A Consistent Adjacency Spectral Embedding for Stochastic Blockmodel Graphs

Daniel L. Sussman [a] , Minh Tang [a] , Donniell E. Fishkind [a] & Carey E. Priebe [a]

[a] Department of Applied Math and Statistics, Johns Hopkins University, Baltimore, MD, 21218-2682

PLEASE SCROLL DOWN FOR ARTICLE

A Consistent Adjacency Spectral Embedding for Stochastic Blockmodel Graphs

Daniel L. SUSSMAN, Minh TANG, Donniell E. FISHKIND, and Carey E. PRIEBE

We present a method to estimate block membership of nodes in a random graph generated by a stochastic blockmodel. We use an embedding procedure motivated by the random dot product graph model, a particular example of the latent position model. The embedding associates each node with a vector; these vectors are clustered via minimization of a square error criterion. We prove that this method is consistent for assigning nodes to blocks, as only a negligible number of nodes will be misassigned. We prove consistency of the method for directed and undirected graphs. The consistent block assignment makes possible consistent parameter estimation for a stochastic blockmodel. We extend the result in the setting where the number of blocks grows slowly with the number of nodes. Our method is also computationally feasible even for very large graphs. We compare our method with Laplacian spectral clustering through analysis of simulated data and a graph derived from Wikipedia documents.

KEY WORDS: Graph; Latent position; Network; Random dot product.

1. BACKGROUND AND OVERVIEW

Network analysis is rapidly becoming a key tool in the analysis of modern datasets in fields ranging from neuroscience to sociology to biochemistry. In each of these fields, there are objects, such as neurons, people, or genes, and there are relationships between objects, such as synapses, friendships, or protein interactions. The formation of these relationships can depend on attributes of the individual objects as well as higher-order properties of the network as a whole. Objects with similar attributes can form communities with similar connective structure, while unique properties of individuals can fine-tune the shape of these relationships. Graphs encode the relationships between objects as edges between nodes in the graph.

Clustering objects based on a graph enables identification of communities and objects of interest as well as illumination of overall network structure. Finding optimal clusters is difficult and will depend on the particular setting and task. Even in moderately sized graphs, the number of possible partitions of nodes is enormous, so a tractable search strategy is necessary. Methods for finding clusters of nodes in graphs are many and varied, with origins in physics, engineering, and statistics; Fortunato (2010) and Fjallstrom (1998) provided comprehensive reviews of clustering techniques. In addition to techniques motivated by heuristics based on graph structure, others have attempted to fit statistical models with inherent community structure to a graph (Snijders and Nowicki 1997; Nowicki and Snijders 2001; Handcock, Raftery, and Tantrum 2007; Airoldi et al. 2008).

These statistical models use random graphs to model relationships between objects; Goldenberg (2009) provided a review of statistical models for networks. A graph consists of a set of nodes, representing the objects, and a set of edges, representing relationships between the objects. The edges can be either directed (ordered pairs of nodes) or undirected (unordered pairs of nodes). In our setting, the node set is fixed and the set of edges is random.

Hoff, Rafferty, and Handcock (2002) proposed what they called a latent space model for random graphs. Under this model, each node is associated with a latent random vector. There may also be additional covariate information that we do not consider in this work. The vectors are independent and identically distributed, and the probability of an edge between two nodes depends only on their latent vectors. Conditioned on the latent vectors, the presence of each edge is an independent Bernoulli trial.

One example of a latent space model is the random dot product graph (RDPG) model (Young and Scheinerman 2007). Under the RDPG model, the probability that an edge between two nodes is present is given by the dot product of their respective latent vectors. For example, in a social network with edges indicating friendships, the components of the vector may be interpreted as the relative interest of the individual in various topics. The magnitude of the vector can be interpreted as how talkative the individual is, with more talkative individuals more likely to form relationships. Talkative individuals interested in the same topics are most likely to form relationships, while individuals who do not share interests are unlikely to form relationships.

We present an embedding motivated by the RDPG model that uses a decomposition of a low rank approximation of the adjacency matrix. The decomposition gives an embedding of the nodes as vectors in a low-dimensional space. This embedding is similar to embeddings used in spectral clustering but operates directly on the adjacency matrix rather than a Laplacian. We discuss a relationship between spectral clustering and our work in Section 7.

Our results are for graphs generated by a stochastic blockmodel (Holland, Laskey, and Leinhardt 1983; Wang and Wong 1987). In this model, each node is assigned to a block, and the probability of an edge between two nodes depends only on their respective block memberships; in this manner, two nodes in the same block are stochastically equivalent. In the context of the latent space model, all nodes in the same block are assigned the same latent vector. An advantage of this model is the clear

Daniel L. Sussman (E-mail: *dsussma3@jhu.edu*), Minh Tang (E-mail: *mtang10@jhu.edu*), Donniell E. Fishkind (E-mail: *def@jhu.edu*), and Carey E. Priebe (E-mail: *cep@jhu.edu*), Department of Applied Math and Statistics, Johns Hopkins University, Baltimore, MD 21218-2682.

and simple block structure, where block membership is determined solely by the latent vector.

Given a graph generated from a stochastic blockmodel, our primary goal is to accurately assign all of the nodes to their correct blocks. Algorithm 1 gives the main steps of our procedure. In summary, these steps involve computing the singular value decomposition of the adjacency matrix, reducing the dimension, coordinate-scaling the singular vectors by the square root of their singular value, and, finally, clustering via minimization of a square error criterion. We note that Step 4 in the procedure is a mathematically convenient stand-in for what might be used in practice. Indeed, the standard K-means algorithm approximately minimizes the square error and we use K-means for evaluating the procedure empirically. This article shows that the node assignments returned by Algorithm 1 are consistent.

Consistency of node assignments means that the proportion of misassigned nodes goes to zero (probabilistically) as the number of nodes goes to infinity. Others have already shown similar consistency of node assignments. Snijders and Nowicki (1997) provided an algorithm to consistently assign nodes to blocks under the stochastic blockmodel for two blocks, and later Condon and Karp (2001) provided a consistent method for equal-sized blocks. Bickel and Chen (2009) showed that maximizing the Newman–Girvan modularity (Newman and Girvan 2004) or the likelihood modularity provides consistent estimation of block membership. Choi, Wolfe, and Airoldi (2010) used likelihood methods to show consistency with rapidly growing numbers of blocks.

Maximizing modularities and likelihood methods are both computationally difficult, but provide theoretical results for rapidly growing numbers of blocks. Our method is related to that of McSherry (2001), in that we consider a low rank approximation of the adjacency matrix, but their results do not provide consistency of node assignments. Rohe, Chatterjee, and Yu (2010) used spectral clustering to show consistent estimation of block partitions with growing number of blocks; in this article, we demonstrate that for both directed and undirected graphs, our proposed embedding allows for accurate block assignment in a stochastic blockmodel. These matrix decomposition methods are computationally feasible, even for graphs with a large number of nodes.

Algorithm 1: The adjacency spectral clustering procedure for directed graphs.

