A limit theorem for scaled eigenvectors of random dot product graphs

Avanti Athreya†, Vince Lyzinski†, David J. Marchette‡, Carey E. Priebe†, Daniel L. Sussman†∗, Minh Tang †
†Applied Math and Statistics Department, Johns Hopkins University,
‡Naval Surface Warfare Center
June 3, 2013

Abstract
We prove a central limit theorem for the components of the largest eigenvector of the adjacency matrix of a one-dimensional random dot product graph whose true latent positions are unknown. In particular, we follow the methodology outlined in Sussman et al. [2013] to construct consistent estimates for the latent positions, and we show that the appropriately scaled differences between the estimated and true latent positions converge to a mixture of Gaussian random variables. As a corollary, we obtain a central limit theorem for the first eigenvector of the adjacency matrix of an Erdős-Renyi random graph. We conjecture an analogous central limit theorem in the case of a higher-dimension random dot product graph, and we illustrate the multi-dimensional case through numerical simulations. A proof of this conjecture will have implications for the development of statistical procedures for random graphs analogous to the results on estimation, hypothesis testing, and clustering in the setting of a mixture of normal distributions in Euclidean space.

Spectral analysis of the adjacency and Laplacian matrices for graphs is of both theoretical [Chung, 1997] and practical [Luxburg, 2007] significance. For instance, the spectrum can be used to characterize the number of connected components in a graph and various properties of random walks on graphs, and the eigenvector corresponding to the second smallest eigenvalue of the Laplacian is used in the solution to a relaxed version of the min-cut problem [Fiedler, 1973]. In our current work, we investigate the second order properties of the eigenvectors corresponding to the largest eigenvalues of the adjacency matrix of a random graph. In particular, we show that under the random dot product graph model [Young and Scheinerman, 2007], the components of the eigenvectors are asymptotically normal and centered around the true latent positions (see Section 3.1). We consider only undirected, loop-free graphs in which the expected number of edges grows as \( \Theta(n^2) \).

This paper is organized as follows: in Section 1 we provide background and give a brief overview of related work. In Section 2 we prove our main central limit theorem for a one-dimensional random dot product graph. We then note two corollaries for special cases of the random dot product graph. In Section 3 we conjecture an analogous limit theorem for the case of a higher-dimensional random dot product graph, and in Section 4 we demonstrate evidence for this conjecture via simulation. Finally in Section 5 we provide further discussion of our results.

1 Background and Related Work

This work is concerned with the eigenvectors corresponding to the largest eigenvalues of the adjacency matrix of a random dot product graph (RDPG). Random dot product graphs are a specific example of latent position

*Corresponding Author: dsussma3@jhu.edu
random graphs [Hoff et al., 2002], in which each vertex is associated with a latent position and, conditioned on the latent positions, the presence or absence of all edges in the graph are independent. The edge presence probability is given by a link function, which is a symmetric function of the two latent positions.

We note briefly that, in a strong sense, latent position graphs are identical to exchangeable random graphs [Aldous, 1981, Hoover, 1979], with the key unifying ingredient being the conditional independence of the edges. A fundamental result on exchangeable graphs is the notion of a graph limit which is constructed via subgraph counts [Diaconis and Janson, 2007]. This work has important consequences in statistical inference, for instance the method of moments for subgraph counts [Bickel et al., 2011]. In a similar spirit, our work derives asymptotic distributions for spectral statistics that have the promise to improve current statistical methodology for random graphs.

Statistical analysis for latent position random graphs has received much recent interest: see Goldenberg et al. [2010] and Fortunato [2010] for reviews of the pertinent literature. Some fundamental results are found in Bickel et al. [2011], Bickel and Chen [2009], Choi et al. [2012] among many others. In the statistical analysis of random position random graphs, a common strategy is to first estimate the latent positions based on some specified link function. For example, in RDPGs [Young and Scheinerman, 2007], the link function is the dot product: namely, the edge probabilities are given by the dot products of the latent positions. Sussman et al. [2013] show that spectral decompositions of the adjacency matrix for a random dot product graphs provide accurate estimates of the underlying latent positions. In this work, we extend the analysis in Sussman et al. [2013] to show a distributional convergence of the residuals between the estimated and true latent positions.

Our work is also influenced by the analysis of the spectra of random graphs [Chung, 1997]. Of special note is the classic paper of Füredi and Komlós [1981], in which the authors show that for an Erdős-Rényi graph with parameter $p$, the appropriately scaled largest eigenvalue of the adjacency matrix converges in law to a normal distribution. Other results of this type are proved for sparse graphs in both the independent edge model [Krivelevich and Sudakov, 2003] and the $d$-regular random graph model [Janson, 2005]. More recently, general bounds for the operator norm of the difference between the adjacency matrix and its expectation have been proved in Oliveira [2009] and Tropp [2011] (see Proposition 2, Eq. (5)).

A final important influence is the field of random matrix theory. A recent result by Tao and Vu [2012] proves a central limit theorem for the eigenvectors of a mean zero random symmetric matrix with independent entries. Tao and Vu [2012] prove a result for eigenvectors corresponding to the bulk of the spectra and Knowles and Yin [2011] prove a similar result for eigenvectors near the “edge” of the spectra. A material difference between these results and our present work, however, is that we consider random matrices whose entries have nonzero mean. For mean zero matrices, the eigenvalues and eigenvectors that are most readily studied are not the largest in magnitude but those in the “bulk” of the spectra, while in our setting, the structure of the mean matrix eases the study of the largest eigenvalues and their corresponding eigenvectors.

As will be seen in Section 2, the key step in our work is to apply the power method to the adjacency matrix, with the initial vector given by the true latent position. Conditioned on the true latent position, this produces a vector whose components are asymptotically normally distributed. Furthermore, the difference between this vector and the true eigenvector of the adjacency matrix is asymptotically negligible, due to a large gap between the largest eigenvalue and the remaining eigenvalues.

**2 Main Theorem**

In this section, we state and prove a central limit theorem for a one-dimensional random dot product graph, defined as follows.

