Electronic Journal of Statistics
Vol. 8 (2014) 2905–2922
ISSN: 1935-7524
DOI: 10.1214/14-EJS978

Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding*

Vince Lyzinski

Human Language Technology Center of Excellence
Johns Hopkins University
Stieff Building, 810 Wyman Park Drive, Baltimore, MD 21211, USA
e-mail: vlyzins1@jhu.edu

Daniel L. Sussman

Department of Statistics
Harvard University
One Oxford Street, Cambridge, MA, 02138, USA
e-mail: daniellsussman@fas.harvard.edu

Minh Tang, Avanti Athreya and Carey E. Priebe

Department of Applied Mathematics and Statistics
Johns Hopkins University
3400 N. Charles St, Baltimore, MD 21218, USA
e-mail: mtang10@jhu.edu; dathrey1@jhu.edu; cep@jhu.edu

Abstract: Vertex clustering in a stochastic blockmodel graph has wide applicability and has been the subject of extensive research. In this paper, we provide a short proof that the adjacency spectral embedding can be used to obtain perfect clustering for the stochastic blockmodel and the degree-corrected stochastic blockmodel. We also show an analogous result for the more general random dot product graph model.

AMS 2000 subject classifications: Primary 62C30; secondary 05C80.
Keywords and phrases: Clustering, stochastic block model, degree corrected stochastic block model.

Received September 2014.

1. Introduction

In many problems arising in the natural sciences, technology, business and politics, it is crucial to understand the specific connections among the objects under study: for example, the interactions between members of a political party; the

*Supported in part by National Security Science and Engineering Faculty Fellowship (NSSEFF), Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE), and the XDATA program of the Defense Advanced Research Projects Agency (DARPA) administered through Air Force Research Laboratory contract FA8750-12-2-0303.

firing of synapses in a neuronal network; or citation patterns in reference literature. Mathematically, these objects and their connections are modeled as graphs, and a common goal is to find clusters of similar vertices within a graph.

Both model-based and heuristic-based techniques have been proposed for clustering the vertices in a graphs [2, 5, 14, 19]. In this paper we focus on probabilistic performance guarantees for spectral-based techniques which have elements of both model- and heuristic-based methods [18, 20]. We study the consistency of mean squared error clustering via the adjacency spectral embedding for three nested classes of models, each an examples of latent position models [7]:

- the stochastic blockmodel where vertices in the same cluster are stochastically equivalent [8],
- the degree-corrected stochastic blockmodel where stochastic equivalence holds up to a scaling factor [9],
- and the random dot product graph where a natural vertex clustering may not exist [27].

The generality of our main result allows for the extension of our asymptotically error-free results from the rather restrictive stochastic blockmodel to more general settings.

Numerous spectral clustering procedures have been proposed and analyzed under various random graph models [4, 10, 17, 18, 20]. For example, Laplacian spectral embedding [18] and adjacency spectral embedding [20] have been shown to yield consistent clustering for the stochastic blockmodel. These results have relied on bounding the Frobenius norm difference between the embedded vertices and associated eigenvectors of the population Laplacian (in Laplacian spectral embedding) or edge probability matrix (in adjacency spectral embedding).

Relying on global Frobenius norm bounds for demonstrating consistent clustering is suboptimal, however, because in general, one cannot rule out that a diminishing but positive proportion of the embedded points contribute disproportionately to the global error. When this occurs, these "outliers" are very likely to be misclustered, and hence the best existing bounds on the Frobenius norm show that at most $O(\log(n))$ vertices will be misclustered (see [18, Theorem 3.1] and [20, Theorem 1]).

In contrast, our main technical result gives a bound (in probability) on the maximum error between *individual* embedded vertices and the associated eigenvectors of the edge probability matrix (see Lemma 5). This lemma is proved for general random dot product graphs and provides the necessary tools to improve the bounds on the error rate of mean squared error clustering in adjacency spectral embedding. The first main clustering result of this paper gives a bound on the probability that a mean square error clustering of the adjacency spectral embedding will be *error-free*, i.e. zero vertices will be misclustered (see Theorem 6).

Due to the generality of our main lemma, we are able to prove an analogous asymptotically error free clustering result in the degree-corrected stochastic blockmodel (see Theorem 13). Note again that the best existing results for

spectral methods in the degree-corrected model assert that at most $O(\log(n))$ vertices will be misclustered [17, Theorem 4.4]. Finally, we prove a very general result that spectral clustering of random dot product graphs is strongly universally consistent in the sense of [16] (see Theorem 15). These extensions underly the wide utility of our approach, and we believe our main lemma to be of independent interest.

We note that the authors of [3], among others, have shown that likelihood-based techniques can be employed to achieve asymptotically error-free clustering in the stochastic blockmodel. However, likelihood based approaches are computationally intractable for very large graphs compared to our present spectral clustering approach.

2. Setting and main theorem

In the first part of this section, we will define the random dot product graph and our main tool, the adjacency spectral embedding. Next, we define the stochastic blockmodel and clustering procedure, and finally, we will state our main theorem and the supporting lemmas.

2.1. Random dot product graphs and the adjacency spectral embedding

The random dot product graph model is a convenient theoretical tool, and spectral properties of the adjacency matrix is well understood. While the stochastic blockmodel relies on an inherently non-geometric construction—indeed, each block in associated with a categorical label, and these labels determine the adjacency probabilities—the random dot product graph relies on a geometric construction in which each block is associated with a point in Euclidean space, i.e. a vector. The dot products of these vectors then determine the adjacency probabilities in the graph.

Definition 1 (Random Dot Product Graph (RDPG))**.** A random adjacency matrix $A \sim \text{RDPG}(X)$ for $X = [X_1, \ldots, X_n]^\top \in \mathcal{X}_d$ where

$$\mathcal{X}_d = \{Z \in \mathbb{R}^{n\times d} : ZZ^\top \in [0,1]^{n\times n}, \text{rank}(Z) = d\}$$

is said to be an instance of a random dot product graph (RDPG) if

$$\mathbb{P}[A|X] = \prod_{i>j} (X_i^\top X_j)^{A_{ij}} (1 - X_i^\top X_j)^{1-A_{ij}}.$$

Remark. In general we will denote the rows of an $n \times d$ matrix M by $M_i^\top$. With this notation, in the above definition $P(A_{ij} = 1) = X_i^\top X_j$ for all $1 \le i < j \le n$. We then define $P = XX^\top$, so that the entries of P give the Bernoulli parameters for edge probabilities.