Input: $\mathbf{A} \in \{0, 1\}^{n \times n}$
Parameters: $d \in \{1, 2, \dots, n\}$, $K \in \{2, 3, \dots, n\}$
Step 1: Compute the singular value decomposition, $\mathbf{A} = \widetilde{\mathbf{U}}'\widetilde{\mathbf{\Sigma}}'\widetilde{\mathbf{V}}'^T$. Let $\widetilde{\mathbf{\Sigma}}'$ have decreasing main diagonal.
Step 2: Let $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ be the first d columns of $\widetilde{\mathbf{U}}'$ and $\widetilde{\mathbf{V}}'$, respectively, and let $\widetilde{\mathbf{\Sigma}}$ be the submatrix of $\widetilde{\mathbf{\Sigma}}'$ given by the first d rows and columns.
Step 3: Define $\widetilde{\mathbf{Z}} = [\widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}^{1/2} | \widetilde{\mathbf{V}}\widetilde{\mathbf{\Sigma}}^{1/2}] \in \mathbb{R}^{n \times 2d}$ to be the concatenation of the coordinate-scaled singular vector matrices.
Step 4: Let $(\hat{\boldsymbol{\psi}}, \hat{\tau}) = \text{argmin}_{\boldsymbol{\psi};\tau} \sum_{u=1}^{n} \|\widetilde{Z}_u - \psi_{\tau(u)}\|_2^2$ give the centroids and block assignments, where $\widetilde{Z}_u$ is the uth row of $\widetilde{\mathbf{Z}}$, $\hat{\boldsymbol{\psi}} \in \mathbb{R}^{K \times d}$ are the centroids and $\hat{\tau}$ is a function from $[n]$ to $[K]$.
return $\hat{\tau}$, the block assignment function.

The remainder of the article is organized as follows. In Section 2, we formally present the stochastic blockmodel, the RDPG model, and our adjacency spectral embedding. In Section 3, we state and prove our main theorem, and in Section 4, we present some useful Corollaries. Sections 2–4 focus only on directed random graphs; in Section 5, we present the model and results for undirected graphs. In Section 6, we present simulations and empirical analysis to illustrate the performance of the algorithm. Finally, in Section 7, we discuss further extensions to the theorem. In the Appendix, we prove some key technical results to prove our main theorem.

2. MODEL AND EMBEDDING

First, we adopt the following conventions. For a matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$, entry i, j is denoted by $\mathbf{M}_{ij}$. Row i is denoted by $\mathbf{M}_i^T \in \mathbb{R}^{1 \times d}$, where $\mathbf{M}_i$ is a column vector. Column j is denoted as $\mathbf{M}_{\cdot j}$ and occasionally we refer to row i as $\mathbf{M}_{i\cdot}$.

The node set is $[n] = \{1, 2, \dots, n\}$. For directed graphs edges are ordered pairs of elements in $[n]$. For a random graph, the node set is fixed and the edge set is random. The edges are encoded in an adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$. For directed graphs, the entry $\mathbf{A}_{uv}$ is 1 or 0 according as an edge from node u to node v is present or absent in the graph. We consider graphs with no loops, meaning $\mathbf{A}_{uu} = 0$ for all $u \in [n]$.

2.1 Stochastic Blockmodel

Our results are for random graphs distributed according to a stochastic blockmodel (Holland, Laskey, and Leinhardt 1983; Wang and Wong 1987), where each node is a member of exactly one block and the probability of an edge from node u to node v is determined by the block memberships of nodes u and v for all $u, v \in [n]$. The model is parameterized by $\mathbf{P} \in [0, 1]^{K \times K}$, and $\boldsymbol{\rho} \in (0, 1)^K$ with $\sum_{i=1}^{K} \rho_i = 1$. K is the number of blocks, which are labeled $1, 2, \dots, K$. The block memberships of all nodes are determined by the random block membership function $\tau : [n] \mapsto [K]$. For all nodes $u \in [n]$ and blocks $i \in [K]$, $\tau(u) = i$ would mean node u is a member of block i; node memberships are independent with $\mathbb{P}[\tau(u) = i] = \rho_i$.

The entry $\mathbf{P}_{ij}$ gives the probability of an edge from a node in block i to a node in block j for each $i, j \in [K]$. Conditioned on τ, the entries of $\mathbf{A}$ are independent, and $\mathbf{A}_{uv}$ is a Bernoulli random variable with parameter $\mathbf{P}_{\tau(u),\tau(v)}$ for all $u \neq v \in [n]$. This gives

$$\mathbb{P}[\mathbf{A}|\tau] = \prod_{u \neq v} \mathbb{P}[\mathbf{A}_{uv} \mid \tau(u), \tau(v)]$$
$$= \prod_{u \neq v} (\mathbf{P}_{\tau(u),\tau(v)})^{\mathbf{A}_{uv}} (1 - \mathbf{P}_{\tau(u),\tau(v)})^{1-\mathbf{A}_{uv}}, \quad (1)$$

with the product over all ordered pairs of nodes.

The row $\mathbf{P}_{i\cdot}$ and column $\mathbf{P}_{\cdot i}$ determine the probabilities of the presence of edges incident to a node in block i. In order that the blocks be distinguishable, we require that different blocks have distinct probabilities so that either $\mathbf{P}_{i\cdot} \neq \mathbf{P}_{j\cdot}$ or $\mathbf{P}_{\cdot i} \neq \mathbf{P}_{\cdot j}$ for all $i \neq j \in [K]$.

Theorem 1 shows that using our embedding (Section 2.3) and a mean square error clustering criterion (Section 2.4), we are able to accurately assign nodes to blocks, for all but a negligible

number of nodes, for graphs distributed according to a stochastic blockmodel.

2.2 Random Dot Product Graphs

We present the RDPG model to motivate our embedding technique (Section 2.3) and provide a second parameterization for stochastic blockmodels (Section 2.5). Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times d}$ be such that $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n]^T$ and $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n]^T$, where $\mathbf{X}_u, \mathbf{Y}_u \in \mathbb{R}^d$ for all $u \in [n]$. The matrices $\mathbf{X}$ and $\mathbf{Y}$ are random and satisfy $\mathbb{P}[\langle \mathbf{X}_u, \mathbf{Y}_v \rangle \in [0, 1]] = 1$ for all $u, v \in [n]$. Conditioned on $\mathbf{X}$ and $\mathbf{Y}$, the entries of the adjacency matrix $\mathbf{A}$ are independent and $\mathbf{A}_{uv}$ is a Bernoulli random variable with parameter $\langle \mathbf{X}_u, \mathbf{Y}_v \rangle$ for all $u \neq v \in [n]$. This gives

$$\mathbb{P}[\mathbf{A} \mid \mathbf{X}, \mathbf{Y}] = \prod_{u \neq v} \mathbb{P}[\mathbf{A}_{uv} \mid \mathbf{X}_u, \mathbf{Y}_v]$$
$$= \prod_{u \neq v} \langle \mathbf{X}_u, \mathbf{Y}_v \rangle^{\mathbf{A}_{uv}} (1 - \langle \mathbf{X}_u, \mathbf{Y}_v \rangle)^{1-\mathbf{A}_{uv}}, \quad (2)$$

where the product is over all ordered pairs of nodes.

