $\lambda_i(P) = 0$. If (6) holds, then for $i, j \leq d+1$, $i \neq j$, and δ satisfying (2) and n sufficiently large, we have

$$|\lambda_i(\mathbf{P}) - \lambda_j(\mathbf{P})| > \delta n. \tag{7}$$

Proof. First, $\lambda_i(P) = \lambda_i(\mathbf{X}\mathbf{X}^\top) = \lambda_i(\mathbf{X}^\top\mathbf{X})$ for $i \leq d$. Note each entry of $\mathbf{X}^\top\mathbf{X}$ is the sum of n independent random variables each in $[-1, 1]$: $\mathbf{X}^\top\mathbf{X}_{ij} = \sum_{l=1}^n X_{li}X_{lj}$. This means we can apply Hoeffding's inequality to each entry of $\mathbf{X}^\top\mathbf{X} - n\mathbb{E}[XX^\top]$ to obtain

$$\mathbb{P}[|(\mathbf{X}^\top\mathbf{X} - n\mathbb{E}[XX^\top])_{ij}| \geq 2\sqrt{n\log n}] \leq \frac{2}{n^2}. \tag{8}$$

Using a union bound we have that $\mathbb{P}[\|\mathbf{X}^\top\mathbf{X} - \mathbb{E}[XX^\top]\|_F \geq 2d^2\sqrt{n\log n}] \leq \frac{2d^2}{n^2}$. Using Weyl's inequality [16], we have the result.

Equation (7) follows from (6) provided

$$2d^2\sqrt{n\log n} < n\delta,$$

which is the case for n large enough. □

This next lemma shows that we can bound the difference between the eigenvectors of $\mathbf{A}$ and $\mathbf{P}$, while our main results are for scaled versions of the eigenvectors.

Lemma 4.4. *With probability greater than $1 - \frac{2d^2+1}{n^2}$, there exists a choice for the signs of the columns of $\mathbf{U_A}$ such that for each $i \leq d$,*

$$\|(\mathbf{U_A})_{\cdot i} - (\mathbf{U_P})_{\cdot i}\|_F \leq \sqrt{\frac{3\log n}{\delta^2 n}}. \tag{9}$$

Proof. This is a result of applying the Davis-Kahan Theorem [8]; see also [23] to $\mathbf{A}$ and $\mathbf{P}$. Propositions 4.2 and 4.3 give that the eigenvalue gap for $\mathbf{P}^2$ is greater than $\delta^2 n^2$, and that $\|\mathbf{A}^2 - \mathbf{P}^2\|_F \leq \sqrt{3n^3 \log n}$ with probability greater than $1 - \frac{2(d^2+1)}{n^2}$. Apply the Davis-Kahan theorem to each eigenvector of $\mathbf{A}$ and $\mathbf{P}$, which are the same as the eigenvectors of $\mathbf{A}^2$ and $\mathbf{P}^2$, respectively, to get

$$\min_{r_i \in \{-1,1\}} \|(\mathbf{U_A})_{\cdot i} - (\mathbf{U_P})_{\cdot i} r_i\|_F \leq \sqrt{\frac{3\log n}{\delta^2 n}}, \tag{10}$$

for each $i \leq d$. The claim then follows by choosing U_A so that $r_i = 1$ minimizes (10) for each $i \leq d$. □

We now have the ingredients to prove our main theorem.

Proof of Theorem 4.1. The following argument assumes that (9) and (7) hold, which occurs with probability greater than $1 - \frac{2(d^2+1)}{n^2}$. By the triangle inequality, we have

$$\begin{aligned}\left\|\mathbf{U_A}\mathbf{S_A}^{1/2} - \mathbf{U_P}\mathbf{S_P}^{1/2}\right\|_F &\leq \left\|\mathbf{U_A}\mathbf{S_A}^{1/2} - \mathbf{U_A}\mathbf{S_P}^{1/2}\right\|_F \\ &\quad + \left\|\mathbf{U_A}\mathbf{S_P}^{1/2} - \mathbf{U_P}\mathbf{S_P}^{1/2}\right\|_F \\ &= \left\|\mathbf{U_A}(\mathbf{S_A}^{1/2} - \mathbf{S_P}^{1/2})\right\|_F \\ &\quad + \left\|(\mathbf{U_A} - \mathbf{U_P})\mathbf{S_P}^{1/2}\right\|_F.\end{aligned} \tag{11}$$

Note that

$$\begin{aligned}&\lambda_i^{1/2}(|\mathrm{A}|) - \lambda_i^{1/2}(\mathbf{P}) \\ &\quad = \frac{\lambda_i^2(|\mathbf{A}|) - \lambda_i^2(\mathbf{P})}{(\lambda_i(|\mathbf{A}|) + \lambda_i(\mathbf{P}))(\lambda_i(|\mathbf{A}|)^{1/2} + \lambda_i(\mathbf{P})^{1/2})},\end{aligned} \tag{12}$$

where the numerator of the right-hand side is less than $\sqrt{3n^3 \log n}$ by Proposition 4.2 and the denominator is greater than $(\delta n)^{3/2}$ by Proposition 4.3. The first term in (11) is thus bounded by $d\sqrt{3\log n/\delta^3}$. For the second term, $(\mathbf{S_P})_{ii} \leq n$ and

$$\|\mathbf{U_A} - \mathbf{U_P}\| \leq d\sqrt{\frac{3\log n}{\delta^2 n}}.$$

We have established that with probability greater than $1 - \frac{2d^2+1}{n^2}$,

$$\left\|\mathbf{U_A}\mathbf{S_A}^{1/2} - \mathbf{U_P}\mathbf{S_P}^{1/2}\right\|_F \leq 2d\sqrt{\frac{3\log n}{\delta^3}}. \tag{13}$$