Note that, as defined above, the rank of P is d. Let $P = [V|\widetilde{V}][S \oplus \widetilde{S}][V|\widetilde{V}]^T$ be the spectral decomposition of P where $[V|\widetilde{V}] \in \mathbb{R}^{n\times n}$ is orthogonal, $V \in \mathbb{R}^{n\times d}$ has orthonormal columns, $S \in \mathbb{R}^{d\times d}$ is diagonal with

$$S(1,1) \geq S(2,2) \geq \cdots \geq S(d,d) > 0 \text{ and } \widetilde{S} = 0.$$

Importantly, we shall assume throughout this paper that the non-zero eigenvalues of P are distinct, i.e., the inequalities above are strict.

It follows that there exists an orthonormal $W \in \mathbb{R}^{d\times d}$ such that $VS^{1/2} = XW$. We thus suppose that $X = VS^{1/2}$; the assumption does not lead to any loss of generality because the distribution of A is invariant under orthogonal transformations of the latent positions and the clustering method considered in this paper is invariant under orthogonal transformations. This relationship between the spectral decomposition of P and the latent positions X for the RDPG model motivates our main tool: the adjacency spectral embedding.

Definition 2 (Adjacency Spectral Embedding (ASE))**.** Let $\hat{V} \in \mathbb{R}^{n\times d}$ have orthonormal columns given by the eigenvectors of A corresponding to the d largest eigenvalues of A according to the algebraic ordering. Let $\hat{S} \in \mathbb{R}^{d\times d}$ be diagonal with diagonal entries given by these eigenvalues in descending order. We define the d-dimensional *adjacency spectral embedding* of A via $\hat{X} = \hat{V}\hat{S}^{1/2}$.

We shall assume, for ease of exposition, that the diagonal entries of $\hat{S}$ are positive. As will be seen later, e.g., Lemma 8, this assumption is justified in the context of random dot product graphs due to the concentration of the eigenvalues of A around those of P. Recall that the rows of $\hat{X}$ will be denoted by $\hat{X}_i^\top$.

2.2. Clustering

We begin by considering the task of clustering in the K-block stochastic blockmodel. This model is typically parameterized by a $K \times K$ matrix of probabilities of adjacencies between vertices in each of the blocks along with the block memberships for each vertex. Here we present an alternative definition in terms of the RDPG model.

Definition 3 ((Positive Semidefinite) K-block Stochastic Blockmodel (SBM))**.** We say an RDPG is an SBM with K blocks if the number of distinct rows in X is K. In this case, we define the block membership function $\tau : [n] \mapsto [K]$ to be a function such that $\tau(i) = \tau(j)$ if and only if $X_i = X_j$. For each $k \in [K]$, let n_k be the number of vertices such that $\tau_i = k$, i.e. the number of vertices in block k.

To ease notation, we will always use K to denote the number of blocks in an SBM, and we will refer to a K-block SBM as simply an SBM when appropriate.

Remark. Note that a general K-block SBM can only be represented in this way if the $K \times K$ matrix of probabilities is positive semidefinite.

Next, we introduce mean square error clustering, which is the clustering sought by K-means clustering.

Definition 4 (Mean Square Error (MSE) Clustering)**.** The MSE clustering of the rows of $\hat{X}$ into K blocks returns

$$\hat{C} := \min_{C \in \mathcal{C}_K} \|C - \hat{X}\|_F, \text{ where}$$
$$\mathcal{C}_K = \{C \in \mathbb{R}^{n \times d} : C \text{ has } K \text{ distinct rows}\},$$

are the optimal cluster centroids for the MSE clustering. We also define the cluster membership function $\hat{\tau} : [n] \mapsto [K]$, which satisfies $\hat{\tau}_i = \hat{\tau}_j$ if and only if $\hat{C}_i = \hat{C}_j$, where $\hat{C}_i^\top$ is the i^{th} row of $\hat{C}$.

In our results, we consider the MSE clustering of the rows of $\hat{X}$ in two contexts, the SBM and RDPG models defined above. We will also consider a variation of the SBM, the degree-corrected SBM, in which we perform MSE clustering on the rows of a projected version of $\hat{X}$.

2.3. Main theorems

Before stating our main results, we indicate our notation for matrix norms and we define constants to be used throughout the remaining text. For a matrix $M \in \mathbb{R}^{n \times d}$ we let $\|M\|_{2\to\infty} = \max_i \|M_i\|_2$, i.e. the maximum of the Euclidean norm of the rows. For a square matrix $M \in \mathbb{R}^{n \times n}$, $\|M\|_2 = \sqrt{\text{largest eigenvalue of } M^\top M}$ denotes the spectral norm. The Frobenius norm of a matrix is denoted by $\|\cdot\|_F$.

We define:

- $\Delta = \max_i \sum_{j \neq i} P_{ij}$ is the maximum of the row sums of P;
- d is the rank of P;
- $\gamma n = \min_{1 \leq i \leq d} |S(i+1, i+1) - S(i,i)| > 0$ is the minimum gap among the distinct eigenvalues of P;
- K is the number of blocks in the SBM.

Note, γn is not necessarily equal to the magnitude of the smallest non-zero eigenvalue of P, as the gaps between consecutive non-zero eigenvalues could be smaller. We now state a technical but highly useful lemma in which we bound the maximum difference between the rows of $\hat{X}$ and the rows of an orthogonal transformation of X.

Lemma 5. *Suppose* $0 < \eta < 1/2$ *is given such that* $\gamma n \geq 4\sqrt{\Delta \log(n/\eta)}$*. Then, with probability at least* $1 - 2\eta$*, one has*

$$\|\hat{X} - X\|_{2\to\infty} \leq \frac{85 d \Delta^3 \log(n/\eta)}{(\gamma n)^{7/2}}. \tag{1}$$

Remark. The parameter Δ is the maximum expected degree for the random graph. In many models, γ will be of the same order of the density of the graph. If the density is very small, the bound in the Eq. (1) will be large. Frequently Δ and

γn are of the same order, so that the bound in Eq. (1) is of order $O(\log(n)/\sqrt{\gamma n})$ for d fixed and η decaying polynomially. See Example 1 for a simple illustration of how this bound can be applied. Finally, note that we always have $\Delta > \gamma n$ so that the condition in Lemma 5 implies that $\Delta > 16\log(n/\eta)$, which allows us to use the results of [15, 25].