2.3 Embedding

The RDPG model motivates the following embedding. By an embedding of an adjacency matrix $\mathbf{A}$, we mean

$$(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) = \underset{(\mathbf{X}^\dagger, \mathbf{Y}^\dagger) \in \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d}}{\operatorname{argmin}} \|\mathbf{A} - \mathbf{X}^\dagger \mathbf{Y}^{\dagger T}\|_F, \quad (3)$$

where d, the target dimensionality of the embedding, is fixed and known and $\| \cdot \|_F$ denotes the Frobenius norm. Though $\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}^T$ may be a poor approximation of $\mathbf{A}$, Theorems 1 and 3 show that such an embedding provides a representation of the nodes that enables clustering of the nodes, provided the random graph is distributed according to a stochastic blockmodel. In fact, if a graph is distributed according to an RDPG model, then a solution to Equation (3) provides an estimate of the latent vectors given by $\mathbf{X}$ and $\mathbf{Y}$. We do not explore properties of this estimate but instead focus on the stochastic blockmodel.

Eckart and Young (1936) provided the following solution to Equation (3). Let $\mathbf{A} = \widetilde{\mathbf{U}}' \widetilde{\mathbf{\Sigma}}' \widetilde{\mathbf{V}}'^T$ be the singular value decomposition of $\mathbf{A}$, where $\widetilde{\mathbf{U}}', \widetilde{\mathbf{V}}' \in \mathbb{R}^{n \times n}$ are orthogonal and $\widetilde{\mathbf{\Sigma}}' \in \mathbb{R}^{n \times n}$ is diagonal, with diagonals $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_n(\mathbf{A}) \geq 0$, the singular values of $\mathbf{A}$. Let $\widetilde{\mathbf{U}} \in \mathbb{R}^{n \times d}$ and $\widetilde{\mathbf{V}} \in \mathbb{R}^{n \times d}$ be the first d columns of $\widetilde{\mathbf{U}}'$ and $\widetilde{\mathbf{V}}'$, respectively, and let $\widetilde{\mathbf{\Sigma}} \in \mathbb{R}^{d \times d}$ be the diagonal matrix with diagonals $\sigma_1(\mathbf{A}), \dots, \sigma_d(\mathbf{A})$. Equation (3) is solved by $\widetilde{\mathbf{X}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}^{1/2}$ and $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{V}}\widetilde{\mathbf{\Sigma}}^{1/2}$.

We refer to $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ as the "scaled adjacency spectral embedding" of $\mathbf{A}$. We refer to $(\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}})$ as the "unscaled adjacency spectral embedding" of $\mathbf{A}$. The adjacency spectral embedding is similar to an embedding which is presented in Marchette, Priebe, and Coppersmith (2011). It is also similar to spectral clustering where the decomposition is on the normalized graph Laplacian.

Theorem 1 uses a clustering of the unscaled adjacency spectral embedding of $\mathbf{A}$ while Corollary 3 extends the result to clustering on the scaled adjacency spectral embedding. Though this embedding is proposed for embedding an adjacency matrix, we use the same procedure to embed other matrices.

2.4 Clustering Criterion

We prove that for a graph distributed according to the stochastic blockmodel, we can use the following clustering criterion on the adjacency spectral embedding of $\mathbf{A}$ to accurately assign nodes to blocks. Let $\mathbf{Z} \in \mathbb{R}^{n \times m}$. We use the following mean square error criterion for clustering the rows of $\mathbf{Z}$ into K blocks,

$$(\hat{\boldsymbol{\psi}}, \hat{\tau}) = \underset{\boldsymbol{\psi}; \tau}{\operatorname{argmin}} \sum_{u=1}^{n} \|\mathbf{Z}_u - \boldsymbol{\psi}_{\tau(u)}\|_2^2, \quad (4)$$

where $\hat{\boldsymbol{\psi}} \in \mathbb{R}^{K \times m}$, $\hat{\boldsymbol{\psi}}_i \in \mathbb{R}^m$ gives the centroid of block i and $\hat{\tau} : [n] \mapsto [K]$ is the block assignment function.

Again, note that other computationally less expensive criterion can also be quite effective. Indeed, in Section 6.1, we achieve misclassification rates that are empirically better than our theoretical bounds using the K-means clustering algorithm, which only attempts to solve Equation (4). Additionally, other clustering algorithms may prove useful in practice, though presently we do not investigate these procedures.

2.5 Stochastic Blockmodel as RDPG Model

We present another parameterization of a stochastic blockmodel corresponding to the RDPG model. Suppose we have a stochastic blockmodel with rank$(\mathbf{P}) = d$. Then there exist $\boldsymbol{v}, \boldsymbol{\mu} \in \mathbb{R}^{K \times d}$ such that $\mathbf{P} = \boldsymbol{v}\boldsymbol{\mu}^T$ and by definition $\mathbf{P}_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{\mu}_j \rangle$. Let $\tau : [n] \mapsto [K]$ be the random block membership function.

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d}$ have row u given by $\mathbf{X}_u^T = \boldsymbol{v}_{\tau(u)}^T$ and $\mathbf{Y}_u^T = \boldsymbol{\mu}_{\tau(u)}^T$, respectively, for all u. Then we have

$$\mathbb{P}[\mathbf{A}_{uv} = 1] = \mathbf{P}_{\tau(u), \tau(v)} = \langle \boldsymbol{v}_{\tau(u)}, \boldsymbol{\mu}_{\tau(v)} \rangle = \langle \mathbf{X}_u, \mathbf{Y}_v \rangle. \quad (5)$$

In this way, the stochastic blockmodel can be parameterized by $\boldsymbol{v}, \boldsymbol{\mu} \in \mathbb{R}^{K \times d}$ and $\boldsymbol{\rho}$ provided that $(\boldsymbol{v}\boldsymbol{\mu}^T)_{ij} \in [0, 1]$ for all $i, j \in [K]$. This viewpoint proves valuable in the analysis and clustering of the adjacency spectral embedding.

Importantly, the distinctness of rows or columns in $\mathbf{P}$ is equivalent to the distinctness of the rows of $\boldsymbol{v}$ or $\boldsymbol{\mu}$. (Indeed note that for $i \neq j$, $\mathbf{P}_{i\cdot} - \mathbf{P}_{j\cdot} = 0$ if and only if $(\boldsymbol{v}_i^T - \boldsymbol{v}_j^T)\boldsymbol{\mu} = 0$, but rank$(\boldsymbol{\mu}) = d$ so $\boldsymbol{v}_i^T = \boldsymbol{v}_j^T$. Similarly, $\mathbf{P}_{\cdot i} = \mathbf{P}_{\cdot j}$ if and only if $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j$.) Also note, using the embedding of Section 2.3 on $\mathbf{P}$ with target dimensionality rank$(\mathbf{P})$ provides such a representation from any given $\mathbf{P}$.