We now will show that an orthogonal transformation will give us the same bound in terms of $\mathbf{X}$. Let $\mathbf{Y} = \mathbf{U_P}\mathbf{S_P}^{1/2}$. Then $\mathbf{Y}\mathbf{Y}^\top = \mathbf{P} = \mathbf{X}\mathbf{X}^\top$ and thus $\mathbf{Y}\mathbf{Y}^\top\mathbf{X} = \mathbf{X}\mathbf{X}^\top\mathbf{X}$. Because $\mathrm{rank}(\mathbf{P}) = d = \mathrm{rank}(\mathbf{X})$, we have that $\mathbf{X}^\top\mathbf{X}$ is nonsingular and hence $\mathbf{X} = \mathbf{Y}\mathbf{Y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$. Let $\mathbf{W} = \mathbf{Y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$. It is straightforward to verify that $\mathrm{rank}(\mathbf{W}) = d$ and that $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$. $\mathbf{W}$ is thus an orthogonal matrix, and $\mathbf{X} = \mathbf{Y}\mathbf{W} = \mathbf{U_P}\mathbf{S_P}^{1/2}\mathbf{W}$. Equation (3) is thus established.

Now, we will prove (4). Note that because the $\{X_i\}$ are i.i.d., the $\{\widehat{X}_i\}$ are exchangeable and hence identically distributed. As a result, each of the random variables $\|\widehat{X}_i - X_i\|$ are identically distributed. Note that for sufficiently large n, by conditioning on the event in (3), we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{X} - \widehat{\mathbf{X}}\|^2] &\leq \left(1 - \frac{2(d^2+1)}{n^2}\right)(2d)^2\frac{3\log n}{\delta^3} + \frac{2(d^2+1)}{n^2}2n \\ &= O\left(\frac{d^2\log n}{\delta^3}\right),\end{aligned} \tag{14}$$

because the worst case bound is $\|\mathbf{X} - \widehat{\mathbf{X}}\|^2 \leq 2n$ with probability 1. We also have that

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{I}\{\|\widehat{X}_i - X_i\|^2 > n^{-\gamma}\}n^{-\gamma}\right] \leq \mathbb{E}[\|\mathbf{X} - \widehat{\mathbf{X}}\|^2], \tag{15}$$

and because the $\|\widehat{X}_i - X_i\|$ are identically distributed, the left-hand side is simply $n^{1-\gamma}\mathbb{P}[\|\widehat{X}_i - X_i\|^2 > n^{-\gamma}]$. □

5 Consistent Vertex Classification

So far we have shown that using the eigen-decomposition of $|\mathbf{A}|$, we can consistently estimate all latent positions simultaneously (up to an orthogonal transformation). One could imagine that this will lead to accurate inference for various exploitation tasks of interest. For example, Sussman et al. [26] explored the use of this embedding for unsupervised clustering of vertices in the simpler stochastic blockmodel setting. In this section, we will explore the implications of consistent latent position estimation in the supervised classification setting. In particular, we will prove that universally consistent classification using k-nearest-neighbors remains valid when we select the neighbors using the estimated vectors rather than the true but unknown latent positions.

First, let us expand our framework. Let $\mathcal{X} \subset \mathbb{R}^d$ be as in Section 3, and let $F_{X,Y}$ be a distribution on $\mathcal{X} \times \{0,1\}$. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}) \overset{i.i.d.}{\sim} F_{X,Y}$, and let $\mathbf{P} \in \mathcal{M}_{n+1}([0,1])$ and $\mathbf{A} \in \mathcal{M}_{n+1}(\{0,1\})$ be as in Section 3. Here, the Y_is are the class labels for the vertices in the graph corresponding to the adjacency matrix $\mathbf{A}$.

We suppose that we observe only $\mathbf{A}$, the adjacency matrix, and $Y_1, \ldots, Y_n$, the class labels for all but the last vertex. Our goal is to accurately classify this last vertex, so for notational convenience define $X := X_{n+1}$ and $Y := Y_{n+1}$. Let the rows of $\mathbf{U_A}\mathbf{S_A}^{1/2}$ be denoted by $\zeta_1^\top, \ldots, \zeta_{n+1}^\top$. The k-nearest-neighbor rule for k odd is defined as follows: For $1 \leq i \leq n$, let $W_{ni}(X) = 1/k$ only if ζ_i is one of the k nearest points to ζ from among $\{\zeta_i\}_{i=1}^n$; $W_{ni}(X) = 0$ otherwise. (We break ties by selecting the neighbor with the smallest index.)

The k-nearest-neighbor rule is then given by $h_n(x) = \mathbb{I}\{\sum_{i=1}^n W_{ni}(X)Y_i > \frac{1}{2}\}$. It is a well-known theorem of Stone [25] that had we observed the original $\{X_i\}$, the k-nearest neighbor rule using the euclidean distance from $\{X_i\}$ to X is universally consistent provided $k \to \infty$ and $k/n \to 0$. This means that for any distribution $F_{X,Y}$,

$$\begin{aligned}\mathbb{E}[L_n] :&= \mathbb{E}[\mathbb{P}[\widetilde{h}_n(X) \neq Y | (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)]] \\ &\to \mathbb{P}[h^*(X) \neq Y] =: L^*,\end{aligned} \tag{16}$$

as $n \to \infty$, where $\widetilde{h}_n$ is the standard k-nearest-neighbor rule trained on the $\{(X_i, Y_i)\}$ and h^* is the (optimal) Bayes rule. This theorem relies on the following very general result, also of Stone [25]; see also [9, Theorem 6.3].