Lemma 5 gives far greater control of the errors than the previous results that were derived for the Frobenius norm $\|\hat{X} - X\|_F$; indeed, the latter bounds do not allow fine control of the errors in the individual rows of $\hat{X}$, and therefore can only bound the number of mis-clustered vertices via $O(\log n)$. Lemma 5, on the other hand, provides exactly this control and, as such, vastly improves the bounds on the error rate of MSE clustering of $\hat{X}$.

Our main theorem is the following result on the probability that mean square error clustering on the rows of $\hat{X}$ is error-free.

Theorem 6 (SBM)**.** *Let $A \sim \text{RDPG}(X)$ be an SBM with K blocks and block membership function τ and suppose $\eta \in (0, 1/2)$. Assume that*

(A0) *the non-zero eigenvalues of $P = XX^\top$ are distinct.*

Denote the bound on $\|\hat{X} - X\|_{2\to\infty}$ in Lemma 5 as $\beta = \beta(d, n, \eta, \gamma)$. Let $\hat{\tau} : [n] \to [K]$ be the optimal MSE clustering of the rows of $\hat{X}$ into K clusters. Let S_K denote the symmetric group on K, and $\pi \in S_K$ a permutation of the blocks. Finally, let $n_{min} = \min_{k\in[K]} n_k$ be the smallest block size. If

(A1) *for all $i, j \in [n]$ if $X_i \neq X_j$ then $\|X_i - X_j\|_2 > 4\beta\sqrt{n/n_{min}}$ and*
(A2) *the eigenvalue gap satisfies $\gamma n > 4\sqrt{\Delta \log(n/\eta)}$,*

then with probability at least $1 - 2\eta$,

$$\min_{\pi\in S_K} |\{i \in [n] : \tau(i) \neq \pi(\hat{\tau}(i))\}| = 0.$$

We remark that assumptions (A1) and (A2) are quite natural: (A1) requires that the rows of X with distinct entries have some minimum separation that is large enough compared to the ratio of the number of vertices to the smallest block size and compared to the bound in Lemma 5. Assumption (A2) on γ ensures a large enough gap in the eigenvalues to use Lemma 5. We note that Lemma 5 is applicable to the sparse setting, i.e., the setting wherein the average degrees of the vertices are of order $\omega(\log^k n)$ for some $k \geq 2$, but that we still need sufficient separation between the distinct rows of X. For a simple illustration of how this theorem can be applied for a concrete model see Example 1. Finally, we admit that assumption (A0) is less natural, but it is a helpful technical restriction and it excludes a small range of parameters.

While Theorem 6 is proven in the SBM setting, we note that our final theorem, Theorem 15, is an analogous clustering result in which we prove strong universal consistency of MSE clustering for more general random dot product graphs.

Finally, we observe that Theorem 6 has both finite-sample and asymptotic implications. In particular, under these model assumptions, for any finite n, the

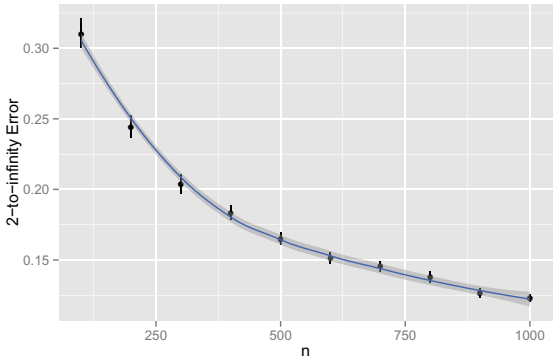

FIG 1. *Mean with standard error bars of* $\|\hat{X} - X\|_{2\to\infty}$ *for each value of* n *for the model in Example 1. The decay in the error is very close to* $O(n^{-1/2})$.

theorem gives a lower bound on the probability of perfect clustering. We do *not* assert—and indeed it is easy to refute—that in the finite sample case, perfect clustering occurs with probability one. Nevertheless, we can choose $\eta = n^{-c}$ for some constant $c \geq 2$, in which case the probability of perfect clustering approaches one as n tends to infinity.

Example 1 (Dense SBM)**.** Here we consider a simple concrete example where we can apply Lemma 5 and Theorem 6. Let $\nu_1 = (.5, .4)^\top$ and $\nu_2 = (0.5, -0.4)^T$ and let $X_i = \nu_1$ for $i = 1, \ldots, n/2$ and $X_i = \nu_2$ for $i = n/2 + 1, \ldots, n$. Hence, this is a two block model and the 2×2 matrix of edge probabilities is given by

$$B = \begin{pmatrix} 0.41 & 0.09 \\ 0.09 & 0.41 \end{pmatrix}.$$

The constants in our theorem are $d = 2$, $\delta = 0.41n + 0.09n = n/2$. The distinct eigenvalues of P are $n/2, 0.32n$, and 0 so the smallest gap is $0.18n$; hence $\gamma = 0.18$. Lemma 5 can be applied as long as $0.18n \geq 4\sqrt{0.5n \log(n/\eta)}$, which will clearly hold for n sufficiently large for any fixed $\eta \in (0, 1/2)$. This also establishes assumptions (A0) and (A1) of Theorem 6.

The implication of Lemma 5 is that

$$\|\hat{X} - X\|_{2\to\infty} \leq \frac{85 \cdot 2 \cdot (0.5n)^3 \log(n/\eta)}{(0.18n)^{7/2}} \approx \frac{8588 \log(n/\eta)}{\sqrt{n}}.$$

While this bound is loose for small to moderate n, the asymptotic implications are clear. Empirically, Figure 1 shows the average $\|\cdot\|_{2\to\infty}$ error for this model as a function of n and we see that the error becomes small much sooner and decays at a rate very close to $O(n^{-1/2})$.

For Theorem 6 we can compute that if $X_i \neq X_j$ then $\|X_i - X_j\|_2 = 0.8$ for all n and hence since $n/n_{min} = 2$ and $\beta = O(\log(n)/\sqrt{n})$, the assumption (A1) will hold for n sufficiently large. Hence, for large enough n, there will be a very high probability that the mean square error clustering will provide perfect performance.