3. MAIN RESULTS

3.1 Notation

We use the following notation for the remainder of this article. Let $\mathbf{P} \in [0, 1]^{K \times K}$ and $\boldsymbol{\rho} \in (0, 1)^K$ be a vector with positive entries summing to unity. Suppose rank$(\mathbf{P}) = d$. Let $\boldsymbol{v}\boldsymbol{\mu}^T = \mathbf{P}$ with $\boldsymbol{v}, \boldsymbol{\mu} \in \mathbb{R}^{K \times d}$. We now define the following constants not depending on n:

1. $\alpha > 0$ such that all eigenvalues of $\boldsymbol{v}^T\boldsymbol{v}$ and $\boldsymbol{\mu}^T\boldsymbol{\mu}$ are greater than α,
2. $\beta > 0$ such that $\beta < \|\boldsymbol{v}_i - \boldsymbol{v}_j\|$ or $\beta < \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$ for all $i \neq j$,
3. $\gamma > 0$ such that $\gamma < \rho_i$ for all $i \in [K]$.

We consider a sequence of random adjacency matrices $\mathbf{A}^{(n)}$ with node set $[n]$ for $n \in \{1, 2, \ldots\}$. The edges are distributed according to a stochastic blockmodel with parameters $\mathbf{P}$ and ρ. Let $\tau^{(n)} : [n] \mapsto [K]$ be the random block membership function, which induces the matrices $\mathbf{X}^{(n)}, \mathbf{Y}^{(n)} \in \mathbb{R}^{n \times d}$ as in Section 2.5. Let $n_i = |\{u : \tau(u) = i\}|$ be the size of block i.

Let $\mathbf{X}\mathbf{Y}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}$ be the singular value of decomposition, with $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n \times d}$ and $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$, so that $(\mathbf{U}, \mathbf{V})$ is the unscaled spectral embedding of the $\mathbf{X}\mathbf{Y}^T$. Let $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ be the adjacency spectral embedding of $\mathbf{A}$ and let $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}})$ be the unscaled adjacency spectral embedding of $\mathbf{A}$. Finally, let $\mathbf{W} \in \mathbb{R}^{n \times 2d}$ be the concatenation $[\mathbf{U}|\mathbf{V}]$ and similarly $\tilde{\mathbf{W}} = [\tilde{\mathbf{U}}|\tilde{\mathbf{V}}]$.

3.2 Main Theorem

The main contribution of this article is the following consistency result in terms of the estimation of the block memberships for each node based on the block assignment function $\hat{\tau}$ that assigns blocks based on $\tilde{\mathbf{W}}$. In the following, an event occurs "almost always" if with probability 1 the event occurs for all but finitely many $n \in \{1, 2, \ldots\}$.

Theorem 1. Under the conditions of Section 3.1, suppose that the number of blocks K and the latent vector dimension d are known. Let $\hat{\tau}^{(n)} : V \mapsto [K]$ be the block assignment function according to a clustering of the rows of $\tilde{\mathbf{W}}^{(n)}$ satisfying Equation (4). Let $\mathcal{S}_K$ be the set of permutations on $[K]$. It almost always holds that

$$\min_{\pi \in \mathcal{S}_K} |\{u \in V : \tau(u) \neq \pi(\hat{\tau}(u))\}| \leq \frac{2^3 3^2 6}{\alpha^5 \beta^2 \gamma^5} \log n. \quad (6)$$

To prove this theorem, we first provide a bound on the Frobenius norm of $\mathbf{A}\mathbf{A}^T - (\mathbf{X}\mathbf{Y}^T)(\mathbf{X}\mathbf{Y}^T)^T$, following Rohe, Chatterjee, and Yu (2010). Using this result and properties of the stochastic blockmodel, we then find a lower bound for the smallest nonzero singular value of $\mathbf{X}\mathbf{Y}^T$ and the corresponding singular value of $\mathbf{A}$. This enables us to apply the Davis–Kahan theorem (Davis and Kahan 1970) to show that the unscaled adjacency spectral embedding of $\mathbf{A}$ is approximately a rotation of the unscaled adjacency spectral embedding of $\mathbf{X}\mathbf{Y}^T$.

Finally, we lower bound the distances between the at most K distinct rows of $\mathbf{U}$ and $\mathbf{V}$. These gaps, together with the good approximation by the embedding of $\mathbf{A}$, are sufficient to prove consistency of the mean square error clustering of the embedded vectors. Most results, except the important Proposition 1 and the main theorem, are proved in the Appendix.

Proposition 1. Let $\mathbf{Q}^{(n)} \in [0, 1]^{n \times n}$ be a sequence of random matrices and let $\mathbf{A}^{(n)} \in \{0, 1\}^{n \times n}$ be a sequence of random adjacency matrices corresponding to a sequence of random graphs on n nodes for $n \in \{1, 2, \ldots\}$. Suppose the probability of an edge from node u to node v is given by $\mathbf{Q}_{uv}^{(n)}$ and that the presence of edges are conditionally independent given $\mathbf{Q}^{(n)}$. Then the following holds almost always:

$$\|\mathbf{A}^{(n)}\mathbf{A}^{(n)T} - \mathbf{Q}^{(n)}\mathbf{Q}^{(n)T}\|_F \leq \sqrt{3}n^{3/2}\sqrt{\log n}. \quad (7)$$

Proof. Note that, conditioned on $\mathbf{Q}$, $\mathbf{A}_{uw}$ and $\mathbf{A}_{vw}$ are independent Bernoulli random variables for all $w \in [n]$ provided $u \neq v$. For each $w \notin \{u, v\}$, $\mathbf{A}_{uw}\mathbf{A}_{vw}$ is a conditionally inde-

pendent Bernoulli with parameter $\mathbf{Q}_{uw}\mathbf{Q}_{vw}$. For $u \neq v$, we have

$$\mathbf{A}\mathbf{A}_{uv}^T - \mathbf{Q}\mathbf{Q}_{uv}^T = \sum_{w \notin \{u,v\}} (\mathbf{A}_{uw}\mathbf{A}_{vw} - \mathbf{Q}_{uw}\mathbf{Q}_{vw}) - \mathbf{Q}_{uu}\mathbf{Q}_{vu} - \mathbf{Q}_{uv}\mathbf{Q}_{vv}. \quad (8)$$

Thus, by Hoeffding's inequality,

$$\mathbb{P}\left[\left(\mathbf{A}\mathbf{A}_{uv}^T - \mathbf{Q}\mathbf{Q}_{uv}^T\right)^2 \geq 2(n-2)\log n + 2n + 4 \mid \mathbf{Q}_n\right] \leq 2n^{-4}. \quad (9)$$

We can integrate over all choices of $\mathbf{Q}_n$ so that Equation (9) holds unconditionally.