Theorem 5.1 [25]. *Assume that for any distribution of X, the weights W_{ni} satisfy the following three conditions:*

1. *There exists a constant c such that for every nonnegative measurable function f satisfying $\mathbb{E}[f(X)] < \infty$,*

$$\mathbb{E}\left[\sum_{i=1}^n W_{ni}(X) f(X_i)\right] \leq c\mathbb{E}[f(X)]. \tag{17}$$

2. *For all $a > 0$,*

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^n W_{ni}(X)\mathbb{I}\{\|X_i - X\| > a\}\right] = 0. \tag{18}$$

3.

$$\lim_{n\to\infty} \mathbb{E}\Big[\max_{1\leq i\leq n} W_{ni}(X)\Big] = 0 \tag{19}$$

Then $h_n(x) = \mathbb{I}\{\sum_i W_{ni}(x) > 1/2\}$ is universally consistent.

Remark 5.2. Recall that the $\{\widehat{X}_i\}$ are defined in Theorem 4.1. Because the $\{\widehat{X}_i\}$ are obtained via an orthogonal transformation of the $\{\zeta_i\}$, the nearest neighbors of $\widehat{X} = \widehat{X}_{n+1}$ are the same as those of ζ. As a result of this and the relationship between $\mathbf{X}$ and $\widehat{\mathbf{X}}$, we work using the $\{\widehat{X}_i\}$, even though these cannot be known without some additional knowledge.

To prove that the k-nearest-neighbor rule for the $\{\widehat{X}_i\}$ is universally consistent, we must show that the corresponding W_{ni} satisfy these conditions. The methods to do this are adapted from the proof presented in [9]. We will outline the steps of the proof, but the details follow *mutatis mutandis* from the standard proof.

First, the following lemma is adapted from [9] by using a triangle inequality argument.

Lemma 5.3. *Suppose $k/n \to 0$. If $X \in \mathrm{supp}(F_X)$, then $\|\widehat{X}_{(k)}(\widehat{X}) - \widehat{X}\| \to 0$ almost surely, where $\widehat{X}_{(k)}(\widehat{X})$ is the kth nearest neighbor of $\widehat{X}$ among $\{\widehat{X}_i\}_{i=1}^n$.*

Condition 3 follows immediately from the definition of the W_{ni}. The remainder of the proof follows with few changes after recognizing that the random variables $\{(X, \widehat{X})\}$ are exchangeable. Overall, we have the following universal consistency result.

Theorem 5.4. *If $k \to \infty$ and $k/n \to 0$ as $n \to \infty$, then the $W_{ni}(X)$ satisfy the conditions of Theorem 5.1 and hence $\mathbb{E}[\mathbb{P}[h_n(\widehat{X}) \neq Y | \mathbf{A}, \{Y_i\}_{i=1}^n] = \mathbb{E}[L_n] \to L_X^*$.*

6 Extensions

The results presented thus far are for the specific problem of determining one unobserved class label for a vertex in a random dot product graph. In fact, the techniques used can be extended to somewhat more general settings without significant additional work.

6.1 Classification

For example, the results in Section 5 are stated in the case that we have observed the class labels for all but one vertex. However, the universal consistency of the k-nearest-neighbor classifier remains valid provided the number of vertices m with observed vertex class labels goes to infinity and $k/m \to 0$ as the number of vertices $n \to \infty$. In other words, we may train the k-nearest neighbor on a smaller subset of the estimated latent vectors provided the size of that subset goes to ∞.

On the other hand, if we fix the number of observed class labels m and the classification rule h_m, and let the number of vertices tend to ∞, then we can show the probability of incorrectly classifying a vertex will converge to $L_m = \mathbb{P}[h_m(Z) \neq Y]$. Additionally, our results also hold when the class labels Y can take more than two but still finitely many values.

In fact, the results in Section 5 and (4) from Theorem 4.1 rely only on the fact that the $\{X_i\}$ are i.i.d. and bounded, the $\{(X_i, \widehat{X}_i)\}$ are exchangeable, and $\|\mathbf{X} - \widehat{\mathbf{X}}\|_F^2$ can be bounded with high probability by a $O(\log n)$ function. The random graph structure provided in our framework is of interest, but it is the total noise bounds that are crucial for the universal consistency claim to hold.

6.2 Latent Position Estimation

In Section 4, we state our results for the random dot product graph model. We can generalize our results immediately by replacing the dot product with a bilinear form, $g(x, y) = x^\top(\mathbf{I}_{d'} \oplus (-\mathbf{I}_{d''}))y$, where $\mathbf{I}_d$ is the $d \times d$ identity matrix. This model has the interpretation that similarities in the first d' dimensions increase the probability of adjacency, while similarities in the last d'' reduce the probability of adjacency. All the results remain valid under this model, and in fact, arguments in [22] can be used to show that the signature of the bilinear form can also be estimated consistently. We also recall that the assumption of distinct eigenvalues for $\mathbb{E}[XX^T]$ can be removed with minor changes. Particularly, Lemma 4.4 applies to groups of eigenvalues, and subsequent results can be adapted without changing the order of the bounds.

This work focuses on undirected graphs and this assumption is used explicitly throughout Section 4. The extension to the directed graph setting is also of interest and may be possible, though currently our proofs rely heavily on the symmetry and hence we do not presently investigate this problem. We also note that we assume the graph has no

loops so that $\mathbf{A}$ is hollow. This assumption can be dropped, and in fact, the impact of the diagonal is asymptotically negligible, provided each entry is bounded. Marchette et al. [19] suggest that augmenting the diagonal may improve latent position estimation for finite samples.