Example 2 (Sparse SBM)**.** In this example, we will illustrate some asymptotic implications of the assumptions of Theorem 6 in a generalization of Example 1. The SBM of the previous example is a specific instance of an SBM with edge probabilities given by

$$B = c\begin{pmatrix} a & b \\ b & a \end{pmatrix}, \tag{2}$$

where $\tau_i = 1$ for $i \in [n/2]$, and $\tau_i = 2$ otherwise. In order for P to be positive semidefinite we need $a > b$, and the subsequent RDPG representation is

$$\nu_1 = \frac{\sqrt{c}}{2}\begin{pmatrix} \sqrt{a+b}+\sqrt{a-b} \\ \sqrt{a+b}-\sqrt{a-b} \end{pmatrix} \text{ and } \nu_2 = \frac{\sqrt{c}}{2}\begin{pmatrix} \sqrt{a+b}-\sqrt{a-b} \\ \sqrt{a+b}+\sqrt{a-b} \end{pmatrix}. \tag{3}$$

We will investigate for what values of a, b and c the assumptions of our theorem are satisfied, noting first that (A0) is automatically satisfied. For the key constants, we can work out that the distinct eigenvalues of P are $nc(a+b)/2$, $nc(a-b)/2$, and 0. This gives $\gamma = c\min\{b,(a-b)/2\}$. In addition, $\Delta = nc(a+b)/2$, so that assumption (A2) reads

$$nc\min\{b,(a-b)/2\} > 4\sqrt{nc(a+b)\log(n/\eta)/2}.$$

We restrict ourselves to the sparse domain, and assume that $c = 1/n$, $a = o(n)$, and $b = o(n)$. Assumption (A2) then becomes

$$\frac{\min\{b^2,(a-b)^2/4\}}{a+b} > 8\log(n/\eta).$$

Here, we have $\|\nu_1 - \nu_2\|_2 = \sqrt{2(a-b)/n}$, and assumption (A1) in this regime is equivalent to

$$\frac{\sqrt{a-b}}{(a+b)^3}\left(\min\{b,(a-b)/2\}\right)^{7/2} > 85\sqrt{n}\log(n/\eta).$$

Highlighting a few special cases, we consider

1. $a = O(1)$ and $b = O(1)$. In this case, our assumptions do not hold. Indeed, in [13] it is established that if $(a-b)^2 < 2(a+b)$, then clustering is impossible, and the same authors recently extended this result in [12] to show that consistent estimation is impossible for any choice of $a = O(1)$ and $b = O(1)$.
2. $b = \Theta(a) = \Theta(a-b)$. In order to satisfy assumptions (A1) and (A2), it suffices that $b = \omega(\sqrt{n}\log(n))$, and (A1) does not hold if $b = o(\sqrt{n}\log(n))$.
3. $b = o(a)$. In order to satisfy assumptions (A1) and (A2), it suffices that $b/\sqrt{a} = \omega(\sqrt{\log(n)})$ and $b^{7/2}/a^{5/2} = \omega(\sqrt{n}\log(n))$. Note that (A2) does not hold if $b/\sqrt{a} = o(\sqrt{\log(n)})$, and (A1) does not hold if $b^{7/2}/a^{5/2} = o(\sqrt{n}\log(n))$.
4. $a - b = o(b)$. In order to satisfy assumptions (A1) and (A2), it suffices that $a - b = \omega(\sqrt{\log(n)b})$ and $a - b = \omega(n^{1/8}\log(n)^{1/4}b^{3/4})$. Note that (A2) does not hold if $a - b = o(\sqrt{\log(n)b})$, and (A1) does not hold if $a - b = o(n^{1/8}\log(n)^{1/4}b^{3/4})$.

3. Proof of Theorem 6

Before we prove Theorem 6, we first collect a sequence of useful bounds from [20, 23, 25]. We then prove two key lemmas.

Proposition 7. *Suppose $A \sim \mathrm{RDPG}(X)$ with $X \in \mathcal{X}_d$ and let Δ and γ be as defined in Lemma 5. For any $\eta \in (0, 1/2)$, if $\gamma n > 4\sqrt{\Delta \log(n/\eta)}$, then the following occur with probability at least $1 - \eta$*

$$\|A - P\|_2 \leq 2\sqrt{\Delta \log(n/\eta)},$$
$$\|\hat{V} - V\|_F^2 \leq 4d\, \frac{\Delta \log(n/\eta)}{\gamma^2 n^2}.$$

In addition, as P is a non-negative matrix, $\|S\|_2 \leq \Delta$. Thus, if $\Delta \geq 16 \log(n/\eta)$, then provided that the above events occur, $\|\hat{S}\|_2 \leq \min\{2\Delta, n\}$.

The next two lemmas from [1] are essential to our argument.

Lemma 8 (Athreya et al. [1])**.** *In the setting of Proposition 7, if the events in Proposition 7 occur, then*

$$\|\hat{S} - S\|_2 \leq 18d \frac{\Delta^2 \log(n/\eta)}{\gamma^2 n^2}.$$

Lemma 9 (Athreya et al. [1])**.** *In the setting of Proposition 7, if the events in Proposition 7 occur, then*

$$\|V^\top \hat{V} - I\|_F \leq \frac{10d\Delta \log(n/\eta)}{\gamma^2 n^2}.$$

We then have the following bound

Lemma 10. *In the setting of Proposition 7, if the events in Proposition 7 occur, then*

$$\|AV\hat{S}^{-1/2} - \hat{X}\|_F \leq \frac{24\sqrt{2}d\Delta^2 \log(n/\eta)}{(\gamma n)^{5/2}}.$$

Proof. Let $E = A - \hat{V}\hat{S}\hat{V}^\top$. Denoting by Z the quantity $AV\hat{S}^{-1/2}$, we have

$$\begin{aligned}\|Z - \hat{X}\|_F &= \|AV\hat{S}^{-1/2} - \hat{V}\hat{S}^{1/2}\|_F \\ &= \|A(V - \hat{V})\hat{S}^{-1/2}\|_F \\ &= \|(\hat{V}\hat{S}\hat{V}^\top + E)(V - \hat{V})\hat{S}^{-1/2}\|_F \leq C_1 + C_2\end{aligned}$$

where C_1 and C_2 are given by

$$C_1 = \|\hat{S}\|_2 \|\hat{V}^\top (V - \hat{V})\|_F \|\hat{S}^{-1/2}\|_2; \quad C_2 = \|E\|_2 \|V - \hat{V}\|_F \|\hat{S}^{-1/2}\|_2.$$

Note that by Proposition 7 and our assumption that $\gamma\sqrt{n} \geq 4\sqrt{\log(n/\eta)}$,

$$\|\hat{S}^{-1}\|_2 \leq (\gamma n - 2\sqrt{n \log n/\eta})^{-1} \leq \frac{2}{\gamma n}.$$

Combining the previous displayed equation and Lemma 9 yields

$$C_1 \leq \frac{20\sqrt{2}d\Delta^2 \log(n/\eta)}{(\gamma n)^{5/2}}.$$

Similarly, we have that $\|E\|_{2\to 2} \leq 2\sqrt{\Delta \log(n/\eta)}$, and combining this with Proposition 7, we bound C_2 by

$$C_2 \leq \frac{4\sqrt{2}d\Delta \log(n/\eta)}{(\gamma n)^{3/2}},$$

from which the desired bound follows. □

We now use Lemma 8, Lemma 10 and Hoeffding's inequality to prove Lemma 5. We note that for any matrices $A \in \mathbb{R}^{k_1\times k_2}$ and $B \in \mathbb{R}^{k_2\times k_2}$, $\|AB\|_{2\to\infty} \leq \|A\|_{2\to\infty}\|B\|_2$.

