

PERFECT CLUSTERING FOR STOCHASTIC BLOCKMODEL GRAPHS VIA ADJACENCY SPECTRAL EMBEDDING

BY VINCE LYZINSKI, DANIEL L. SUSSMAN, MINH TANG,
AVANTI ATHREYA, AND CAREY E. PRIEBE

Johns Hopkins University

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.

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

The stochastic blockmodel and its extensions are key tools for modeling communities and groups in graphs [5, 6]. Given a stochastic blockmodel graph, one way to cluster the vertices is to estimate the block memberships of all vertices. Many techniques have been proposed for estimating the block memberships, including maximization of modularity [8], likelihood maximization [2, 4], Bayesian methods [12] and spectral techniques [11, 13]. In this paper, we focus on the latter, particularly on clustering the vertices using the adjacency spectral embedding.

*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.

AMS 2000 subject classifications: Primary 62C30; secondary 05C80

Keywords and phrases: Clustering, stochastic block model, degree corrected stochastic block model

The authors of [3] have shown that likelihood-based techniques can be employed to achieve error-free vertex clustering with probability tending to one as n , the number of vertices, tends to infinity. For clustering using spectral methods, however, the best existing results assert only that with probability tending to one, at most $O(\log(n))$ vertices will be misclustered. We provide here the necessary theory to bring the error estimates for these two methods in line, showing that spectral clustering will have asymptotically perfect performance for the stochastic blockmodel.

Our main result is Theorem 5, which gives a bound, dependent on model assumptions, on the probability that a mean square error clustering of the adjacency spectral embedding will be error-free. We also discuss the asymptotic implications of these results. To prove this theorem, we prove two key lemmas about a more general model. Consequently, we are able to prove an analogous result for clustering in the degree corrected stochastic blockmodel [7]. Finally, we prove a more general result that spectral clustering of random dot product graphs is strongly universally consistent in the sense of [9].

2. Setting and Main Theorem. In our present setting, we deviate slightly from the standard characterization of the stochastic blockmodel and focus, instead, on a particular latent position parametrization of the model, namely the random dot product graph [17]. Random dot product graphs are a convenient theoretical tool, and their spectral properties are well-understood. This setting replaces the non-geometric constructions, in which each block is associated with a categorical label, with a geometric construction in which each block is associated with a point in Euclidean space.

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

For an RDPG, we also define $P = XX^\top$ and let $P = [V|\tilde{V}][S \oplus \tilde{S}][V|\tilde{V}]^\top$ be the spectral decomposition of P with orthonormal $[V|\tilde{V}] \in \mathbb{R}^{n \times n}$, $V \in \mathbb{R}^{n \times d}$,

$S \in \mathbb{R}^{d \times d}$, and

$$S(1, 1) \geq S(2, 2) \geq \dots \geq S(d, d) > 0$$

$$\tilde{S} = 0.$$

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

It easily follows that there exists an orthonormal W such that $VS^{1/2} = XW$. We thus suppose that $X = VS^{1/2}$. This assumption does not lead to any loss of generality since, for the inference tasks in this paper, we will be performing clustering using the mean square error criterion, which is invariant under orthogonal transformations. Similarly define $\hat{V} \in \mathbb{R}^{n \times d}$, and $\hat{S} \in \mathbb{R}^{d \times d}$. We define $\hat{X} = \hat{V}\hat{S}^{1/2}$ to be the adjacency spectral embedding (ASE) of A . We shall assume, for ease of exposition, that the diagonal entries of \hat{S} are positive. As will be seen later, e.g., Lemma 7, 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 .

We now define the stochastic blockmodel in terms of random dot product graphs.

DEFINITION 2 ((Positive Semidefinite) 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 $\tau_i = k$.

REMARK. Our definition of an SBM is somewhat atypical: SBMs are traditionally defined by the $K \times K$ matrix of probabilities of adjacencies between vertices in each of the blocks. The two definitions coincide provided this matrix is positive semidefinite.

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

DEFINITION 3 (Mean Square Error (MSE) Clustering). The MSE clustering of the rows of \hat{X} into K blocks returns

$$\hat{C} := \min_{C \in \mathcal{C}_K} \|C - \hat{X}\|_F, \text{ where}$$

$$\mathcal{C}_K = \{C \in \mathbb{R}^{n \times d} : C \text{ has } K \text{ distinct rows}\},$$

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

We use the notation $\|\cdot\|_{\alpha \rightarrow \beta}$ to represent the α -to- β operator norm: that is,

$$\|M\|_{\alpha \rightarrow \beta} = \sup_{x \neq 0} \frac{\|Mx\|_{\beta}}{\|x\|_{\alpha}} = \sup_{\|x\|_{\alpha}=1} \|Mx\|_{\beta}.$$

When $\alpha = \beta$, we write $\|\cdot\|_{\alpha}$ in place of $\|\cdot\|_{\alpha \rightarrow \alpha}$.