**Definition 1** (Random Dot Product Graph $(d=1)$). For a distribution $F$ on $[0, 1]$, we say that $(X, A) \sim \text{RDPG}(F)$ if the following hold. Let $X_1, \ldots, X_n \sim_{\text{iid}} F$ and define

$$X = [X_1, X_2, \ldots, X_n]^\top \in \mathbb{R}^{n \times 1} \text{ and } P = XX^\top \in [0, 1]_{\text{sym}}^{n \times n}. \tag{1}$$

2
The $X_i$’s are the latent positions for the random graph. The matrix $A \in \{0,1\}^{n \times n}$ is defined to be a symmetric, hollow matrix such that for all $i < j$, conditioned on $X_i$ and $X_j$,

$$A_{ij} \overset{ind}{\sim} \text{Bern}(X_i X_j).$$  \hfill (2)

We remark that this one-dimensional model is a slight modification of the rank 1 inhomogeneous random graph model studied as an example in [Bollobás et al., 2007].

To estimate $X$, we define $\hat{X} = \frac{\lambda^{1/2} \hat{V}}{\lambda}$, where $\lambda = \lambda_1(A)$ is the largest eigenvalue of $A$ and $\hat{V}$ its associated eigenvector, normalized to be of unit length. We define $V = \frac{\hat{X}}{\|\hat{X}\|}$, so $V$ is the normalized true latent positions. Let $\delta = E[X_i^2]$ be the second moment of the latent positions.

Throughout this work, we will need explicit control on the differences, in Frobenius norm, between $X$ and $\hat{X}$ and $V$ and $\hat{V}$. We state here the necessary bounds, which are proven in Sussman et al. [2013] and Oliveira [2009].

**Proposition 2.** Let $\delta = E[X_i^2]$, $V = X/\|X\|_2$ and $\hat{V} = \hat{X}/\|\hat{X}\|_2$. With probability greater than $1 - \frac{5}{n^2}$,

$$\|X - \hat{X}\|_2 \leq \sqrt{12 \log n/\delta},$$  \hfill (3)

$$\|V - \hat{V}\|_2 \leq \sqrt{\frac{3 \log n}{n\delta^2}},$$  \hfill (4)

and  

$$\|XX^\top - A\|_{2\to2} \leq \sqrt{12n \log n}$$  \hfill (5)

and for $n$ sufficiently large, the events above imply that with probability greater than $1 - \frac{5}{n^2}$,

$$\frac{\delta n}{2} \leq \|P\|_{2\to2} \leq n \quad \text{and} \quad \frac{\delta n}{2} \leq \|A\|_{2\to2} \leq n.$$  \hfill (6)

where $\|\cdot\|_{2\to2}$ represents the $L^2$ operator norm.

Our aim in this section is to prove the following limit theorem.

**Theorem 3.** Let $(X,A) \sim \text{RDPG}(F)$ and let $\hat{X}$ be our estimate for $X$. Let $\Phi(z,\sigma^2)$ denote the normal cumulative distribution function, with mean zero and variance $\sigma^2$, evaluated at $z$. Then for each component $i$ and any $z \in \mathbb{R}$,

$$P \left\{ \sqrt{n} \left( \hat{X}_i - X_i \right) \leq z \right\} \to \Phi(z,\delta^{-2}\sigma^2(x_i))dF(x_i)$$

where $\sigma^2(x_i) = x_iE[X_i^2] - x_i^2E[X_i^4]$ and $\delta = E[X_i^2]$. That is, the sequence of random variables $\sqrt{n} \left( \hat{X}_i - X_i \right)$ converges in distribution to a mixture of normals. We denote this mixture by $\mathcal{N}(0,\delta^{-2}\sigma^2(X_i))$.

As an immediate consequence, we obtain the following corollary for the eigenvectors of an Erdős-Renyi random graph. For the Erdős-Rényi graph, the $X_i$ have a degenerate distribution; namely, there is some $p \in (0,1)$ such that $X_i = \sqrt{p}$ for all $i$.

**Corollary 4.** For an Erdős-Renyi $(p)$ graph, the following Central Limit Theorem holds:

$$\sqrt{n} \left( \hat{X}_i - \sqrt{p} \right) \overset{d}{\to} \mathcal{N}(0,1-p).$$

To prove Theorem 3, we will need Proposition 2 and a succession of simpler lemmas. To begin, we apply one step of the power method, with initial vector $V$. In particular, let $Y = \frac{1}{\lambda}AV$ be the vector in $\mathbb{R}^n$ with components $Y_i = \frac{1}{\lambda} \sum_{j=1}^n A_{ij}V_j$.  

3
Proposition 5. Taking $X_i = x_i$ as given, we have

$$\sqrt{n} \left( Y_i - \frac{\|X\|_2}{\tilde{\lambda}_1/2} \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \delta^{-2}\sigma^2(x_i))$$

where $\sigma^2(\cdot)$ and $\delta$ are as in Theorem 3.

Proof. Observe that

$$\sqrt{n} \left( Y_i - \frac{\|X\|_2}{\tilde{\lambda}_1/2} \right) = \frac{\sqrt{n}}{\tilde{\lambda}_1/2} \left( \sum_{j=1}^n A_{ij} V_j - \|X\|_2 x_i \right)$$

$$= \frac{\sqrt{n}}{\tilde{\lambda}_1/2 \|X\|_2} \left( \sum_{j=1}^n A_{ij} X_j - x_i \|X\|_2^2 \right)$$

$$= \left( \frac{n}{\tilde{\lambda}_1/2 \|X\|_2} \right) \left( \frac{1}{\sqrt{n}} \left[ \sum_{i \neq j} (A_{ij} - x_i X_j) X_j \right] - x_i^3 \right).$$