Proof of Lemma 5. Since $X = PVS^{-1/2}$ we can add and subtract the matrix $AV\hat{S}^{-1/2}$ and $AVS^{-1/2}$ to rewrite $\hat{X} - X$ as

$$\hat{X} - X = A(\hat{V} - V)\hat{S}^{-1/2} + AV(\hat{S}^{-1/2} - S^{-1/2}) + (A - P)VS^{-1/2}.$$

Lemma 10 bounds the first term in terms of the Frobenius norm which is a bound for the $2 \to \infty$ norm. For the second term, we have

$$\hat{S}^{-1/2} - S^{-1/2} = (S - \hat{S})(S^{1/2} + \hat{S}^{1/2})^{-1}(\hat{S}^{-1/2}S^{-1/2}),$$

as both $\hat{S}$ and S are diagonal matrices. Applying Lemma 8 thus yields

$$\begin{aligned}\|\hat{S}^{-1/2} - S^{-1/2}\|_2 &\leq \|S - \hat{S}\|_2(3/2\sqrt{\gamma n})(1/2\gamma n)\\ &\leq 24d\frac{\Delta^2 \log(n/\eta)}{(\gamma n)^{7/2}},\end{aligned}$$

and hence

$$\begin{aligned}\|AV(\hat{S}^{-1/2} - S^{-1/2})\|_{2\to\infty} &\leq \|AV\|_{2\to\infty}\|\hat{S}^{-1/2} - S^{-1/2}\|_2\\ &\leq 48d\frac{\Delta^3 \log(n/\eta)}{(\gamma n)^{7/2}}.\end{aligned}$$

We now bound the third term. Let Z_{ij} denote the i,jth entry, and $Z_i^\top$ the ith row, of the $n \times d$ matrix $(A - P)V$. Observe that

$$\|(A - P)V\|_{2\to\infty} = \max_i \|Z_i\|_2 \leq \sqrt{d}\max_{i,j}|Z_{i,j}|.$$

Next, since

$$Z_{ij} = \sum_{k=1}^{n}(A_{ik} - P_{ik})(V)_{kj},$$

we see that Z_{ij} is a sum of n independent, mean zero random variables $R_k(ij) = (A_{ik} - P_{ik})(V)_{kj}$, and $|R_k(ij)| \leq |(V)_{kj}|$. Since V has orthonormal columns, $\sum_k (V)_{kj}^2 = 1$. Therefore, Hoeffding's inequality implies

$$\mathbb{P}\Big(|Z_{ij}| > \sqrt{\tfrac{1}{2}\log(2nd/\eta)}\ \Big) \leq \frac{\eta}{nd}.$$

Since there are nd entries Z_{ij}, a simple union bound ensures that

$$\mathbb{P}\Big(\max_{i,j}|Z_{ij}| > \sqrt{\tfrac{1}{2}\log(2nd/\eta)}\ \Big) \leq \eta,$$

and consequently that

$$\mathbb{P}\Big(\|(A-P)V\|_{2\to\infty} > \sqrt{\tfrac{d}{2}\log(2nd/\eta)}\ \Big) \leq \eta.$$

The third term can therefore be bounded as

$$\|(A-P)VS^{-1/2}\|_{2\to\infty} \leq \|(A-P)V\|_{2\to\infty}\|S^{-1/2}\|_2 \leq \sqrt{\frac{d\log(2nd/\eta)}{2\gamma n}}$$

with probability at least $1-\eta$. Combining the bounds for the above three terms yields Lemma 5. □

Proof of Theorem 6. We assume that the event in Lemma 5 occurs and show that this implies the result. Since X has K distinct rows, it follows that

$$\|\hat{C} - \hat{X}\|_F \leq \|X - \hat{X}\|_F \leq \beta\sqrt{n}.$$

Let $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_K$ be L^2-balls with radii $2r$ around the K distinct rows of X. By the assumptions in Theorem 6, these balls are disjoint. Suppose there exists $k \in [K]$ such that $\mathcal{B}_k$ does not contain any rows of $\hat{C}$. Then $\|\hat{C} - X\|_F > 2r\sqrt{n_{min}}$, as for each k, no row of $\hat{C}$ is within $2r$ of the (at least n_{min}) rows of X in $\mathcal{B}_k$. This implies that

$$\begin{aligned}\|\hat{C} - \hat{X}\|_F &\geq \|\hat{C} - X\|_F - \|\hat{X} - X\|_F \\ &> 2r\sqrt{n_{min}} - \beta\sqrt{n} \\ &> 2\beta\sqrt{\frac{n}{n_{min}}}\sqrt{n_{min}} - \beta\sqrt{n} = \beta\sqrt{n},\end{aligned}$$

a contradiction. Therefore, $\|\hat{C} - X\|_{2\to\infty} \leq 2r$. Hence, by the pigeonhole principle, each ball $\mathcal{B}_k$ contains precisely one distinct row of $\hat{C}$.