For the diagonal entries, $(\mathbf{A}\mathbf{A}_{uu}^T - \mathbf{Q}\mathbf{Q}_{uu}^T)^2 \leq n^2$ always. The diagonal terms and the $2n + 4$ terms from Equation (9) all sum to at most $3n^3 + 4n^2 \leq n^3 \log n$ for n large enough. Combining these inequalities, we get the inequality

$$\mathbb{P}\left[\|\mathbf{A}\mathbf{A}^T - \mathbf{Q}\mathbf{Q}^T\|_F^2 \geq 3n^3 \log n\right] \leq 2n^{-2}. \quad (10)$$

Applying the Borel–Cantelli Lemma gives the result. $\qquad\square$

Taking $\mathbf{Q} = \mathbf{X}\mathbf{Y}^T$ gives the following immediate corollary.

Corollary 1. It almost always holds that

$$\|\mathbf{A}\mathbf{A}^T - \mathbf{X}\mathbf{Y}^T(\mathbf{X}\mathbf{Y}^T)^T\|_F \leq \sqrt{3}n^{3/2}\sqrt{\log n} \quad (11)$$

and

$$\|\mathbf{A}^T\mathbf{A} - (\mathbf{X}\mathbf{Y}^T)^T\mathbf{X}\mathbf{Y}^T\|_F \leq \sqrt{3}n^{3/2}\sqrt{\log n}. \quad (12)$$

The next two results provide bounds on the singular values of $\mathbf{X}\mathbf{Y}^T$ and $\mathbf{A}$ based on lower bounds for the eigenvalues of $\mathbf{P}$ and the block membership probabilities.

Lemma 1. It almost always holds that $\alpha\gamma n \leq \sigma_d(\mathbf{X}\mathbf{Y}^T)$ and it always holds that $\sigma_{d+1}(\mathbf{X}\mathbf{Y}^T) = 0$ and $\sigma_1(\mathbf{X}\mathbf{Y}^T) \leq n$.

Corollary 2. It almost always holds that

$$\alpha\gamma n \leq \sigma_d(\mathbf{A}) \quad \text{and} \quad \sigma_{d+1}(\mathbf{A}) \leq 3^{1/4}n^{3/4}\log^{1/4} n \quad (13)$$

and it always holds that $\sigma_1(\mathbf{A}) \leq n$.

We note that Corollary 2 immediately suggests a consistent estimator of the rank of $\mathbf{X}\mathbf{Y}^T$ given by $\hat{d} = \max\{d' : \sigma_{d'}(\mathbf{A}) > 3^{1/4}n^{3/4}\log^{1/4} n\}$. Presently we do not investigate the use of this estimator and assume that the $d = \text{rank}\mathbf{P}$ is known.

The following is the version of the Davis–Kahan theorem (Davis and Kahan 1970) as stated in Rohe, Chatterjee, and Yu (2010).

Theorem 2 (Davis and Kahan). Let $\mathbf{H}, \mathbf{H}' \in \mathbb{R}^{n \times n}$ be symmetric, suppose $\mathcal{S} \subset \mathbb{R}$ is an interval, and suppose for some positive integer d that $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{n \times d}$ are such that the columns of $\mathbf{W}$ form an orthonormal basis for the sum of the eigenspaces of $\mathbf{H}$ associated with the eigenvalues of $\mathbf{H}$ in $\mathcal{S}$ and that the columns of $\mathbf{W}'$ form an orthonormal basis for the sum of the eigenspaces of $\mathbf{H}'$ associated with the eigenvalues of $\mathbf{H}'$ in $\mathcal{S}$. Let δ be the minimum distance between any eigenvalue of $\mathbf{H}$ in $\mathcal{S}$ and any eigenvalue of $\mathbf{H}$ not in $\mathcal{S}$. Then there exists an orthogonal matrix $\mathbf{R} \in \mathbb{R}^{d \times d}$ such that $\|\mathbf{W}\mathbf{R} - \mathbf{W}'\|_F \leq \frac{\sqrt{2}}{\delta}\|\mathbf{H} - \mathbf{H}'\|_F$.

For completeness, we provide a brief discussion of this important result in Section A.2 in the Appendix. Applying Theorem 2

and Lemma 1 to $\mathbf{A}\mathbf{A}^T$ and $\mathbf{X}\mathbf{Y}^T(\mathbf{X}\mathbf{Y}^T)^T$, we have the following result.

Lemma 2. It almost always holds that there exists an orthogonal matrix $\mathbf{R} \in \mathbb{R}^{2d \times 2d}$ such that $\|\mathbf{W}\mathbf{R} - \widetilde{\mathbf{W}}\| \leq \sqrt{2}\frac{\sqrt{6}}{\alpha^2\gamma^2}\sqrt{\frac{\log n}{n}}$.

Recall that $\mathbf{X}\mathbf{Y}^T = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$. We now provide bounds for the gaps between the at most K distinct rows of $\mathbf{U}$ and $\mathbf{V}$.

Lemma 3. It almost always holds that, for all u, v such that $\mathbf{X}_u \neq \mathbf{X}_v$, $\|\mathbf{U}_u - \mathbf{U}_v\| \geq \beta\sqrt{\alpha\gamma}n^{-1/2}$. Similarly, for all $\mathbf{Y}_u \neq \mathbf{Y}_v$, $\|\mathbf{V}_u - \mathbf{V}_v\| \geq \beta\sqrt{\alpha\gamma}n^{-1/2}$. As a result, $\|\mathbf{W}_u - \mathbf{W}_v\| \geq \beta\sqrt{\alpha\gamma}n^{-1/2}$ for all u, v such that $\tau(u) \neq \tau(v)$.

We now have the necessary ingredients to show our main result.

Proof of Theorem 1. Let $\hat{\boldsymbol{\psi}}$ and $\hat{\tau}$ satisfy the clustering criterion for $\widetilde{\mathbf{W}}$ (where $\widetilde{\mathbf{W}} = [\widetilde{\mathbf{U}}|\widetilde{\mathbf{V}}]$ takes the role of $\mathbf{Z}$ in Section 2.4). Let $\mathbf{C} \in \mathbb{R}^{n \times 2d}$ have row u given by $C_u = \hat{\psi}_{\hat{\tau}(u)}$. Then Equation (4) gives that $\|\mathbf{C} - \widetilde{\mathbf{W}}\|_F \leq \|\mathbf{W}\mathbf{R} - \widetilde{\mathbf{W}}\|_F$ as $\mathbf{W}$ has at most K distinct rows. Thus, Lemma 2 gives that

$$\|\mathbf{C} - \mathbf{W}\mathbf{R}\|_F \leq \|\mathbf{C} - \widetilde{\mathbf{W}}\|_F + \|\widetilde{\mathbf{W}} - \mathbf{W}\mathbf{R}\|_F$$
$$\leq 2^{3/2}\frac{\sqrt{6}}{\alpha^2\gamma^2}\sqrt{\frac{\log n}{n}}. \quad (14)$$

Let $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_K$ be balls of radius $r = \frac{\beta}{3}\sqrt{\alpha\gamma}n^{-1/2}$ each centered around the K distinct rows of $\mathbf{W}$. By Lemma 3, these balls are almost always disjoint.