We note some basic properties of $\alpha \rightarrow \beta$ norms. In particular, $\|M\|_{2 \rightarrow \infty} = \max_i \|M_i\|_2$. Indeed,

$$\|M\|_{2 \rightarrow \infty} = \sup_{\|x\|_2=1} \|Mx\|_{\infty} = \sup_{\|x\|_2=1} \max_i |M_i x| = \max_i \|M_i\|_2.$$

In addition, for any matrix A and B whose product AB is well-defined, $\|ABx\|_{\beta} \leq \|A\|_{\alpha \rightarrow \beta} \|Bx\|_{\alpha}$ by the definition of the $\|\cdot\|_{\alpha \rightarrow \beta}$ norm and this implies

$$\|AB\|_{\alpha \rightarrow \beta} = \sup_{x \neq 0} \frac{\|ABx\|_{\beta}}{\|x\|_{\alpha}} \leq \sup_{x \neq 0} \frac{\|A\|_{\alpha \rightarrow \beta} \|Bx\|_{\alpha}}{\|x\|_{\alpha}} = \|A\|_{\alpha \rightarrow \beta} \|B\|_{\alpha},$$

i.e., $\|AB\|_{\alpha \rightarrow \beta} \leq \|A\|_{\alpha \rightarrow \beta} \|B\|_{\alpha}$.

We now state a technical lemma in which we bound the maximum difference between the rows of \hat{X} and the rows of an orthogonal transformation of X .

LEMMA 4. *Let Δ and γ be defined as*

$$\Delta = \max_i \sum_{j \neq i} P_{ij}$$

$$\gamma = \min_{1 \leq i \leq d} |S(i+1, i+1) - S(i, i)|/n$$

Δ is then the maximum of the row sums of P and γn is the minimum gap among the distinct eigenvalues of P . 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

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

Lemma 4 gives far greater control of the errors than previous results that were derived for the Frobenius norm $\|\hat{X} - X\|_F$; the latter bounds do not

allow fine control of the errors in the individual rows of \hat{X} . Lemma 4, 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 5. *Let $A \sim \text{RDPG}(X)$ and is an SBM with K blocks and block membership function τ and suppose $\eta \in (0, 1/2)$. Denote the bound on $\|\hat{X} - X\|_{2 \rightarrow \infty}$ in Lemma 4 as $\beta = \beta(d, n, \eta, \gamma)$. Let $r > 0$ be such that for all $i, j \in [n]$ with $X_i \neq X_j$, $\|X_i - X_j\|_2 > 4r$. Let $\hat{\tau} : [n] \rightarrow [K]$ 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 $r > \beta\sqrt{n/n_{\min}}$ and $\gamma n > 4\sqrt{\Delta \log(n/\eta)}$ then with probability at least $1 - 2\eta$,*

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

We remark that these conditions are quite natural: we require that the rows of X with distinct entries have some minimum separation, and that this separation is not too small compared to the smallest block size. Our assumption on γ ensures a large enough gap in the eigenvalues to use Lemma 4. We note that Lemma 4 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 $k \geq 2$. Finally, we observe that the theorem has both finite-sample and asymptotic implications. In particular, under these model assumptions, for any finite n , the 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.

3. Proof of Theorem 5. Before we prove Theorem 5, we first collect a sequence of useful bounds from [13, 15, 16]. We then prove two key lemmas.

PROPOSITION 6. *Suppose $A \sim \text{RDPG}(X)$ with $X \in \mathcal{X}_d$ and let Δ and γ be as defined in Lemma 4. For any $\eta \in (0, 1/2)$, if $\gamma n > 4\sqrt{\Delta \log(n/\eta)}$, then the following occur with probability at least $1 - \eta$*

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

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 7. *In the setting of Proposition 6, if the events in Proposition 6 occur, then*

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

LEMMA 8. *In the setting of Proposition 6, if the events in Proposition 6 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 9. *In the setting of Proposition 6, if the events in Proposition 6 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

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

We have

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

by Proposition 6 and our assumption that $\gamma\sqrt{n} \geq 4\sqrt{\log(n/\eta)}$. We now apply Lemma 9 to derive the bound

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

Similarly, we have that $\|E\|_{2 \rightarrow 2} \leq 2\sqrt{\Delta \log(n/\eta)}$. Therefore, we can 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. \square

We now prove Lemma 4 by dominating the $\|\cdot\|_{2 \rightarrow \infty}$ operator norm by the Frobenius norm and applying Hoeffding's inequality.