If $X_i = X_j$, then both $\hat{C}_i$ and $\hat{C}_j$ are elements of $\mathcal{B}_{\tau(i)}$, and since there is exactly one distinct row of $\hat{C}$ in $\mathcal{B}_k$, $\hat{C}_i = \hat{C}_j$. Conversely, if $\hat{C}_i \neq \hat{C}_j$, then X_i and X_j are in disjoint balls $\mathcal{B}_k$ and $\mathcal{B}_{k'}$ for some $k, k' \in [K]$, implying that $X_i \neq X_j$. Thus, $X_i = X_j$ if and only if $\hat{C}_i = \hat{C}_j$, proving the theorem. □

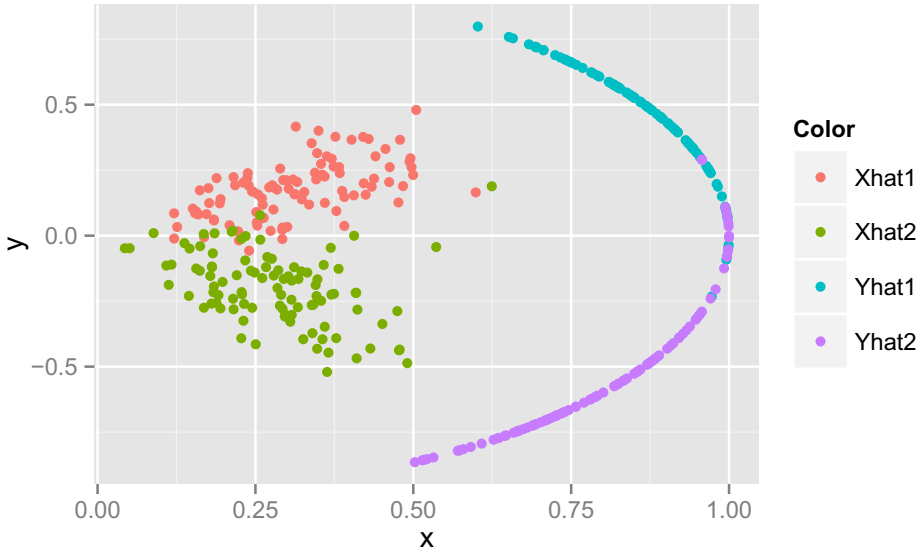

FIG 2. *Example of $\hat{X}$ (the red and green points) and $\hat{Y}$ (the blue and purple points) for a 2-block DCSBM with latent positions $y_1 = [1/5, 2\sqrt{6}/5]$ and $y_2 = [2\sqrt{6}/5, 1/5]$ and 100 vertices in each block. The c_i's are i.i.d. Uniform(0.2,0.5).*

4. Degree corrected SBM

In this section we extend our results to the degree corrected SBM [9].

Definition 11 (Degree Corrected Stochastic Blockmodel (DCSBM))**.** We say an RDPG is a DCSBM with K blocks if there exist K unit vectors $y_1, \ldots, y_K \in \mathbb{R}^d$ such that for each $i \in [n]$, there exists $k \in [K]$ and $c_i \in (0, 1)$ such that $X_i = c_i y_k$.

Remark. This model is inherently more flexible than the standard SBM because it allows for vertices within each block/community to have different expected degrees. This flexibility has made it a popular choice for modeling network data [9].

For this model, we introduce $Y \in \mathbb{R}^{n \times d}$ via $Y_i^\top = y_{\tau(i)}^\top$, so that each row of Y has unit L^2-norm. As demonstrated in [17], a key to spectrally clustering DCSBM graphs is to project the spectral embedding onto the unit sphere, yielding an estimate of Y rather than an estimate of X. As such, let $\hat{Y} = \text{diag}(\hat{X}\hat{X}^\top)^{-1/2}\hat{X}$ where diag$(\cdot)$ denotes the operation of setting all the off-diagonal elements of the argument to 0. If we denote the unit sphere in $\mathbb{R}^d$ by $\mathcal{S} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$, then $\hat{Y}$ is the projection of $\hat{X}$ on $\mathcal{S}$. See Figure 2 for a simple example of this projection step.

Our next lemma is the analogue of Lemma 5 in the DCSBM setting, allowing us to tightly control the errors in the individual rows of $\hat{Y}$.

Lemma 12. *In the setting of Lemma 5, let the $n \times d$ matrices $\widetilde{Y}, \hat{Y} \in \mathcal{S}$ be the projections of X and $\hat{X}$, respectively, onto $\mathcal{S}$. Let $c_{min} = \min_{i\in[n]} \|X_i\|_2$. If $\|\hat{X} - X\|_{2\to\infty} \leq \beta$, then*

$$\|\hat{Y} - \widetilde{Y}\|_{2\to\infty} \leq \frac{2\beta}{c_{min}}.$$

Proof. We have $\widetilde{Y}_i = (\|X_i\|_2)^{-1} X_i$ and $\hat{Y}_i = (\|\hat{X}_i\|_2)^{-1}\hat{X}_i$. Straightforward calculations then yield

$$\begin{aligned}
\|\widetilde{Y}_i - \hat{Y}_i\|_2 &= \Big\| \frac{X_i}{\|X_i\|_2} - \frac{\hat{X}_i}{\|\hat{X}_i\|_2} \Big\|_2 \\
&= \Big\| \frac{X_i}{\|X_i\|_2} - \frac{\hat{X}_i}{\|X_i\|_2} + \frac{\hat{X}_i}{\|X_i\|_2} - \frac{\hat{X}_i}{\|\hat{X}_i\|_2} \Big\|_2 \\
&\leq \frac{\|X_i - \hat{X}_i\|_2}{\|X_i\|_2} + \|\hat{X}_i\|_2 \Big(\frac{1}{\|X_i\|_2} - \frac{1}{\|\hat{X}_i\|_2} \Big) \leq 2\frac{\|X_i - \hat{X}_i\|_2}{\|X_i\|_2} \leq \frac{2\beta}{c_{\min}}
\end{aligned}$$

as desired. □

As in Theorem 6, this allows us to bound the probability of error-free MSE clustering.

Theorem 13 (Degree-corrected SBM)**.** *Suppose $A \sim \mathrm{RDPG}(X)$ and is a DCSBM with K blocks and block membership function τ and suppose $\eta \in (0, 1/2)$. Let $y_1, \ldots, y_K$ be the K unit vectors for the DCSBM and let c_{min} denote the smallest scaling factor. Let γ, β be as in Theorem 6. Suppose $r > 0$ is such that for all $i, j \in [K]$, $\|y_i - y_j\|_2 > 4r$. Let $\hat{\tau} : [n] \to [K]$ be the optimal MSE clustering of the rows of $\hat{Y}$, the projection of $\hat{X}$ onto $\mathcal{S}$, into K clusters. Finally, let $n_{min} = \min_{k\in[K]} n_k$ be the smallest block size. If*

$$r > (2\beta\sqrt{n/n_{min}})/c_{min} \text{ and } \gamma n > 4\sqrt{\Delta \log(n/\eta)},$$

then with probability at least $1 - 2\eta$,

$$\min_{\pi\in S_K} |\{i \in [n] : \tau(i) \neq \pi(\hat{\tau}(i))\}| = 0.$$

The proof of this theorem follows *mutatis mutandis* from the proof of Theorem 6.