Now note that almost always the number of rows u such that $\|\mathbf{C}_u - \mathbf{W}_u\mathbf{R}\| > r$ is at most $\frac{2^3 3^2 6}{\alpha^5 \beta^2 \gamma^5}\log n$. If this were not so, then infinitely often we would have

$$\|\mathbf{C} - \mathbf{W}\mathbf{R}\|_F > \frac{2^3 3^2 6}{\alpha^5 \beta^2 \gamma^5}\log n \frac{\beta}{3}\sqrt{\alpha\gamma}n^{-1/2}$$
$$= 2^{3/2}\frac{\sqrt{6}}{\alpha^2\gamma^2}\sqrt{\frac{\log n}{n}}, \quad (15)$$

in contradiction to Equation (14). Since $n_i > \gamma n > \frac{2^3 3^2 6}{\alpha^5 \beta^2 \gamma^5}\log n$ almost always, each ball $\mathcal{B}_i$ can contain exactly one of the K distinct rows of $\mathbf{C}$. This gives the number of misclassifications as $\frac{2^3 3^2 6}{\alpha^5 \beta^2 \gamma^5}\log n$ as desired. □

This gives that a clustering of the concatenation of the matrices $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ from the singular value decomposition gives an accurate block assignment. One may also cluster the scaled singular vectors given by $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ without a change in the order of the number of misclassifications.

4. EXTENSIONS

Corollary 3. Under the conditions of Theorem 1, let $\hat{\tau} : V \rightarrow [K]$ be a clustering of $\widetilde{\mathbf{Z}} = [\widetilde{\mathbf{X}}|\widetilde{\mathbf{Y}}]$. Then it almost always holds that

$$\min_{\pi \in \mathcal{S}_K} |\{u \in V : \pi(\hat{\tau}(u)) \neq \tau(u)\}| \leq \frac{2^3 3^2 6}{\alpha^6 \beta^2 \gamma^6}\log n. \quad (16)$$

The proof relies on the fact that the square roots of the singular values are all of the same order and differ by a multiplicative factor of at most $\sqrt{\alpha\gamma}$.

We now present consistent estimators of the parameters $\mathbf{P}$ and ρ for the stochastic blockmodel. Consider the following estimates

$$\hat{n}_k = |\{u : \hat{\tau}(u) = k\}|, \quad \hat{\rho}_k = \frac{\hat{n}_k}{n} \quad (17)$$

and

$$\hat{\mathbf{P}}_{ij} = \begin{cases} \dfrac{1}{\hat{n}_i\hat{n}_j} \displaystyle\sum_{(u,v)\in\hat{\tau}^{-1}(i)\times\hat{\tau}^{-1}(j)} \mathbf{A}_{uv}, & \text{if } i \neq j \text{ or,} \\[2ex] \dfrac{1}{\hat{n}_i^2 - \hat{n}_i} \displaystyle\sum_{(u,v)\in\hat{\tau}^{-1}(i)\times\hat{\tau}^{-1}(j)} \mathbf{A}_{uv}, & \text{if } i = j. \end{cases} \quad (18)$$

This gives the following corollary.

Corollary 4. Under the conditions of Theorem 1,

$$\min_{\pi \in \mathcal{S}_K} |\rho_i - \hat{\rho}_{\pi(i)}| \xrightarrow{\text{a.s.}} 0 \quad (19)$$

and

$$\min_{\pi \in \mathcal{S}_K} |\hat{\mathbf{P}}_{\pi(i)\pi(j)} - \mathbf{P}_{ij}| \xrightarrow{\text{a.s.}} 0 \quad (20)$$

for all i, $j \in [K]$ as $n \rightarrow \infty$.

The proof is immediate from Theorem 1 and the law of large numbers.

If we take $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\mu}})$ to be the adjacency spectral embedding of $\hat{\mathbf{P}}$, then we also have that $\hat{\boldsymbol{v}}$ and $\hat{\boldsymbol{\mu}}$ provide consistent estimates for $(\boldsymbol{v}, \boldsymbol{\mu})$, the adjacency spectral embedding of $\mathbf{P}$, in the following sense.

Corollary 5. Under the conditions of Theorem 1, with probability 1, there exists a sequence of orthogonal matrices $\mathbf{R}_1^{(n)}, \mathbf{R}_2^{(n)} \in \mathbb{R}^{d \times d}$ such that

$$\left\|\hat{\boldsymbol{v}} - \boldsymbol{v}\mathbf{R}_1^{(n)}\right\|_F \rightarrow 0 \quad \text{and} \quad \left\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\mathbf{R}_2^{(n)}\right\|_F \rightarrow 0. \quad (21)$$

The proof relies on applications of the Davis–Kahan theorem in a similar way to Lemma 2.

5. UNDIRECTED VERSION

We now present the undirected version of the stochastic blockmodel and state the main result. The setting and notation are from Section 3.1.

For the undirected version of the stochastic blockmodel, the matrix $\mathbf{P}$ is symmetric and $\mathbf{P}_{ij} = \mathbf{P}_{ji}$ gives the probability of an edge between a node in block i and a node in block j for each i, $j \in [K]$. Conditioned on τ, $\mathbf{A}_{uv}$ is a Bernoulli random variable with parameter $\mathbf{P}_{\tau(u),\tau(v)}$ for all $u \neq v \in [n]$. As $\mathbf{A}$ is symmetric, all entries of $\mathbf{A}$ are not independent, but the entries are independent provided two entries do not correspond to the same undirected edge.

For the undirected version, a reparameterization of the stochastic block model as a RDPG model as in Section 2.5 is not always possible. However, we can find $\boldsymbol{v}$, $\boldsymbol{\mu} \in \mathbb{R}^{K \times d}$ such that $\boldsymbol{v}\boldsymbol{\mu}^T = \mathbf{P}$ and $\boldsymbol{v}$ and $\boldsymbol{\mu}$ have equal columns up to a possible change in sign in each column. This means that the rows of $\boldsymbol{v}$ and $\boldsymbol{\mu}$ are distinct so it is not necessary to cluster on the concatenated embeddings. Instead, we consider clustering the rows
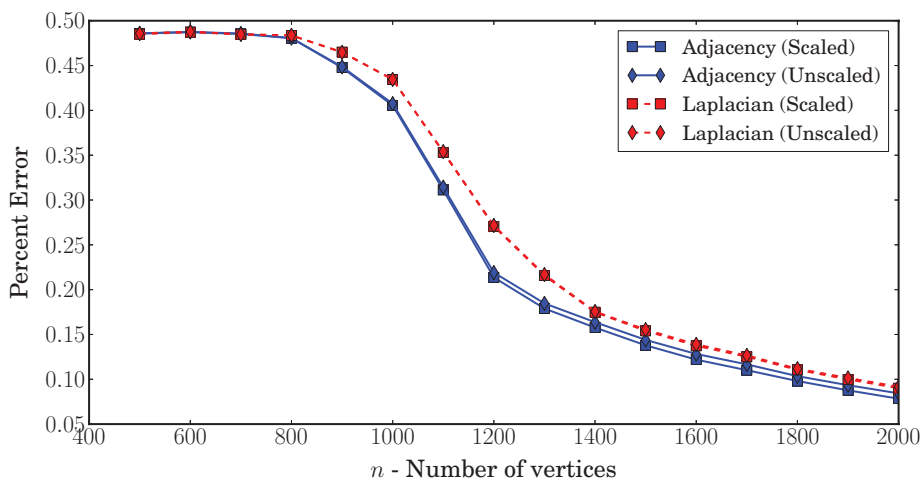
Figure 1. Mean error for 100 Monte Carlo replicates using K-means on four different embedding procedures.

of $\widetilde{\mathbf{U}}$ or $\widetilde{\mathbf{X}}$, which improves the misclassification rate by a factor of two.