In [23], the number of blocks in the stochastic blockmodel, which is related to d in our setting [26], is allowed to grow with n; our work can also be extended to this setting. In this case and in the case of sparse graphs, it will be the interaction between the rate of growth of d and the rate that δ vanishes that controls the bounds in Theorem 4.1. Additionally, the consistency of k-nearest-neighbors when the dimension grows is less well understood and results such as Stone's Theorem 5.1 do not apply.

In addition to keeping d fixed, we also assume that d is known. Fishkind et al. [11] and Sussman et al. [26] suggest consistent methods to estimate the latent space dimension. The results in [22] can also be used to derive thresholds for eigenvalues to estimate d.

Finally, [11] and [19] also consider that the edges may be attributed; for example, if edges represent a communication, then the attributes could represent the topic of the communication. The attributed case can be thought of as a set of adjacency matrices, and we can embed each separately and concatenate the embeddings. Fishkind et al. [11] argue that this method works under the attributed stochastic blockmodel and similar arguments could likely be used to extend the current work.

6.3 Extension to the Laplacian

The eigen-decomposition of the graph Laplacian is also widely used for similar inference tasks. In this section, we argue informally that our results extend to the Laplacian. We will consider a slight modification of the standard normalized Laplacian as defined in [23]. This modification scales the Laplacian in [23] by $n-1$ so that the first d eigenvalues of our matrix are $O(n)$ rather then $O(1)$ for the standard normalized Laplacian.

Let $\mathbf{L} := \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is diagonal with $\mathbf{D}_{ii} := \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{A}_{ij}$. Additionally, let $\mathbf{Q} := \bar{\mathbf{D}}^{-1/2}\mathbf{P}\bar{\mathbf{D}}^{-1/2}$, where $\bar{\mathbf{D}}$ is diagonal with

$$\begin{aligned}\bar{\mathbf{D}}_{ii} :&= \frac{1}{n-1}\mathbb{E}\left[\sum_{j=1}^{n}\mathbf{A}_{ij}|\ \mathbf{X}\right] = \frac{1}{n-1}\sum_{j\neq i}\mathbf{P}_{ij} \\ &= \frac{1}{n-1}\sum_{j\neq i}\langle X_i, X_j\rangle.\end{aligned} \tag{20}$$

Finally, define $q : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}^d$ as $q(x,y) := \frac{x}{\sqrt{\langle x,y\rangle}}$, $Z_i := q(X_i, \frac{1}{n}\sum_{j\neq i}X_j)$, and $\widetilde{Z}_i := q(X_i, \mathbb{E}(X))$.

Because the pairwise dot products of the rows of $\bar{\mathbf{D}}^{-1/2}\mathbf{X}$ are the same as the entries of $\mathbf{Q}$, the scaled eigenvectors of $\mathbf{Q}$ must be an orthogonal transformation of the $\{Z_i\}$. Furthermore, note that for large n, Z_i, and $\widetilde{Z}_i$ will be close with high probability because $\frac{1}{n}\sum_{j\neq i}X_j \xrightarrow{a.s} \mathbb{E}[X]$ and the function $q(X_j,\cdot)$ is smooth almost surely. Additionally, the $\{\widetilde{Z}_i\}$ are i.i.d. and $q(\cdot, \mathbb{E}[X])$ is one-to-one so that the Bayes optimal error rate is the same for the $\{\widetilde{Z}_i\}$ as for the $\{X_i\}$: $L^*_X = L^*_{\widetilde{Z}}$. If the further assumption that the minimum expected degree among all vertices is greater than $\sqrt{2n}/\sqrt{\log n}$ holds, then the assumptions of Theorem 2.2 in [23] are satisfied.

Let $\widehat{Z}_i$ denote the ith row of the matrix $\mathbf{U_L}\mathbf{S_L}$ defined analogously in Section 3, and let $\widetilde{\mathbf{Z}}$ be the matrix with row i given by $\widetilde{Z}_i^\top$. Using the results in [23] and similar tools to those we have used thus far, one can show that $\min_{\mathbf{W}}\|\mathbf{U_L}\mathbf{S_L}\mathbf{W} - \widetilde{\mathbf{Z}}\|^2$ can be bounded with high probability by a function in $O(\log n)$. As discussed above, this is sufficient for k-nearest-neighbors trained on $\{(\widehat{Z}_i, Y_i)\}$ to be universally consistent. In this paper, we do not investigate the comparative values of the eigen-decompositions for the Laplacian versus the adjacency matrix, but one factor may be the properties of the map q defined above as applied to different distributions on $\mathcal{X}$.

7 Experiments

In this section, we present empirical results for a graph derived from Wikipedia links as well as simulations for an example wherein the $\{X_i\}$ arise from a Dirichlet distribution.

7.1 Simulations

To demonstrate our results, we considered a problem where perfect classification is possible. Each $X_i : \Omega \mapsto \mathbb{R}^2$ is distributed according to a Dirichlet distribution with parameter $\alpha = [2,2,2]^\top$ where we keep just the first two coordinates. The class labels are determined by the X_i with $Y_i = \mathbb{I}\{X_{i1} < X_{i2}\}$ so in particular $L^* = 0$.

For each $n \in \{100, 200, \ldots, 2,000\}$, we simulated 500 instances of the $\{X_i\}$ and sample the associated random graph. For each graph, we used our technique to embed each vertex in two dimensions. To facilitate comparisons, we used the matrix $\mathbf{X}$ to construct the matrix $\widehat{\mathbf{X}}$ via transformation by the optimal orthogonal $\mathbf{W}$. Fig. 1 illustrates our embedding for $n = 2{,}000$ with each point corresponding to a row of $\widehat{\mathbf{X}}$ with points colored according the class labels $\{Y_i\}$. To demonstrate our results from Section 4, Fig. 2 shows the average square error in the latent position estimation per vertex.