The scaled sum

$$\frac{1}{\sqrt{n}} \left( \sum_{j \neq i} (A_{ij} - x_i X_j) X_j \right)$$

is a sum of independent, identically distributed random variables each with mean zero and variance

$$\sigma^2(x_i) = x_i \mathbb{E}[X_j^3] - x_i^3 \mathbb{E}[X_j^4].$$

The classical Lindeberg-Feller Central Limit Theorem and Slutsky’s Theorem [Chung, 1974, Theorem 7.2.1 and Theorem 4.4.6] imply that

$$\frac{1}{\sqrt{n}} \left[ \sum_{i \neq j} (A_{ij} - x_i X_j) X_j \right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(x_i)). \quad (7)$$

Furthermore, the Strong Law and Proposition 2 imply that

$$\frac{n}{\sqrt{\tilde{\lambda}_1/2 \|X\|_2}} \xrightarrow{a.s.} \delta^{-1},$$

so another application of Slutsky’s Theorem allows us to conclude that

$$\sqrt{n} \left( Y_i - \frac{\|X\|_2}{\tilde{\lambda}_1/2} \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \delta^{-2}\sigma^2(x_i)).$$

The next lemma shows that the scaling factor $\|X\|_2/\tilde{\lambda}_1/2$ is asymptotically close to one.

Lemma 6. In the setting of Theorem 3, $\sqrt{n} \left( 1 - \frac{\|X\|_2}{\tilde{\lambda}_1/2} \right) \xrightarrow{a.s.} 0.$

Proof. First, note that

$$1 - \frac{\|X\|_2}{\tilde{\lambda}_1/2} = \frac{\tilde{\lambda} - \|X\|_2^2}{\tilde{\lambda}_1/2 (\|X\|_2 + \tilde{\lambda}_1/2)}.$$

By Proposition 2, the denominator

$$\tilde{\lambda}_1/2 (\|X\|_2 + \tilde{\lambda}_1/2) \geq \frac{\delta n}{2}$$
with probability at least $1 - \frac{5}{n^2}$. Since

$$
\|X\|_2^2 = \lambda_1(P) = V^T PV \quad \text{and} \quad \hat{\lambda} = \hat{V}^T A \hat{V},
$$

it follows that

$$
\|X\|_2^2 - \hat{\lambda} \leq |V^T PV - V^T AV| + |V^T AV - \hat{V}^T A \hat{V}|.
$$

For the second term on the right-hand-side of (8), first observe that

$$
|\hat{V}^T (V - \hat{V})| = \frac{1}{2} |\hat{V}^T V - \hat{V}^T \hat{V} + V^T \hat{V} - V^T V| = \frac{1}{2} \|V - \hat{V}\|^2 \leq \frac{3 \log n}{2n \delta^2}.
$$

It follows that

$$
|V^T AV - \hat{V}^T A \hat{V}| = |(V - \hat{V})^T A(V - \hat{V}) + 2(V - \hat{V})^T A \hat{V}|
$$

$$
= |(V - \hat{V})^T A(V - \hat{V}) + 2\hat{\lambda}(V - \hat{V})^T \hat{V}|
$$

$$
\leq 2\hat{\lambda}\|V - \hat{V}\|_2 \leq \frac{6 \log n}{\delta},
$$

with probability at least $1 - \frac{5}{n^2}$ by Proposition 2. For the first term in Eq. 8, we have

$$
|V^T AV - V^T PV| = |\sum_{i,j}(A_{ij} - P_{ij})V_i V_j| \leq 2|\sum_{i<j}(A_{ij} - P_{ij})V_i V_j| + |\sum_i P_{ii}V_i^2|.
$$

Put $V_{\max}^2 = \max_i V_i^2$. Proposition 2 guarantees that with probability at least $1 - \frac{5}{n^2}$, $V_{\max}^2 \leq \frac{4}{n\delta}$. The second summand on the right-hand side of (11) is bounded by $nV_{\max}^2$. If we condition on $X$, which is equivalent to conditioning on $V$, the first summand in (11) is the sum of independent mean-zero random variables, all bounded by $V_{\max}^2$. Define $D(n)$ to be the event

$$
D(n) = \left\{ 2|\sum_{i<j}(A_{ij} - P_{ij})V_i V_j| \geq V_{\max}^2 n \sqrt{8 \log n} \right\}
$$

An application of Hoeffding’s inequality ensures that the following bound holds on the conditional probability of $D(n)$ given $V$:

$$
P[D(n)|V] \leq 2 \exp \left( -\frac{2n \log n}{n - 1} \right).
$$

The upper bound in (13) is independent of $V$. Therefore,

$$
P[D(n)] \leq 2 \exp \left( -\frac{2n \log n}{n - 1} \right).
$$

Let $D(n)^c$ denote the complement of $D(n)$. Again, $V_{\max}^2 \leq \frac{4}{n\delta}$ with probability at least $1 - \frac{5}{n^2}$, so

$$
P\left[D(n)^c \cap \left\{ V_{\max}^2 \leq \frac{4}{n\delta} \right\} \right] \geq 1 - \frac{C}{n^2}.
$$

Now, on $D(n)^c \cap \left\{ V_{\max}^2 \leq \frac{4}{n\delta} \right\}$,

$$
|2 \sum_{i<j}(A_{ij} - P_{ij})V_i V_j| \leq \frac{4\sqrt{8 \log n}}{\delta},
$$

and the desired bound follows. □
Remark 7. Note that the proof of the previous lemma shows that with high probability, $|\hat{\lambda} - \|X\|_2^2| \leq C \log n$ for some $C$ depending only on the distribution $F$. This bound is similar in kind to the central limit theorem, proved in Füredi and Komlós [1981], for the largest eigenvalue of an Erdős-Rényi random graph. Alon et al. [2002] also provide similar concentration rates for the first eigenvalue, namely that $|\lambda - E[\lambda]|$ can be tightly controlled, which is somewhat different from our result. This result also greatly improves on the bound one obtains using only the operator norm bound of Oliveira [2009] (see Proposition 2).

Finally, we prove a bound on the $L^2$ distance between $\hat{\lambda}^{-1/2}AV$ and $\hat{X}$.