Theorem 3. Under the undirected version of the stochastic blockmodel, suppose that the number of blocks K and the latent vector dimension d are known. Let $\hat{\tau} : V \mapsto [K]$ be a block assignment function according to a clustering of the rows of $\widetilde{\mathbf{U}}$ satisfying the criterion in Equation (4). It almost always holds that

$$\min_{\pi \in \mathcal{S}_K} |\{u \in V : \tau(u) \neq \pi(\hat{\tau}(u))\}| \leq \frac{2^2 3^2 6}{\alpha^5 \beta^2 \gamma^5} \log n. \quad (22)$$

Corollary 3 holds when clustering on $\widetilde{\mathbf{X}}$, with the same factor of 2 improvement in misclassification rate. Corollaries 10 and 11 also hold without change.

6. EMPIRICAL RESULTS

We evaluated this procedure and compared it to the spectral clustering procedure of Rohe, Chatterjee, and Yu (2010) for both simulated data (Section 6.1) and using a Wikipedia hyperlink graph (Section 6.2).

6.1 Simulated Data

To illustrate the effectiveness of the adjacency spectral embedding, we simulate random undirected graphs generated from the following stochastic blockmodel:

$$\mathbf{P} = \begin{pmatrix} 0.42 & 0.42 \\ 0.42 & 0.5 \end{pmatrix} \text{ and } \boldsymbol{\rho} = (0.6, 0.4)^T. \quad (23)$$

For each $n \in \{500, 600, \dots, 2000\}$, we simulated 100 Monte Carlo replicates from this model conditioned on the fact that $|\{u \in [n] : \tau(u) = i\}| = \rho_i n$ for each $i \in \{1, 2\}$. In this model, we assume that $d = 2$ and $K = 2$ are known.

We evaluated four different embedding procedures, and for each embedding, we used K-means clustering, which attempts to iteratively find the solution to Equation (4), to generate the node assignment function $\hat{\tau}$. The four embedding procedures are the scaled and unscaled adjacency spectral embedding as well as the scaled and unscaled Laplacian spectral embedding. The

Laplacian spectral embedding uses the same spectral decomposition but works with the normalized Laplacian [as defined in Rohe, Chatterjee, and Yu (2010)] rather than the adjacency matrix. The normalized Laplacian is given by $\mathbf{L} = \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$ where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is diagonal with $\mathbf{D}_{vv} = \deg(v)$, the degree of node v.

We evaluated the performance of the node assignments by computing the percentage of misassigned nodes, $\min_{\pi \in S_2} |\{u \in [n] : \tau(u) \neq \pi(\hat{\tau}(u))\}|/n$, as in Equation (6). Figure 1 demonstrates that the performance of K-means on all four embeddings improves with increasing number of nodes. It also demonstrates (via a paired Wilcoxon test) that for these model parameters, the adjacency embedding is superior to the Laplacian embeddings for large n. In fact, for $n \geq 1400$, we observed that for each simulated graph, the scaled adjacency embedding always performed better than both Laplacian embeddings. We note that these model parameters were specifically constructed to demonstrate a case where the adjacency embedding is superior to the Laplacian embedding.

Figure 2 shows an example of the scaled adjacency (left) and the scaled Laplacian (right) spectral embeddings. The graph has 2000 nodes and the points are colored according to their block membership. The dashed line shows the discriminant boundary given by the K-means algorithm with $K = 2$.

6.2 Wikipedia Graph

For this data, each node in the graph corresponds to a Wikipedia page and the edges correspond to the presence of a hyperlink between two pages (in either direction). We consider this as an undirected graph. Every article within two hyperlinks of the article "Algebraic Geometry" was included as a node in the graph. This resulted in $n = 1382$ nodes. Additionally, each document, and hence each node, was manually labeled as one of the following: Category, Person, Location, Date, and Math.

To illustrate the utility of this algorithm, we embedded this graph using the scaled adjacency and Laplacian procedures. Figure 3 shows the two embeddings for $d = 2$. The points are colored according to their manually assigned labels. First we note that on the whole, the two embeddings look moderately different. In fact, for the adjacency embedding, one can see that
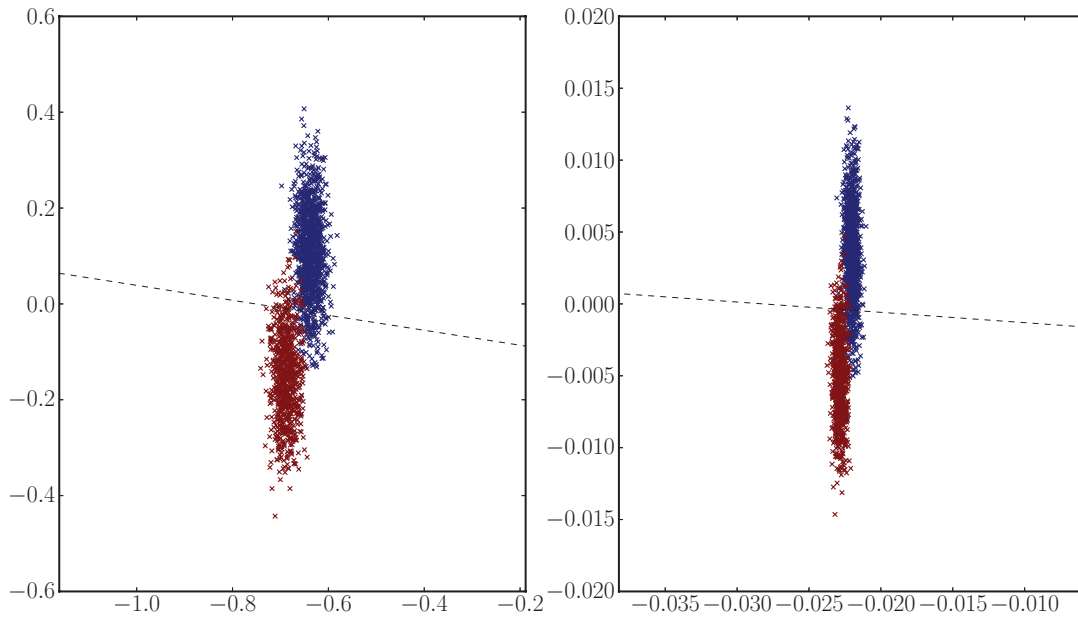
Figure 2. Scatterplots of the scaled adjacency (left) and Laplacian (right) embeddings of a 2000 node graph.

the orange points are well separated from the remaining data. On the other hand, with the Laplacian embedding, we can see that the red points are somewhat separated from the remaining data. The dashed lines show the result boundary as determined by K-means with $K = 2$.