For each graph, we used leave-one-out cross validation to evaluate the error rate for k-nearest-neighbors for $k = 2\lfloor\sqrt{n}/4\rfloor + 1$. We suppose that we observe all but 1 class label as in Section 5. Fig. 3 shows the classification error rates. The black line shows the classification error when classifying using $\widehat{\mathbf{X}}$ while the red line shows the classification error when classifying using $\mathbf{X}$. Unsurprisingly, classifying using $\widehat{\mathbf{X}}$ gives worse performance. However, we still see steady improvement as the number of vertices increases, as predicted by our universal consistency result. Indeed, this figure suggests that the rates of convergence may be similar for both $\mathbf{X}$ and $\widehat{\mathbf{X}}$.

7.2 Wikipedia Graph

For this data [18], http://www.cis.jhu.edu/~parky/Data/Wiki/, each vertex in the graph corresponds to a Wikipedia page, and the edges correspond to the presence of a hyperlink between two pages (in either direction). We consider this as an undirected graph. Every article within two hyperlinks of the article "Algebraic Geometry" was included as a vertex in the graph. This resulted in $n = 1{,}382$ vertices. Additionally, each document, and hence each vertex, was manually labeled as one of the following: Category (119), Person (372), Location (270), Date (191), and Math (430).

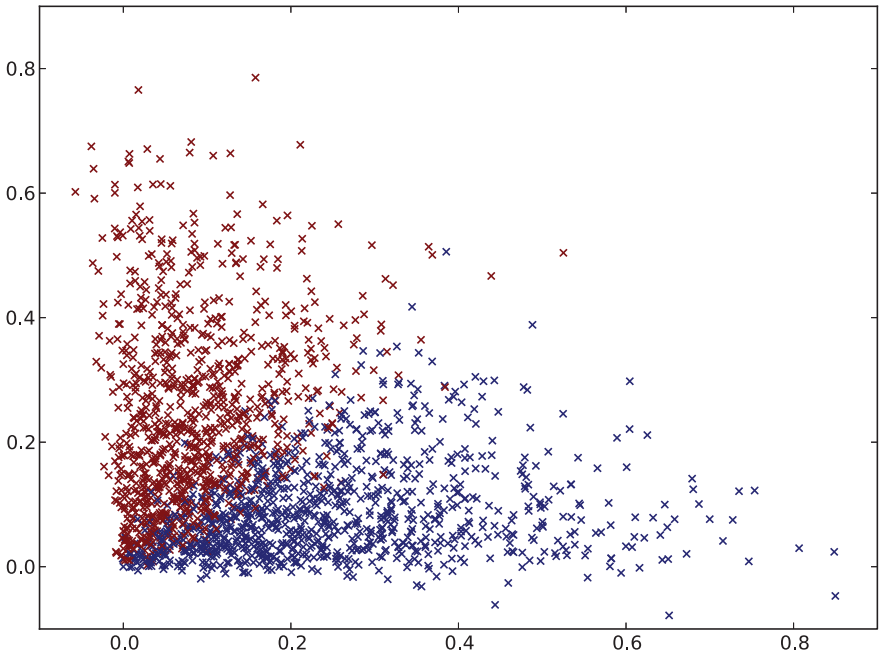


Fig. 1. An example of estimated latent position $\{\widehat{X}_i\}$ for the distribution described in Section 7.1. Each point is colored according to class labels $\{Y_i\}$. For the original latent position $\{X_i\}$, the two classes would be perfectly separated by the line $y = x$. In this figure the two classes are nearly separated but have some overlap. Note also that some estimated positions are outside the support of the original distribution.

To investigate the implications of the results presented thus far, we performed a pair of illustrative investigations. First, we used our technique on random induced subgraphs and used leave-one-out cross validation to estimate error rates for each subgraph. We used $k = 9$ and $d = 10$ and performed 100 Monte Carlo iterates of random induced subgraphs with $n \in \{100, 200, \ldots, 1,300\}$ vertices. Fig. 4 shows the mean classification error estimates using leave-one-out cross validation on each randomly selected subgraph. Note, the chance error rate is $1 - 430/1{,}382 = 0.689$.

We also investigated the performance of our procedure for different choices of d, the embedding dimension, and k, the number of nearest neighbors. Because this data has five classes, we use the standard k-nearest-neighbor algorithm and break ties by choosing the first label as ordered above. Using leave-one-out cross validation, we calculated an estimated error rate for each $d \in \{1, \ldots, 50\}$ and $k \in \{1, 5, 9, 13, 17\}$. The results are shown in Fig. 5. This figure suggests that our technique will be robust to different choices of k and d within some range.