5. Strong universal consistency

We next show how our methodology can be used to prove strong universal consistency of K-means clustering (as considered in [16]) in the general RDPG setting. Specifically, suppose that $\{X_1, \ldots, X_n\}$ is a sample of independent observations from some common compactly supported distribution F on $\mathbb{R}^d$. Denote by F_n the empirical distribution of the $\{X_i\}$, and let C be a set containing K or fewer

points. Suppose that $\phi\colon [0,\infty) \mapsto [0,\infty)$ is a continuous, nondecreasing function with $\phi(0) = 0$. Now define $\Phi(C, F_n)$ and $\Phi(C, F)$ by

$$\Phi(C, F_n) = \int \big(\min_{c\in C} \phi(\|x - c\|)\big) dF_n(x),$$
$$\Phi(C, F) = \int \big(\min_{c\in C} \phi(\|x - c\|)\big) dF(x).$$

The problem of K-means mean square error clustering given $\{X_1, \ldots, X_n\}$ can then be viewed as the minimization of $\Phi(A, F_n)$ for $\phi(r) = r^2$ over all sets C containing K or fewer elements. The strong consistency of K-means clustering corresponds then to the following statement.

Theorem 14 ([16]). *Suppose that for each $k = 1, \ldots, K$, there is a unique set $\bar{C}_k$ for which*

$$\Phi(\bar{C}_k, F) = \inf_{C\colon |C|=k} \Phi(C, F).$$

For any given $\{X_1, \ldots, X_n\}$, denote by C_n a minimizer of $\Phi(C, F_n)$ over all sets C containing K or fewer elements. Then $C_n \to \bar{C}_K$ almost surely and $\Phi(C_n, F_n) \to \Phi(\bar{C}_K, F)$ almost surely.

We now state the counterpart to Theorem 14 for the RDPG setting.

Theorem 15 (RDPG). *Let $A \sim \mathrm{RDPG}(X)$, where the latent positions are sampled from some common compactly supported distribution F. Let $\hat{F}_n$ be the empirical distribution of the $\{\hat{X}_i\}_{i=1}^n$. Denote by $\hat{C}_n$ a minimizer of $\Phi(C, \hat{F}_n)$ over all sets C containing K or fewer elements. Then provided that the conditions in Theorem 14 holds for F, $\hat{C}_n \to \bar{C}_K$ almost surely, and furthermore, $\Phi(\hat{C}_n, \hat{F}_n) \to \Phi(\bar{C}_K, F)$ almost surely.*

Proof. We can suppose, without loss of generality, that F is a distribution on a totally bounded set, say Ω. Let $\mathcal{G}$ denote the family of functions of the form $g_C(x) = \min_{c\in C} \phi(\|x - c\|_2)$ where C ranges over all subsets of Ω containing K or fewer points. The theorem is equivalent to showing

$$\sup_{g\in\mathcal{G}} \left| \int g \, d\hat{F}_n - \int g \, dF \right| \xrightarrow{\text{a.s.}} 0.$$

By Theorem 14, we know that

$$\sup_{g\in\mathcal{G}} \left| \int g \, dF_n - \int g \, dF \right| \xrightarrow{\text{a.s.}} 0.$$

and so the theorem holds provided that

$$\sup_{g\in\mathcal{G}} \left| \int g \, d\hat{F}_n - \int g \, dF_n \right| = \sup_{g\in\mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} \big\{ \min_{c\in C} \phi(\|\hat{X}_i - c\|_2) - \min_{c\in C} \phi(\|X_i - c\|_2) \big\} \right| \xrightarrow{\text{a.s.}} 0.$$

Let s_i denote the summand in the above display. We then have the following bound

$$|s_i| \le \max_{c\in C} \{ \big| \phi(\|\hat{X}_i - c\|_2) - \phi(\|X_i - c\|_2) \big| \};$$

and hence

$$\begin{aligned}\Big|\frac{1}{n}\sum_{i=1}^{n} s_i\Big| &\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{c\in C}|\phi(\|\hat{X}_i - c\|_2) - \phi(\|X_i - c\|_2)| \\ &\leq K \max_i \max_{c\in C} |\phi(\|\hat{X}_i - c\|_2) - \phi(\|X_i - c\|_2)|.\end{aligned}$$

We thus have the bound

$$\sup_{g\in\mathcal{G}}\Big|\frac{1}{n}\sum_{i=1}^{n} s_i\Big| \leq K \sup_{c\in\Omega}\max_i |\phi(\|\hat{X}_i - c\|_2) - \phi(\|X_i - c\|_2)|.$$

Now, by Lemma 5, $\sup_i \|\hat{X}_i - X_i\|_2$ converges to 0 almost surely. Since ϕ is continuous on a compact set, it is uniformly continuous. Thus

$$\sup_{c\in\Omega}\sup_i |\phi(\|\hat{X}_i - c\|_2) - \phi(\|X_i - c\|_2)| \xrightarrow{\text{a.s.}} 0$$

as desired. □

6. Discussion

Lemma 5 provides a bound on the 2-to-∞ norm of the difference between $\hat{X}$ and X. The ability to control the errors of individual rows of $\hat{X}$ allows us to prove asymptotically almost surely perfect clustering in the SBM and DCSBM, a substantive improvement on the best existing spectral clustering results.

Our approach can be easily modified to prove several extensions of Theorem 6. For example, we can consider the special case in which the constants are fixed in n and $n_{min} = \Theta(n)$, whereupon the conditions of Theorem 6 are all satisfied for n sufficiently large. In this setting, we can further suppose that there are K positions $\xi_1, \ldots, \xi_k$ and $\mathbb{P}[X_i = \xi_k] = \pi_k$ for some $\pi_k > 0$, i.e. a mixture of point masses. In other words, this is an stochastic block model with independent, identically distributed block memberships (that are not fixed) across vertices. Proving that the number of errors converges almost surely to zero is then an easy application of Theorem 6. Furthermore, our methods can be extended to alternate clustering procedures, such as Gaussian mixture modeling (see [22]) or hierarchical clustering.

Indeed, one can construct many examples where perfect performance is achieved asymptotically (see Examples 1 and 2). We will not detail all regimes explicitly, but rather note that this theory can be easily applied to handle a growing number of blocks, possibly impacting d, γ and n_{min}, and moderately sparse regimes, impacting γ.