PROOF OF LEMMA 4. Since $X = PV S^{-1/2}$ we can add and subtract the matrix $AV \hat{S}^{-1/2}$ and $AV S^{-1/2}$ 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 9 bounds the first term in terms of the Frobenius norm which is a bound for the $2 \rightarrow \infty$ 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. We thus have

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

and hence

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

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

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

Next, since

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

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

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

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

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

and consequently that

$$\mathbb{P}\left(\|(A - P)V\|_{2 \rightarrow \infty} > \sqrt{\frac{d}{2} \log(2nd/\eta)}\right) \leq \eta.$$

The third term can therefore be bounded as

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

with probability at least $1 - \eta$. Combining the bounds for the above three terms yields Lemma 4. \square

PROOF OF THEOREM 5. We assume that the event in Lemma 4 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 $\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_K$ be L^2 -balls with radii $2r$ around the K distinct rows of X . By the assumptions in Theorem 5, these balls are disjoint. Suppose now that 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 no row of \hat{C} is within $2r$ of a row of X in \mathcal{B}_k , of which there are at least n_{\min} . 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, $\|\hat{C} - X\|_{2 \rightarrow \infty} \leq 2r$. 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. \square

4. Degree Corrected SBM. In this section we extend our results to the degree corrected SBM [7].

DEFINITION 10 (Degree Corrected Stochastic Blockmodel (DCSBM)). We say an RDPG is a DCSBM with K blocks if there exist K unit vectors $y_1, \dots, y_K \in \mathbb{R}^d$ such that for each $i \in [n]$, there exists $k \in [K]$ and $c_i \in (0, 1)$ such that $X_i = c_i y_k$.

For this model, we introduce $Y \in \mathbb{R}^{n \times d}$ to be such that $Y_i = y_{\tau(i)}$, so each row of Y has unit L^2 -norm. Similarly, let $\hat{Y} = \text{diag}(\hat{X}\hat{X}^\top)^{-1/2}\hat{X}$ where $\text{diag}(\cdot)$ denotes the operation of setting all the off-diagonal elements of its 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}^n .

LEMMA 11. *In the setting of Lemma 4, let the $n \times d$ matrices $\tilde{Y}, \hat{Y} \in \mathcal{S}^n$ be the projections of X and \hat{X} , respectively, onto \mathcal{S}^n . Let $c_{\min} = \min_{i \in [n]} \|X_i\|_2$. If $\|\hat{X} - X\|_{2 \rightarrow \infty} \leq \beta$, then*

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

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

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

as desired. \square

This again leads to the following theorem about the probability of error-free clustering.

THEOREM 12. *Suppose $A \sim \text{RDPG}(X)$ and is a DCSBM with K blocks and block membership function τ and suppose $\eta \in (0, 1/2)$. Let y_1, \dots, 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 5. Suppose $r > 0$ is such that for all $i, j \in [K]$, $\|y_i - y_j\|_2 > 4r$. Let $\hat{\tau} : [n] \rightarrow [K]$ be the optimal MSE clustering of the rows of \hat{Y} , the projection of \hat{X} onto \mathcal{S}^n , into K clusters. Finally, let $n_{\min} = \min_{k \in [K]} n_k$ be the smallest block size. If $r > (2\beta\sqrt{n/n_{\min}})/c_{\min}$ 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 5.

5. Strong universal consistency. The problem strong consistency of K -means clustering is considered in [9]. Specifically, suppose that $\{X_1, \dots, X_n\}$ 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 $\phi : [0, \infty) \mapsto [0, \infty)$ is a continuous, nondecreasing function with $\phi(0) = 0$. Now define $\Phi(C, F_n)$ and $\Phi(C, F)$ by

$$\begin{aligned} \Phi(C, F_n) &= \int \left(\min_{c \in C} \phi(\|x - c\|) \right) dF_n(x) \\ \Phi(C, F) &= \int \left(\min_{c \in C} \phi(\|x - c\|) \right) dF(x) \end{aligned}$$

The problem of K -means mean square error clustering given $\{X_1, \dots, X_n\}$ 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 13 ([9]). *Suppose that for each $k = 1, \dots, K$, there is a unique set \bar{C}_k for which*

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

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

We now state the counterpart to Theorem 13 for the RDPG setting.