For comparison, we considered an alternative method to classify vertices. For this method, each vertex $i \in [n]$ is associated with a feature vector $W_i \in \mathbb{N}^5$, where $W_{ij} = \#\{\ell : A_{ik} = 1, Y_\ell = j\}$. Leave-one-out cross validation estimates of classification error using k-nearest-neighbors on the W_i with $k \in \{1, 5, \ldots, 17\}$ ranged from 0.39 to 0.43. Using normalized versions of the W_i, leave-one-out error estimates ranged from 0.40 to 0.46. The similar

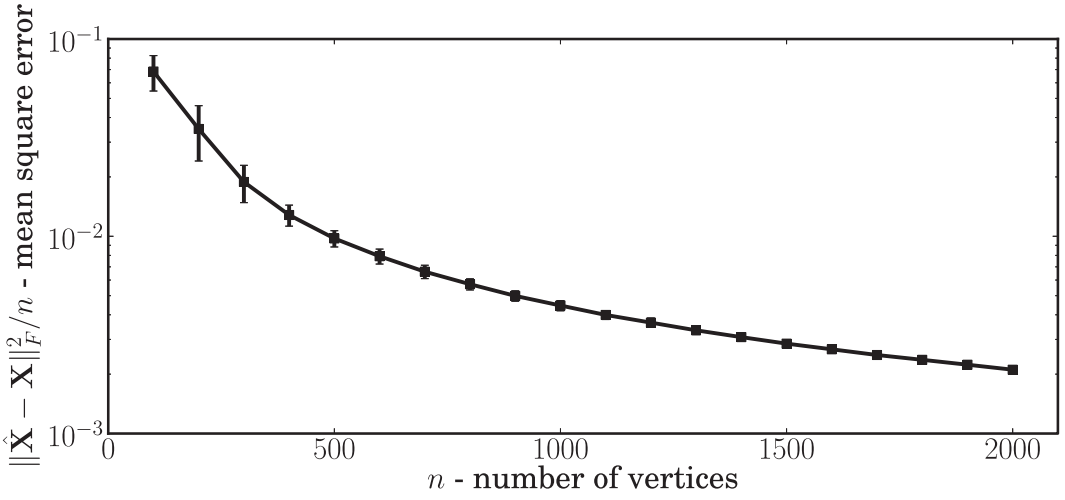


Fig. 2. Mean square error versus number of vertices. This figure shows the mean square error in latent position estimation per vertex, given by $\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2/n$, for the simulation described in Section 7.1. The error bars are given by the standard deviation of the average square error over 500 Monte Carlo replicates for each n. On average, the estimated latent positions converge rapidly to the true latent positions as the number of vertices in the graph increases.

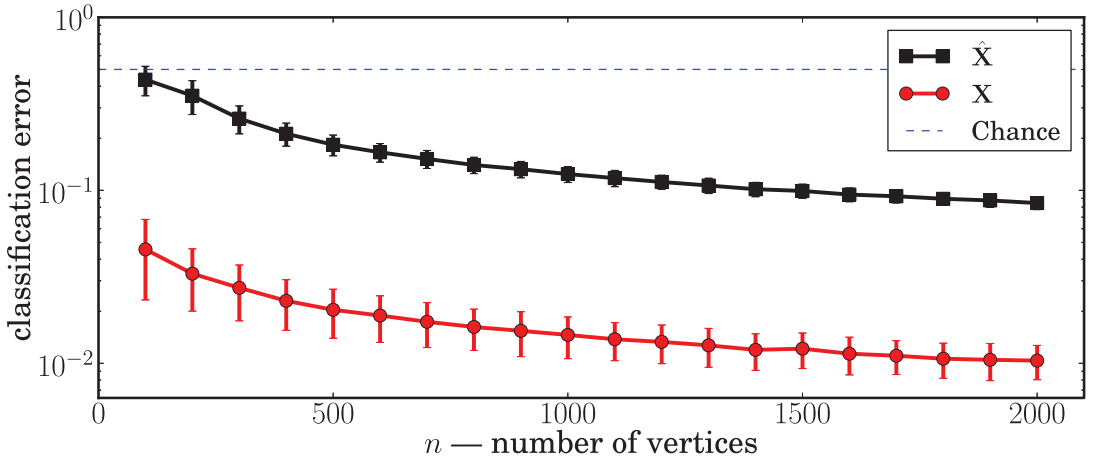


Fig. 3. Leave-one-out cross validation classification error estimates using k-nearest neighbors for the simulations described in Section 7.1. The black line show the classification error when classifying using $\widehat{\mathbf{X}}$, while the red line shows the error rates when classifying using $\mathbf{X}$. Error bars show the standard deviation over the 500 Monte Carlo replicates. Chance classification error is 0.5; $L^* = 0$. This figure suggests the rates of convergence may be similar for both $\mathbf{X}$ and $\widehat{\mathbf{X}}$.