We believe Lemma 5 to be of independent interest apart from clustering. Indeed, Lemma 5 is a key result in proving consistency of a divide-and-conquer seeded graph matching procedure [11]. The lemma also leads to an easy proof of the strong consistency of k-nearest-neighbors for vertex classification, thereby extending the results of [21]. The lemma is a key component of the construction

of a consistent two-sample graph hypothesis test [24]. Additionally, we are exploring the implications of the lemma on parameter estimation for more general latent position random graphs.

The DCSBM is inherently more general than the SBM, and has key properties useful in modeling group structures in graphs. In [17], the authors provide complementary results for spectral analysis of the DCSBM without requiring lower bounds on the degrees; however, in turn, they obtain less-than-perfect clustering. Our results are the first to show that, depending on model parameters, the probability of perfect clustering tends to one as the number of vertices tends to infinity. The keys to the easy extension of these results to more general models are Lemmas 5 and 10, stated here in the RDPG setting.

For a general RDPG, there may not be a "natural" community structure. Nonetheless, the strong universal consistency result of Theorem 15 ensures that clustering the embedded graph will be asymptotically equivalent to the clustering the true latent positions. Finding the k-centers of the estimated latent positions provides one way to approximate the distribution of the latent positions as a mixture of point masses corresponding to an SBM, where the k distinct latent positions are given by the k-centers. Approximating a more general graph distribution as a stochastic blockmodel has been studied by Wolfe and Olhede [26] and Choi and Wolfe [6], and here we have detailed one spectral solution to this problem. If k is chosen appropriately, these approximations yield suitably parsimonious distributions that can be used for understanding large complex graphs, without requiring the estimation of a correspondingly complex distribution.

Acknowledgments

This work is partially supported by a National Security Science and Engineering Faculty Fellowship (NSSEFF), Johns Hopkins University Human Language Technology Center of Excellence, and the XDATA program of the Defense Advanced Research Projects Agency. Lastly, we would like to thank the anonymous referees for their helpful comments and suggestions and for suggesting Example 2.

References

[1] Athreya, A., Lyzinski, V., Marchette, D. J., Priebe, C. E., Sussman, D. L., and Tang, M., A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, accepted for publication. 2913

[2] Bickel, P. J. and Chen, A., A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA*, 106:21068–21073, 2009. 2906

[3] Bickel, P. J., Chen, A., and Levina, E., The method of moments and degree distributions for network models. *Ann. Statist.*, 39:38–59, 2011. MR2906868 2907

[4] Chaudhuri, K., Chung, F., and Tsiatas, A., Spectral partitioning of graphs with general degrees and the extended planted partition model. In *Proceedings of the 25th Conference on Learning Theory*, 2012. 2906

[5] Choi, D. S., Wolfe, P. J., and Airoldi, E. M., Stochastic blockmodels with a growing number of classes. *Biometrika*, 99:273–284, 2012. MR2931253 2906

[6] Choi, D. and Wolfe, P. J., Co-clustering separately exchangeable network data. *Ann. Statist.*, 42(1):29–63, 2014. MR3161460 2920

[7] Hoff, P. D., Raftery, A. E., and Handcock, M. S., Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.*, 97(460):1090–1098, 2002. MR1951262 2906

[8] Holland, P. W., Laskey, K., and Leinhardt, S., Stochastic blockmodels: First steps. *Social Networks*, 5:109–137, 1983. MR0718088 2906

[9] Karrer, B. and Newman, M. E. J., Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83, 2011. MR2788206 2906, 2916

[10] Lei, J. and Rinaldo, A., Consistency of spectral clustering in stochastic blockmodels. *Ann. Statist.*, 43:215–237, 2015. MR3285605 2906

[11] Lyzinski, V., Sussman, D. L., Fishkind, D. E., Pao, H., Chen, L., Vogelstein, J. T., and Priebe, C. E., Spectral clustering for divide-and-conquer graph matching. Arxiv preprint at http://arxiv.org/abs/1310.1297, 2014. 2919

[12] Mossel, E., Neeman, J., and Sly, A., Consistency thresholds for binary symmetric block models. ArXiv preprint at http://arxiv.org/abs/1407.1591, 2014. 2912

[13] Mossel, E., Neeman, J., and Sly, A., Stochastic block models and reconstruction. *Probab. Theory Related Fields*, in press. 2912

[14] Newman, M. E. J., Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103(23):8577–8582, 2006. 2906

[15] Oliveira, R. I., Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. Arxiv preprint at http://arxiv.org/abs/0911.0600, 2010. 2910

[16] Pollard, D., Strong consistency of k-means clustering. *Ann. Statist.*, 9:135–140, 1981. MR0600539 2907, 2917, 2918

[17] Qin, T. and Rohe, K., Regularized spectral clustering under the degree-corrected stochastic blockmodel. *NIPS*, 2013. 2906, 2907, 2916, 2920

[18] Rohe, K., Chatterjee, S., and Yu, B., Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, 39:1878–1915, 2011. MR2893856 2906

[19] Snijders, T. A. B. and Nowicki, K., Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *J. Classification*, 14:75–100, 1997. MR1449742 2906

[20] Sussman, D. L., Tang, M., Fishkind, D. E., and Priebe, C. E., A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.*, 107(499):1119–1128, 2012. MR3010899 2906, 2913

[21] Sussman, D. L., Tang, M., and Priebe, C. E., Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Trans. Pattern. Anal.*, 36:48–57, 2014. 2919

[22] Suwan, S., Lee, D. S., Tang, R., Sussman, D. L., Tang, M., and Priebe, C. E., Empirical bayes estimation for the stochastic blockmodel. Arxiv preprint at arxiv.org/abs/1405.6070, 2014. 2919

[23] Tang, M., Sussman, D. L., and Priebe, C. E., Universally consistent vertex classification for latent position graphs. *Ann. Statist.*, 41:1406–1430, 2013. MR3113816 2913

[24] Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., and Priebe, C. E., A semiparametric two-sample hypothesis testing problem for random dot product graphs. ArXiv preprint at http://arxiv.org/abs/1403.7249, 2014. 2920

[25] Tropp, J. A., Freedman's inequality for matrix martingales. *Electron. Commun. Probab*, 16:262–270, 2011. MR2802042 2910, 2913

[26] Wolfe, P. J. and Olhede, S. C., Nonparametric graphon estimation. ArXiv preprint at http://arxiv.org/abs/1309/5936, 2013. 2920

[27] Young, S. and Scheinerman, E., Random dot product graph models for social networks. In *Proceedings of the 5th International Conference on Algorithms and Models for the Web-Graph*, pages 138–149, 2007. MR2504912 2906