**Proposition 8.** Let $Y = \hat{\lambda}^{-1/2}AV$. Provided the events in Theorem 2 occur, we have

$$\|Y - \hat{X}\| \leq \frac{24 \log n}{\sqrt{n \delta^3}}.$$ 

**Proof.** Let $E = A - P = A - XX^T$. We have

$$\|Y - \hat{X}\|_2 = \hat{\lambda}^{-1/2} \|A(V - \hat{V})\|_2 = \hat{\lambda}^{-1/2} \|([\hat{\lambda}V\hat{V}^T + E](V - \hat{V}))\|_2 \leq \hat{\lambda}^{1/2} \|V\hat{V}^T(V - \hat{V})\| + \hat{\lambda}^{-1/2} \|E(V - \hat{V})\|_2$$

Using Eq. (9) and

$$\sqrt{n \delta^2 / 2} \leq \hat{\lambda}^{1/2} \leq 2 \sqrt{n \delta},$$

the first term is bounded above by $\frac{12 \log n}{\sqrt{n \delta^3}}$. Proposition 2 guarantees that $\|E\|_2 \leq \sqrt{12n \log n}$, and therefore

$$\hat{\lambda}^{-1/2} \|E(V - \hat{V})\|_2 \leq \hat{\lambda}^{-1/2} \|E\|_2 \|V - \hat{V}\|_2 \leq \frac{2}{\sqrt{n \delta}} \sqrt{12n \log n} \sqrt{\frac{3 \log n}{n \delta^2}} \leq \frac{12 \log n}{\sqrt{n \delta^3}},$$

from which the desired bound follows.

We are now equipped to prove our limit theorem:

**Proof of Theorem 3.** Integrating over the possible realizations of $X_i$ in Proposition 5 and applying the dominated convergence theorem, we deduce that

$$P \left\{ \sqrt{n} \left( Y_i - \frac{\|X\|_2}{\lambda^{1/2}} X_i \right) \leq z \right\} \rightarrow \Phi(z, \delta^{-2} \sigma^2(x_i))dF(x_i).$$

This establishes that

$$\sqrt{n} \left( Y_i - \frac{\|X\|_2}{\lambda} X_i \right) \rightarrow \mathcal{N}(0, \sigma^2(X_i)).$$

Markov’s inequality, the exchangeability of $\{Y_i - \hat{X}_i\}_{i=1}^n$, and the bounds in Prop. 8 allow us to conclude that

$$P[\sqrt{n}|Y_i - \hat{X}_i| > \epsilon] \leq \frac{\mathbb{E}[n(Y_i - \hat{X}_i)^2]}{\epsilon^2} = \frac{\mathbb{E}[\|Y - \hat{X}\|_2^2]}{\epsilon^2} \leq \frac{C \log^2 n}{\epsilon^2 n}.$$ 

Hence $\sqrt{n}(Y_i - \hat{X}_i)$ converges to zero in probability. Observe that

$$\sqrt{n} \left( \hat{X}_i - \frac{\|X\|_2}{\lambda^{1/2}} X_i \right) = \sqrt{n} \left( \hat{X}_i - Y_i \right) + \sqrt{n} \left( Y_i - \frac{\|X\|_2}{\lambda^{1/2}} X_i \right),$$

and Theorem 3 follows from the convergence in probability of the first summand, the distributional convergence of the second summand, and Lemma 6.

\qed
2.1 Corollaries

In this section, we prove three corollaries, each of which is either a special case or an extension of our main theorem. First, we demonstrate that in the stochastic blockmodel, if we condition on \(X_i = x\), the residuals converge to the correct mixture component; second, we prove a similar result in the case where the latent position distribution has a density and we condition on \(X_i\) belonging to any set of positive \(F\)-measure, where \(F\) is the distribution of the latent positions; and finally, we prove a central limit theorem for the distribution of any fixed number \(k\) of the residuals, \(\sqrt{n}(\hat{X}_i - X_i)\), for \(1 \leq i \leq k\). We begin with the first corollary, in which we obtain appropriate convergence to the correct mixture component for the stochastic blockmodel, as defined in the statement of Corollary 9 below.

**Corollary 9.** In the setting of Theorem 3, let \(X = \text{supp}(F) \subset [0, 1] \) be the support of the distribution of the \(X_i\) and suppose that \(|X| = m < \infty\). Suppose for each \(x \in X\), we have that \(P[X_i = x] = \pi_x > 0\). Then for all \(x \in X\), if we condition on \(X_i = x\), we obtain

\[
P\left(\sqrt{n}\left(\hat{X}_i - x\right) \leq z \middle| X_i = x\right) \to \Phi(z, \delta^{-2}\sigma^2(x)) \tag{14}
\]

where \(\sigma^2(\cdot)\) and \(\delta\) are as in Theorem 3.

**Proof.** Let \(p_n(x, \epsilon) = P\left[\sqrt{n}|\hat{X}_i - Y_i| > \epsilon \middle| X_i = x\right]\) where \(Y = \lambda^{-1/2}AV\) is as in Proposition 8. By Proposition 5 and Slutsky’s Theorem, we need only show that for all \(\epsilon > 0\) and \(x \in X\),

\[
p_n(x, \epsilon) \to 0 \text{ as } n \to \infty, \tag{15}
\]

because this yields

\[
\sqrt{n}\left(\hat{X}_i - \frac{\|X\|_2}{\lambda^{1/2}}\right) \overset{d}{\to} N(0, \delta^{-2}\sigma^2(x)). \tag{16}
\]

First, by Markov’s inequality, Proposition 8, and the exchangeability of the sequence \(\hat{X}_i - Y_i\), we have

\[
P[\sqrt{n}|\hat{X}_i - Y_i| > \epsilon] \leq \frac{E[\|X - Y\|_2^2]}{\epsilon^2} \leq \frac{C\log^2 n}{n\epsilon^2}
\]

for some \(C > 0\) depending only on the distribution \(F\). Let \(\pi_{\text{min}} = \min_{x \in X} \pi_x\). We then have

\[
P[\sqrt{n}|\hat{X}_i - Y_i| > \epsilon] = \sum_{x' \in X} \pi_{x'}p_n(x', \epsilon) \geq \pi_{\text{min}}p_n(x, \epsilon)
\]

for all \(x \in X\). This implies

\[
p_n(x, \epsilon) \leq \frac{C\log^2 n}{n\pi_{\text{min}}\epsilon^2},
\]

which proves Eq. (15). The same argument can be applied to show that

\[
P\left[\sqrt{n}\left(1 - \frac{\|X\|_2}{\lambda^{1/2}}\right) > \epsilon \middle| X_i = x\right] \to 0
\]

for all \(\epsilon > 0\) and \(x \in X\). An application of Slutsky’s Theorem completes the proof.