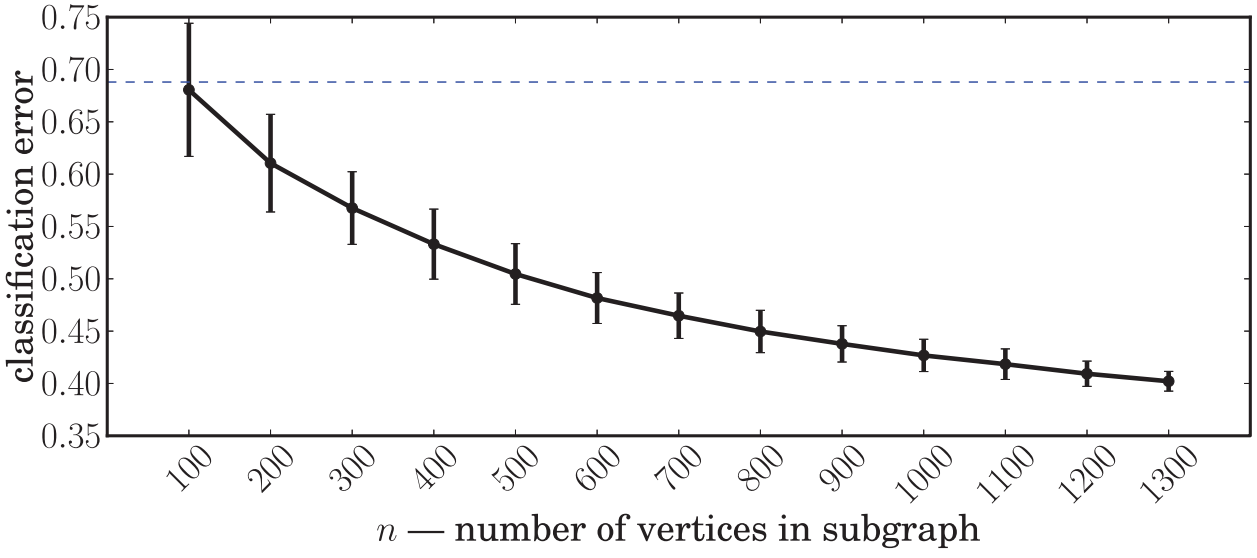


Fig. 4. Error rate using leave-one-out cross validation for random induced subgraphs. Chance classification error is ≈ 0.688, shown in blue. This illustrates the improvement vertex classification as the number of vertices and the number of observed class labels increases.

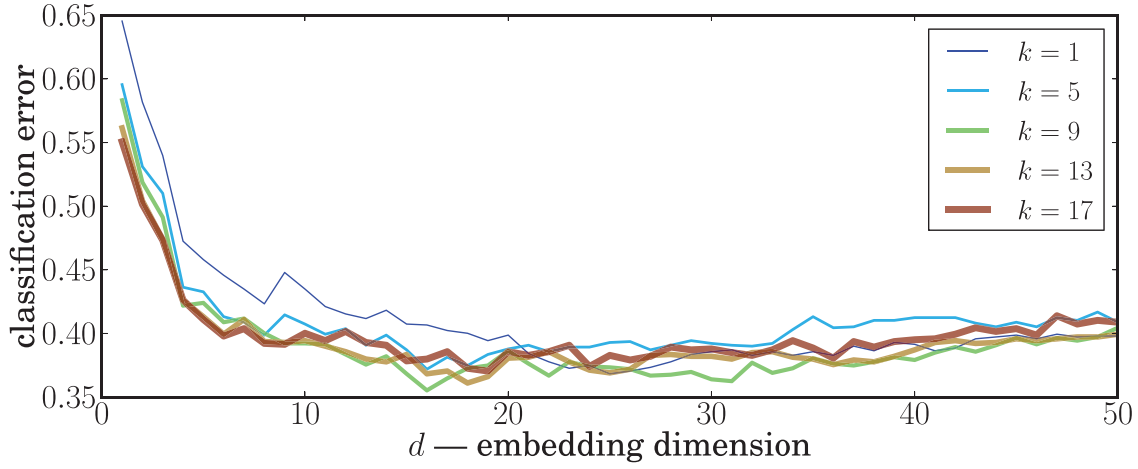


Fig. 5. Leave-one-out error rate plotted against the embedding dimension d for different choices of k (see legend). Each line corresponds to a different choice for the number of nearest neighbors k. All results are better than chance ≈ 0.688. We see that method is robust to changes of k and d near the optimal range.

performance estimates suggest that the classifiers may have similar properties and further investigation of this naive classifier may be of interest. Overall, the lowest estimated error rate was 0.35 using the embedding with $k = 9$ and $d = 16$, though a statistical comparison of the procedures was not considered.

8 Conclusion

Overall, we have shown that under the random dot product graph model, we can consistently estimate the latent positions provided they are independent and identically distributed. We have shown further that these estimated positions are also sufficient to consistently classify vertices. We have shown that this method works well in simulations and can be useful in practice for classifying documents based on their links to other documents.

Acknowledgments

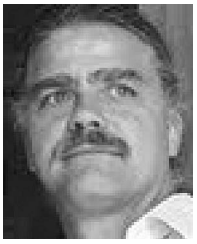
This work was partially supported by a Natinal Security Science and Engineering Faculty Fellowship (NSSEFF), by the Acheson J. Duncan Fund for the Advancement of Research in Statistics, and by the Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE).