THEOREM 14. *Let $A \sim \text{RDPG}(X)$ 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 13 holds for F , $\hat{C}_n \rightarrow \bar{C}_K$ almost surely, and furthermore, $\Phi(\hat{C}_n, \hat{F}_n) \rightarrow \Phi(\bar{C}_K, F)$ 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 $g_C(x) = \min_{c \in C} \phi(\|x - c\|)$ 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 13, 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

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

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

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

and hence

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

We thus have the bound

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

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

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

as desired. \square

6. Discussion. Lemma 4 bounds on the 2-to- ∞ norm of the difference between \hat{X} and a linearly transformed X . In our proof, we exploit the fact that the clustering criterion is relatively insensitive to such linear transformations.

If we are interested in asymptotics, we can take η to be n^{-c} for some c and so $\beta = O(d\Delta^3 \log(n)/(\gamma n)^{7/2})$. We can also consider the special case in which the constants are fixed in n and $n_{\min} = \Theta(n)$, whereupon the conditions of Theorem 5 are all satisfied for n sufficiently large. For example, if we suppose there are K positions ξ_1, \dots, ξ_k and $\mathbb{P}[X_i = \xi_k] = \pi_k$ for some $\pi_k > 0$, i.e. a mixture of point masses, then this is an stochastic block model with independent, identically distributed block memberships across vertices; it is not hard to show that the number of errors converges almost surely to zero.

Indeed, one can construct many examples where perfect performance is achieved asymptotically. We will not detail all regimes explicitly, but rather note that this theorem can be applied to a growing number of blocks, possibly impacting d , γ and n_{\min} , as well as moderately sparse regimes, impacting γ .

The DCSBM is more flexible than the SBM, but still has key properties useful in modeling group structures in graphs. In [10], 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 4 and 9, stated here in the RDPG setting. The generality of this setting renders these lemmas applicable to a number of other models and inference tasks. For example, the above results can be used to show that the K -nearest neighbor classifier of the adjacency spectral embedding is a strongly universally consistent vertex classifier, thereby extending the results of [14].

REFERENCES

- [1] ATHREYA, A., LYZINSKI, V., MARCHETTE, D. J., PRIEBE, C. E., SUSSMAN, D. L. and TANG, M. (2013). A central limit theorem for scaled eigenvectors of random dot product graphs. arXiv preprint at <http://arxiv.org/abs/1305.7388>.
- [2] BICKEL, P. J. and CHEN, A. (2009). 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** 21068–73.

- [3] BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Annals of Statistics* **39** 38–59.
- [4] CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284.
- [5] GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning* **2** 129–233.
- [6] HOLLAND, P. W., LASKEY, K. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137.
- [7] KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E* **83**.
- [8] NEWMAN, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103** 8577–8582.
- [9] POLLARD, D. (1981). Strong consistency of K -means clustering. *Annals of Statistics* **9** 135–140.
- [10] QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. arXiv preprint at <http://arxiv.org/abs/1309.4111>.
- [11] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics* **39** 1878–1915.
- [12] SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification* **14** 75–100.
- [13] SUSSMAN, D. L., TANG, M., FISHKIND, D. E. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association* **107** 1119–1128.
- [14] SUSSMAN, D. L., TANG, M. and PRIEBE, C. E. (2014). Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36** 48–57.
- [15] TANG, M., SUSSMAN, D. L. and PRIEBE, C. E. (2013). Universally consistent vertex classification for latent position graphs. *Annals of Statistics* **41** 1406 – 1430.
- [16] TROPP, J. A. (2011). Freedmans inequality for matrix martingales. *Electronic Communications in Probability* **16** 262–270.
- [17] YOUNG, S. and SCHEINERMAN, E. (2007). Random dot product graph models for social networks. In *Algorithms and models for the web-graph* 138–149. Springer.

DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS,
 JOHNS HOPKINS UNIVERSITY,
 3400 N. CHARLES ST, BALTIMORE, MD 21218, USA.
 E-MAIL: ([vlyzins1](mailto:vlyzins1@jhu.edu), [dsussma3](mailto:dsussma3@jhu.edu), [mtang10](mailto:mtang10@jhu.edu), [dathrey1](mailto:dathrey1@jhu.edu), [cep](mailto:cep@jhu.edu))@jhu.edu