This next corollary has essentially the same proof as the previous one; here, however, the set of possible latent positions need not be finite. We condition, instead, on the true position belonging to a fixed set \(B\) for which \(P(X_i \in B)\) is strictly positive (but we note that this probability can be arbitrarily small).
Corollary 10. In the setting of Theorem 3, suppose that $\mathcal{B} \subset [0,1]$ is such that $\mathbb{P}[X_i \in \mathcal{B}] > 0$. If we condition on the event $\{X_i \in \mathcal{B}\}$, we obtain

$$P \left[ \sqrt{n} \left( \hat{X}_i - X_i \right) \leq z \mid X_i \in \mathcal{B} \right] \to \frac{1}{P(X_i \in \mathcal{B})} \int_{\mathcal{B}} \Phi(z, \delta^{-2}\sigma^2(x)) dF(x_i)$$

(17)

where $\sigma^2(\cdot)$ and $\delta$ are as in Theorem 3.

In other words, if we condition on an event of positive probability, the convergence in distribution is to a mixture of normals where the mixture is over only the conditioned event. Our last corollary asserts that our main theorem can be extended to distributional convergence of any finite collection of estimated latent positions. Indeed, we prove that the residuals $\sqrt{n}(\hat{X}_i - X_i)$ are asymptotically jointly normal and asymptotically uncorrelated, and hence asymptotically independent.

Corollary 11. Suppose $X$ and $\hat{X}$ are as in Theorem 3. Let $k \in \mathbb{N}$ be any fixed positive integer; let $i_1, i_2, \ldots, i_k \in \mathbb{N}$ be any fixed set of indices and let $z_1, z_2, \ldots, z_k \in \mathbb{R}$ be fixed. Then

$$\lim_{n \to \infty} \mathbb{P}[\sqrt{n}(\hat{X}_{i_1} - X_{i_1}) \leq z_1, \sqrt{n}(\hat{X}_{i_2} - X_{i_2}) \leq z_2, \ldots, \sqrt{n}(\hat{X}_{i_k} - X_{i_k}) \leq z_k] = \prod_{j=1}^{k} \int_{\mathcal{X}} \Phi(z_j, \delta^{-1}\sigma^2(x_j)) dF(x_j)$$

(18)

where $\Phi(\cdot, \sigma^2)$ denotes the cumulative distribution function (cdf) for a normal with mean zero and variance $\sigma^2$. Again, $\sigma^2(\cdot)$ and $\delta$ are as in Theorem 3.

In other words, for any finite collection of indices, the residuals between $\hat{X}$ and $X$ converge to independent mixtures of multivariate normals which we will denote $\mathcal{N}(0, \delta^{-2}\sigma^2(X_j))$:

$$\left( \begin{array}{c} \sqrt{n}(\hat{X}_{i_1} - X_{i_1}) \\ \sqrt{n}(\hat{X}_{i_2} - X_{i_2}) \\ \vdots \\ \sqrt{n}(\hat{X}_{i_k} - X_{i_k}) \end{array} \right) \xrightarrow{\mathcal{D}} \bigotimes_{j=1}^{k} \mathcal{N}(0, \delta^{-1}\sigma^2(X_j)) \text{ as } n \to \infty.$$

(19)

Sketch of proof. The proof essentially follows from an application of the Cramér-Wold theorem. For ease of notation, we consider the $k = 2$ case, where we can take $i \neq i'$ and condition on $X_i = x_i$ and $X_{i'} = x_{i'}$; the case for general $k$ follows similarly. This simplifies the covariance computation to the following:

$$\text{Cov} \left( \frac{1}{\sqrt{n}} \left[ \sum_{j \neq i} (A_{ij} - x_iX_j)X_j \right], \frac{1}{\sqrt{n}} \left[ \sum_{j' \neq i'} (A_{i'j'} - x_{i'}X_{j'})X_{j'} \right] \right)$$

$$= \frac{1}{n} \sum_{j=1}^{n} \sum_{j'=1}^{n} \text{Cov}((A_{ij} - x_iX_j)X_j, (A_{i'j'} - x_{i'}X_{j'})X_{j'}).$$

Now, if $j, j' \neq \{i, i'\}$, the summand is zero because the terms are independent. This leaves only 4 possible non-zero summands, so the covariance is bounded by $\frac{4}{n}$. The remainder of the proof follows mutatis mutandis as a consequence of Slutsky’s Theorem and the bounds established in Proposition 2 and Proposition 8.

3 A conjecture for higher-dimensional random dot product graphs

3.1 Setting for the higher-dimensional case

In this section, we consider a higher-dimensional random dot product graph, and give a precise conjecture for the limiting distribution of the differences between the estimated and true latent positions. We begin with the construction of our estimate for the underlying latent positions.
Definition 12 (Random Dot Product Graph (d-dimensional)). Let $F$ be a distribution on a set $\mathcal{X} \subset \mathbb{R}^d$ satisfying $\langle x, x' \rangle \in [0, 1]$ for all $x, x' \in \mathcal{X}$. We say $(X, A) \sim \text{RDPG}(F)$ if the following hold. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$ and define
\[ X = [X_1, X_2, \ldots, X_n]^\top \in \mathbb{R}^{n \times d} \quad \text{and} \quad P = XX^\top \in [0, 1]^{n \times n}. \] The $X_i$ are the latent positions for the random graph. The matrix $A \in \{0, 1\}^{n \times n}$ is defined to be a symmetric, hollow matrix such that for all $i < j$, conditioned on $X_i, X_j$, \[ A_{ij} \overset{i.i.d.}{\sim} \text{Bern}(X_i^\top X_j). \] As in Section 2, we seek to demonstrate an asymptotically normal estimate of $X$, the matrix of latent positions $X_1, \ldots, X_n$. However, we first note that the model as specified above is non-identifiable: if $W \in \mathbb{R}^{d \times d}$ is orthogonal, then $XW$ generates the same distribution over adjacency matrices. As a result, we will often consider uncentered principal components (UPCA) of $X$.