References

[1] E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing, "Mixed Membership Stochastic Blockmodels," *J. Machine Learning Research,* vol. 9, pp. 1981-2014, 2008.
[2] D.J. Aldous, "Representations for Partially Exchangeable Arrays of Random Variables," *J. Multivariate Analysis,* vol. 11, no. 4, pp. 581-598, 1981.
[3] P.J. Bickel and A. Chen, "A Nonparametric View of Network Models and Newman-Girvan and Other Modularities," *Proc. Nat'l Academy of Sciences USA,* vol. 106, no. 50, pp. 21068-21073, 2009.
[4] P.J. Bickel, A. Chen, and E. Levina, "The Method of Moments and Degree Distributions for Network Models," *Annals of Statistics,* vol. 39, no. 5, pp. 38-59, 2011.
[5] D.S. Choi, P.J. Wolfe, and E.M. Airoldi, "Stochastic Blockmodels with a Growing Number of Classes," *Biometrika,* vol. 99, no. 2, pp. 273-284, 2012.
[6] F. Chung, L. Lu, and V. Vu, "The Spectra of Random Graphs with Given Expected Degrees," *Internet Math.,* vol. 1, no. 3, pp. 257-275, 2004.
[7] F. RK Chung, *Spectral Graph Theory.* Am. Math. Soc., 1997.
[8] C. Davis and W. Kahan, "The Rotation of Eigenvectors by a Pertubation. III," *SIAM J. Numerical Analysis,* vol. 7, pp. 1-46, 1970.
[9] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition.* Springer Verlag, 1996.
[10] M. Fiedler, "Algebraic Connectivity of Graphs," *Czechoslovak Math. J.,* vol. 23, no. 2, pp. 298-305, 1973.
[11] D.E. Fishkind, D.L. Sussman, M. Tang, J.T. Vogelstein, and C.E. Priebe, "Consistent Adjacency-Spectral Partitioning for the Stochastic Block Model When the Model Parameters Are Unknown," rapid post, 2012.
[12] M.S. Handcock, A.E. Raftery, and J.M. Tantrum, "Model-Based Clustering for Social Networks," *J. Royal Statistical Soc.: Series A (Statistics in Soc.),* vol. 170, no. 2, pp. 301-354, 2007.
[13] P.D. Hoff, A.E. Raftery, and M.S. Handcock, "Latent Space Approaches to Social Network Analysis," *J. Am. Statistical Assoc.,* vol. 97, no. 460, pp. 1090-1098, 2002.
[14] P.W. Holland, K. Laskey, and S. Leinhardt, "Stochastic Blockmodels: First Steps," *Social Networks,* vol. 5, no. 2, pp. 109-137, 1983.
[15] D.N. Hoover, "Relations on Probability Spaces and Arrays of Random Variables," *Inst. for Advanced Study, Princeton,* rapid post, 1979.
[16] R. Horn and C. Johnson, *Matrix Analysis.* Cambridge Univ. Press, 1985.
[17] O. Kallenberg, *Probabilistic Symmetries and Invariance Principles.* Springer Verlag, 2005.
[18] Z. Ma, D.J. Marchette, and C.E. Priebe, "Fusion and Inference from Multiple Data Sources in a Commensurate Space," *Statistical Analysis and Data Mining,* vol. 5, no. 3, pp. 187-193, 2012.
[19] D. Marchette, C.E. Priebe, and G. Coppersmith, "Vertex Nomination via Attributed Random Dot Product Graphs," *Proc. 57th ISI World Statistics Congress,* 2011.
[20] F. McSherry, "Spectral Partitioning of Random Graphs," *Proc. IEEE CS Int'l Conf. Cluster Computing,* pp. 529-537, 2001.
[21] C.L.M. Nickel, "Random Dot Product Graphs: A Model for Social Networks," PhD thesis, Johns Hopkins Univ., 2006.
[22] R.I. Oliveira, "Concentration of the Adjacency Matrix and of the Laplacian in Random Graphs with Independent Edges," rapid post, 2009.
[23] K. Rohe, S. Chatterjee, and B. Yu, "Spectral Clustering and the High-Dimensional Stochastic Blockmodel," *Annals of Statistics,* vol. 39, no. 4, pp. 1878-1915, 2011.
[24] T.A.B. Snijders and K. Nowicki, "Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure," *J. Classification,* vol. 14, no. 1, pp. 75-100, 1997.
[25] C.J. Stone, "Consistent Nonparametric Regression," *Annals of Statistics,* vol. 5, no. 4, pp. 595-620, 1977.
[26] D.L. Sussman, M. Tang, D.E. Fishkind, and C.E. Priebe, "A Consistent Adjacency Spectral Embedding for Stochastic Blockmodel Graphs," *J. Am. Statistical Assoc.,* vol. 107, no. 499, pp. 1119-1128, 2012.
[27] S. Young and E. Scheinerman, "Random Dot Product Graph Models for Social Networks," *Proc. Fifth Int'l Conf. Algorithms and Models for the Web-Graph,* pp. 138-149, 2007.

Daniel L Sussman received the BA degree in mathematics from Cornell University in 2008. From 2009-2010 he researched medical imaging under a Intramural Research Fellow Award at the Clinical Center of the National Institutes of Health. He is currently a third year PhD student in the Applied Math and Statistics Department at Johns Hopkins University. He has been awarded the Charles and Catherine Counselman Fellowship and the Newman Family Fellowship. His research interests include the statistical inference on random graphs and high-dimensional data analysis with applications in brain networks and space archeology.

Minh Tang received the BS degree from Assumption University, Thailand, in 2001, the MS degree from the University of Wisconsin, Milwaukee, in 2004, and the PhD degree from Indiana University, Bloomington, in 2010. He is currently a postdoctoral fellow in the Department of Applied Mathematics and Statistics at Johns Hopkins University. His research interests include statistical pattern recognition, dimensionality reduction, and high-dimensional data analysis.



Carey E Priebe received the BS degree in mathematics from Purdue University in 1984, the MS degree in computer science from San Diego State University in 1988, and the PhD degree in information technology (computational statistics) from George Mason University in 1993. From 1985 to 1994 he worked as a mathematician and scientist in the US Navy research and development laboratory system. Since 1994 he has been a professor in the Department of Applied Mathematics and Statistics, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland. At Johns Hopkins, he holds joint appointments in the Department of Computer Science, the Department of Electrical and Computer Engineering, the Center for Imaging Science, the Human Language Technology Center of Excellence, and the Whitaker Biomedical Engineering Institute. He is a past president of the Interface Foundation of North America-Computing Science and Statistics, a past chair of the American Statistical Association Section on Statistical Computing, a past vice president of the International Association for Statistical Computing, and is on the editorial boards of the *Journal of Computational and Graphical Statistics*, *Computational Statistics and Data Analysis*, and *Computational Statistics*. His research interests include computational statistics, kernel and mixture estimates, statistical pattern recognition, statistical image analysis, dimensionality reduction, model selection, and statistical inference for high-dimensional and graph data. He is a senior member of the IEEE, a Lifetime Member of the Institute of Mathematical Statistics, an Elected Member of the International Statistical Institute, and a fellow of the American Statistical Association.