Definition 13 (UPCA). Let $X$ and $P$ be as in Definition 12. Then $P$ is symmetric and positive semidefinite and has rank at most $d$. Hence, $P$ has an spectral decomposition given by $P = U_P S_P U_P^\top$, where $U_P \in \mathbb{R}^{n \times d}$ has orthonormal columns and $S_P$ is a diagonal matrix with positive decreasing entries along the diagonal. The UPCA of $X$ is then $U_P S_P^{1/2}$. Note that $U_P S_P^{1/2} = X W_n$ for some random orthogonal matrix $W_n \in \mathbb{R}^{d \times d}$. We will denote the UPCA of $X$ as $\tilde{X}$.

Remark 14. We denote the second moment matrix for $X_i$ by $\Delta = \mathbb{E}[X_i X_i^\top]$. We assume without loss of generality that this matrix is diagonal, so $\Delta = \text{diag}(\delta_1, \delta_2, \ldots, \delta_d)$. For the remainder of this work we will assume that the eigenvalues of $\Delta$ are distinct and positive, so that $\Delta$ has a strictly decreasing diagonal.

Our estimate for $X$, or specifically our estimate for the UPCA of $X$, is given by a spectral embedding defined below, which is motivated by the observation that $A$ is essentially a noisy version of $P$.

Definition 15 (Embedding of $A$). Suppose that $A$ is as in Definition 12. Let $A = \tilde{U}_A S_A U_A^\top$ be the (full) spectral decomposition of $A$. Then our estimate for the UPCA of $X$ is $\tilde{X} = U_A S_A^{1/2}$, where $S_A \in \mathbb{R}^{d \times d}$ is the diagonal matrix with the $d$ largest eigenvalues (in magnitude) of $A$ and $U_A \in \mathbb{R}^{n \times d}$ is the matrix with orthonormal rows of the corresponding eigenvectors.

As in the one-dimensional case, we will again need explicit control on the differences, in Frobenius norm, between $X$ and $\tilde{X}$, as well as $U_A$ and $U_P$. These are the higher-dimensional analogues of the bounds in Proposition 2 and are also proven in Sussman et al. [2013] and Oliveira [2009].

Proposition 16. Suppose $X, A, \tilde{X}$ and $\tilde{X}$ are as defined above. Recall that $\delta_d$ denotes the smallest eigenvalue of $\Delta = \mathbb{E}[X_1 X_1^\top]$. With probability greater than $1 - \frac{2d^2 + 3}{n^2}$,
\[ \|\tilde{X} - X\|_F = \|U_A S_A^{1/2} - U_P S_P^{1/2}\|_F \leq 2d \sqrt{\frac{3 \log n}{\delta_d^2}}, \] \[ \|U_A - U_P\|_F \leq d \sqrt{\frac{3 \log n}{n \delta_d^2}}, \] and \[ \|XX^\top - A\|_{2 \rightarrow 2} \leq \sqrt{12 n \log n}. \]

We note that for $n$ sufficiently large, the events above imply that with probability greater than $1 - \frac{2d^2 + 3}{n^2}$, \[ \frac{\delta_d n}{2} \leq \|S_P\|_{2 \rightarrow 2} \leq n \quad \text{and} \quad \frac{\delta_d n}{2} \leq \|S_A\|_{2 \rightarrow 2} \leq n. \]
3.2 A conjecture for the limiting distribution of the residuals in higher dimensions

We now conjecture the form of the limiting distribution of the scaled differences between the estimated and true latent positions. In what follows, $X \in \mathbb{R}^{n \times d}$, where $d > 1$. However, the method of proof we employed in the one-dimensional case cannot be easily generalized to the $d > 1$ case. In the $d = 1$ case, we bound the difference between $\hat{X}$ and the vector $Y$, the latter a single scaled step of the power method, by exploiting the fact that $\hat{V}^T V$ is simply a real number, so $(\hat{V}^T V)^T = V^T \hat{V}$. In the multi-dimensional case, we need to bound the difference between $\hat{X}$ and $A U_p S_A^{-1/2}$, but the matrix $U_p U_A$ need not be symmetric.

**Conjecture 17.** Suppose $X$ and $A$ are as in Definition 12, $\hat{X}$ is the UPCA of $X$ and $\hat{X}$ is the embedding of $A$. Let $k \in \mathbb{N}$ be any fixed positive integer; let $i_1, i_2, \ldots, i_k \in \mathbb{N}$ be any fixed set of indices; and let $z_1, z_2, \ldots, z_k \in \mathbb{R}^d$ be any fixed vectors. Then

$$
\lim_{n \to \infty} \mathbb{P}[\sqrt{n}(\hat{X}_{i_1} - \bar{X}_{i_1}) \leq z_1, \sqrt{n}(\hat{X}_{i_2} - \bar{X}_{i_2}) \leq z_2, \ldots, \sqrt{n}(\hat{X}_{i_k} - \bar{X}_{i_k}) \leq z_k] = \prod_{j=1}^k \int_X \Phi(z_j, \Sigma(x_j))dF(x_j)
$$

where $\leq$ denotes elementwise inequality and $\Phi(\cdot, \Sigma)$ denotes the cumulative distribution function (cdf) for a $d$-variate normal with mean zero and covariance matrix $\Sigma$. The covariance matrix is given by the function

$$
\Sigma(x) = \Delta^{-1} \mathbb{E}[X_j^T X_j (x^T X_j)^2] \Delta^{-1},
$$

where $\Delta$ is the second moment matrix.

In other words, for any finite collection of indices, the residuals between the estimate and the UPCA of $X$ converge to independent mixtures of multivariate normals. As mentioned above, our proof technique breaks down in the multi-dimensional analogue of Proposition 8. The remainder of the proof can be adapted to the multi-dimensional setting. For example, our bounds on the difference $|\lambda - \|X\|_2^2|$ discussed in Remark 7 can be adapted to the multi-dimensional setting to show tight concentration of the matrices $S_A$ and $S_P$; specifically, not only does the largest eigenvalue concentrate, but the top $d$ eigenvalues all concentrate rapidly. Finally, we note that the corollaries presented in Section 2.1 all have $d$-dimensional analogues that follow from our conjecture. In the next section, we describe a simulation that provides empirical support for the conjecture.

4 Simulations

To illustrate Conjecture 17, we consider random graphs generated according to a stochastic block model (SBM) with parameters $B = \begin{bmatrix} 0.42 & 0.42 \\ 0.2 & 0.5 \end{bmatrix}$ and $\pi = (0.6, 0.4)$. (27)

In this model, each node is either in block 1 (with probability 0.6) or block 2 (with probability 0.4). Adjacency probabilities are determined by the entries in $B$ based on the block memberships of the incident vertices. The above SBM corresponds to a random dot product model in $\mathbb{R}^2$ where the distribution $F$ of the latent positions is a mixture of point masses located at $x_1 \approx (0.63, -0.14)$ (with prior probability 0.6) and $x_2 \approx (0.69, 0.13)$ (with prior probability 0.4).

We sample an adjacency matrix $A$ for graphs on $n$ vertices from the above model for various choices of $n$. For each graph $G$, let $\hat{X} \in \mathbb{R}^{n \times 2}$ denote the embedding of $A$ and let $\hat{X}_i$ denote the $i$-th row of $\hat{X}$. In Figure 1, we plot the $n$ rows of $\hat{X}$ for the various choice of $n$. The points are colored according to the block membership of the corresponding vertex in the SBM. The ellipses show the 95% level curves for the distribution of $\hat{X}_i$ for each block as specified by the conjectured limiting distribution, namely the ellipse such that $\mathbb{P}[\hat{X}_i \in \text{Ellipse } k|X_i = x_k] = .95$ for $k = 1, 2$. 

10
We then estimate the covariance matrices for the residuals. The theoretical covariance matrices as given by Conjecture 17 are given in the last line of Table 1, where $\Sigma_1$ and $\Sigma_2$ are the covariance matrices for the residual $\sqrt{n}(\hat{X}_i - X_i)$ when $X_i$ is from the first block and second block, respectively. The empirical covariance matrices, denoted $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$, are computed by evaluating the sample covariance of the rows of $\sqrt{n}\hat{X}_i$ corresponding to vertices in block 1 and 2 respectively. The estimates of the covariance matrices are given in Table 1. We see that as $n$ increases, the sample covariances tend toward the conjectured limiting covariance matrix given in the last row.

We also investigate the effects of the multivariate normal assumption in Conjecture 17 on inference procedures. It is shown in Sussman et al. [2012, 2013] that the approach of embedding a graph into some Euclidean space, followed by inference (for example, clustering or classification) in that space can be consistent. However, these consistency results are, in a sense, only first-order results. In particular, they demonstrate only that the error of the inference procedure converges to 0 as the number of vertices in the graph increases. We now illustrate how Conjecture 17 may lead to a more refined error analysis.

We construct a sequence of random graphs on $n$ vertices, where $n \in \{1000, 1250, 1500, \ldots, 4000\}$, following the stochastic blockmodel with parameters as given above in Eq. (27). For each graph $G_n$ on $n$ vertices, we embed $G_n$ and cluster the (embedded) vertices of $G_n$ via a Gaussian mixture model (GMM) and K-Means. GMM was accomplished using the MCLUST implementation [Fraley and Raftery, 1999] of the Gaussian mixture model for clustering. We then measure the classification error of the clustering solution. We repeat this procedure 100 times to obtain an estimate of the misclassification rate. The results are plotted in Figure 2. For comparison, we also plot the Bayes optimal classification error rate under the assumption that the embedded points do indeed follow a multivariate normal mixture with covariance matrices $\Sigma_1$ and $\Sigma_2$. 

Figure 1: Plot of the estimated latent positions for $n \in \{1000, 2000, 4000, 8000\}$. Dashed ellipses give the 95\% level curves for the distributions predicted in Conjecture 17.
Table 1: For each $n \in \{2000, 4000, 8000, 16000\}$, we show the sample covariance matrix for $\sqrt{n}(\hat{X}_i - X_i)$ for each block. The last line shows the conjectured covariance for the limiting distribution.

| $n$  | $\hat{\Sigma}_1$         | $\hat{\Sigma}_2$         |
|------|--------------------------|--------------------------|
| 2000 | $\begin{bmatrix} 0.58 & 0.54 \\ 0.54 & 16.56 \end{bmatrix}$ | $\begin{bmatrix} 0.58 & 0.75 \\ 0.75 & 16.28 \end{bmatrix}$ |
| 4000 | $\begin{bmatrix} 0.58 & 0.63 \\ 0.63 & 14.87 \end{bmatrix}$ | $\begin{bmatrix} 0.59 & 0.71 \\ 0.71 & 15.79 \end{bmatrix}$ |
| 8000 | $\begin{bmatrix} 0.60 & 0.61 \\ 0.61 & 14.20 \end{bmatrix}$ | $\begin{bmatrix} 0.58 & 0.54 \\ 0.54 & 14.23 \end{bmatrix}$ |
| 16000| $\begin{bmatrix} 0.59 & 0.58 \\ 0.58 & 13.96 \end{bmatrix}$ | $\begin{bmatrix} 0.61 & 0.69 \\ 0.69 & 13.92 \end{bmatrix}$ |
| $\infty$ | $\begin{bmatrix} 0.59 & 0.55 \\ 0.55 & 13.07 \end{bmatrix}$ | $\begin{bmatrix} 0.60 & 0.59 \\ 0.59 & 13.26 \end{bmatrix}$ |

Figure 2: Comparison of classification error for GMM, K-Means, Bayes optimal error rate, and STFP. The classification errors for each $n \in \{1000, 1250, 1500, \ldots, 4000\}$ were obtained by averaging 100 Monte Carlo iterations and are plotted on a log$_{10}$ scale. The plot indicates that the assumption of mixture of multivariate normals can yield non-negligible improvement in the inference procedure. The STFP curve shows an upper bound on the error rate as derived in Sussman et al. [2012].

$\Sigma_2$ as given above in the last line of Table 1. We also plot the misclassification rate of $\frac{C \log n}{n}$ as given in Sussman et al. [2012] (STFP) where the constant $C$ was chosen to match the misclassification rate of $K$-means clustering for $n = 1000$. We note that for the number of vertices considered here, i.e., for $n \leq 4000$, the upper bound for the constant $C$ from Sussman et al. [2012] will give a vacuous upper bound of the order of $10^6$ for the misclassification rate in this example.
5 Discussion

From our simulations above and previous experience, it is natural to conjecture that under the stochastic blockmodel, the estimated latent positions provided by the scaled eigenvectors should follow a multivariate normal distribution. The added value of the conjecture in Section 3 is a specification of the limiting covariance matrix. The simulations in Section 4 show that the predicted covariance of the residuals corresponds closely to the empirically observed covariance matrices. Though this is just one example, other empirical investigations of this kind have all supported our conjecture.

Our demonstration of the clustering accuracy of GMM shows how our conjecture may impact statistical inference for random graphs. First, we see that the empirical error rates are much lower than those proved in previous work on spectral methods [Rohe et al., 2011, Sussman et al., 2012, Fishkind et al., 2013]. Indeed, using both the K-Means algorithm and GMM, the average clustering error decreases at an exponential rate as opposed to the \( \log(n)/n \) bounds shown in the previous work. Furthermore, the rate of decrease for GMM clustering closely mirrors the Bayes optimal error rate that would be achieved if the estimated latent positions were exactly distributed according to the conjectured distribution and the parameters of this distribution were known.

These results suggest that further investigations using our conjecture could lead to much more accurate bounds on the empirical error rates of GMM clustering. We believe that a proof of this conjecture and further work regarding distributions of spectral statistics for stochastic blockmodels will lead to foundational statistical procedures analogous to the results on estimation, hypothesis testing, and clustering in the setting of mixtures of normal distributions in Euclidean space. As opposed to the procedures for stochastic blockmodels currently in the literature, which are often either ad hoc or highly computationally expensive, the relatively simple nature of our spectral procedure allows for computationally efficient statistical methodology.

Extensions of this work to a wider class of exchangeable graphs are also of interest. Though not all exchangeable random graphs can be represented as random dot product graphs, random dot product graphs can approximate any exchangeable graph in the following sense: given a sufficiently regular link function, there exists a feature map from the original latent position space to \( \ell_2 \), such that the link function applied to the original latent positions is equal to the inner product applied to the feature-mapped positions in \( \ell_2 \). Tang et al. [2013] argue that by increasing the dimension of the estimated latent positions, it is possible to estimate these feature-mapped latent positions in a way that allows for consistent subsequent inference. Though this larger class of models is not considered here, we believe this is strong motivation to study the random dot product graph model and its eigenvalues and eigenvectors.

References

D. J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

N. Alon, M. Krivelevich, and V. H. Vu. On the concentration of eigenvalues of random symmetric matrices. *Israel Journal of Mathematics*, 131(1):259–267, 2002.

P. J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences of the United States of America*, 106(50):21068–73, 2009.

P. J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *Annals of Statistics*, 39(5):38–59, 2011.

B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

D. S. Choi, P. J. Wolfe, and E. M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012.
F. R. K. Chung. *Spectral Graph Theory*, volume 92. American Mathematical Society, 1997.

K. L. Chung. *A course in probability theory*, volume 3. Academic Press New York, 1974.

P. Diaconis and S. Janson. Graph limits and exchangeable random graphs. *arXiv preprint arXiv:0712.2749*, 2007.

M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973.

Doniell F. Fishkind, Daniel L. Sussman, Minh Tang, Joshua T. Vogelstein, and Carey E. Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, 2013.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010. ISSN 03701573. doi: 10.1016/j.physrep.2009.11.002.

C. Fraley and A. E. Raftery. MCLUST: Software for model-based cluster analysis. *Journal of Classification*, 16:297–306, 1999.

Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1(3):233–241, 1981.

A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

D. N. Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 1979.

S. Janson. The first eigenvalue of random graphs. *Combinatorics, Probability and Computing*, 14(5-6):815–828, 2005.

A. Knowles and J. Yin. Eigenvector distribution of wigner matrices. *Probability Theory and Related Fields*, pages 1–40, 2011.

M. Krivelevich and B. Sudakov. The largest eigenvalue of sparse random graphs. *Combinatorics, Probability and Computing*, 12(01):61–72, 2003.

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

R. I. Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *Arxiv preprint ArXiv:0911.0600*, 2009.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics*, 39(4):1878–1915, 2011.

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

D. L. Sussman, M. Tang, and C. E. Priebe. Universally consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (Accepted)*, 2013.

M. Tang, D. L. Sussman, and C. E. Priebe. Universally consistent vertex classification for latent positions graphs. *Annals of Statistics (Accepted)*, 2013.
T. Tao and V. Vu. Random matrices: Universal properties of eigenvectors. *Random Matrices: Theory and Applications*, 1(01), 2012.

J. A. Tropp. Freedmans inequality for matrix martingales. *Electron. Commun. Probab*, 16:262–270, 2011.

S. Young and E. Scheinerman. Random dot product graph models for social networks. *Algorithms and models for the web-graph*, pages 138–149, 2007.