To evaluate the performance, we considered the five different tasks of identifying one block and grouping the remaining blocks together. For each of the five blocks, we compared each of the one-versus-all block labels with the estimated labels from K-means, with $K = 2$, on the two embeddings. Table 1 shows the number of incorrectly assigned nodes, as in Equation (6), as well as the adjusted Rand index (ARI) (Hubert and Arabie 1985). The ARI has the property that the optimal value is 1 and

a value of zero indicates the expected value if the labels were assigned randomly.

We can see from this table that K-means on the adjacency embedding identifies the separation of the Date block from the other four, while on the Laplacian embedding, K-means identifies the separation of the Math block from the other four. This indicates that for this dataset (and indeed more generally), the choice of embedding procedure will depend greatly on the desired exploitation task.

We note that for both embeddings, the clusters generated using K-means, with $K = 5$, poorly reflect the manually assigned block memberships. We have not investigated beyond the illustrative two-dimensional embeddings.
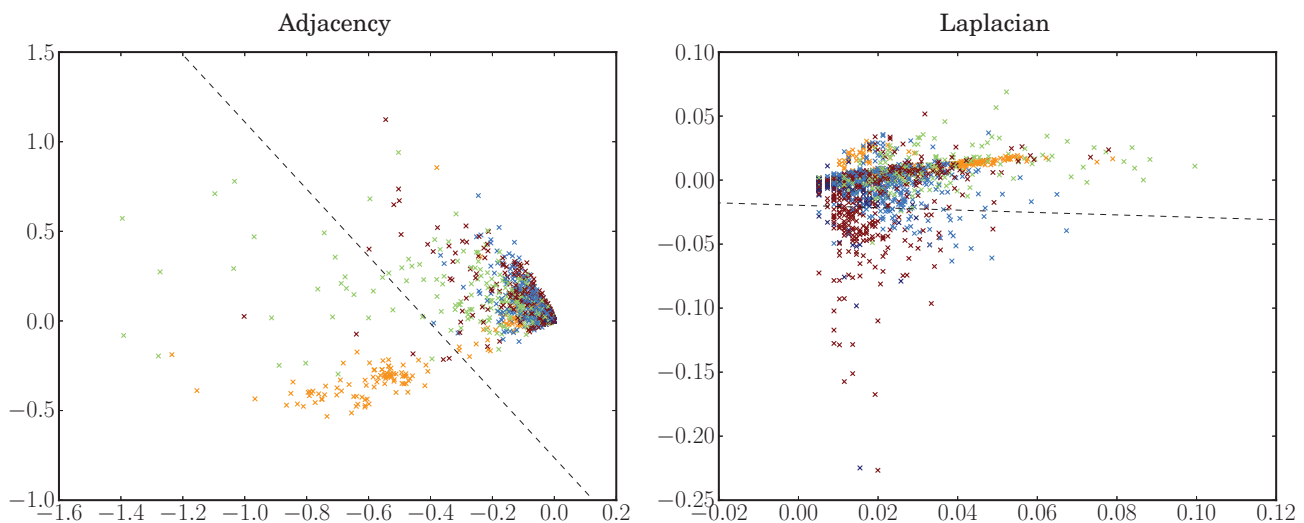

Figure 3. Scatterplots for the Wikipedia graph. The left panel shows the scaled adjacency embedding, and the right panel shows the scaled Laplacian embedding. Each point is colored according to the manually assigned labels. The dashed line represents the discriminant boundary determined by K-means with $K = 2$.

Table 1. One versus all comparison of each block against the estimated K-means block assignments with $K = 2$

	Category (119)		Person (372)		Location (270)		Date (191)		Math (430)	
	Error	ARI	Error	ARI	Error	ARI	Error	ARI	Error	ARI
A	242	−0.08	495	−0.07	341	0.01	**130**	**0.47**	543	0.06
L	299	−0.02	495	−0.02	476	−0.1	401	−0.10	**350**	**0.19**

NOTE: A, adjacency embedding; L, Laplacian embedding. Bold formatting indicates highest ARI, for each row (A versus L).

7. DISCUSSION

Our simulations demonstrate that for a particular example of the stochastic blockmodel, the proportion of misassigned nodes will rapidly become small. Though our bound shows that the number of misassigned nodes will not grow faster than $O(\log n)$, in some instances, this bound may be very loose. We also demonstrate that using the adjacency embedding over the Laplacian embedding can provide performance improvements in some settings. It is also clear from Figure 2 that the use of other unsupervised clustering techniques, such as Gaussian mixture modeling, will likely lead to further performance improvements.

On the Wikipedia graph, the two-dimensional embedding demonstrates that the adjacency embedding procedure provides an alternative to the Laplacian embedding and the two may have fundamentally different properties. Both the Date and the Math block have some differentiating structure in the graph, but these structures are illuminated more in one embedding than the other. This analysis suggests that further investigations into comparisons between the adjacency embedding and the Laplacian embeddings will be fruitful.

Our empirical analysis indicates that while the adjacency spectral embeddings and the Laplacian spectral embeddings are strongly related, the two embeddings may emphasize different aspects of the particular graph. Rohe, Chatterjee, and Yu (2010) used similar techniques to show consistency of block assignment on the Laplacian embedding and achieved the same asymptotic rates of misclassification. Indeed, if one considers the embedding given by $\mathbf{D}^{-1/2}\widetilde{\mathbf{X}}$, then this embedding will be very close to the scaled Laplacian embedding and may provide a link between the two procedures.

Note that consistent block assignments are possible using either the singular vectors or the scaled version of the singular vectors. The singular vectors themselves are essentially a whitened version of scaled singular vectors. Since the singular vectors are orthogonal, the estimated covariance of rows of the scaled vectors is proportional to the diagonal matrix given by the singular values of $\mathbf{A}$. This suggests that clustering using a criterion invariant to coordinate-scalings and rotations will likely have similar asymptotic properties.

Critical to the proof is the bound provided by Proposition 2. Since this bound does not depend on the method for generating $\mathbf{Q}$, it suggests that extensions to this theorem are possible. One such extension is to take the number of blocks $K = K_n$ to go slowly to infinity. For K_n growing, the parameters α, β, and γ are no longer constant in n, so we must impose conditions on these parameters. If we take d fixed and assume that these parameters go to zero slowly, it is possible to allow $K_n = n^\epsilon$ for ϵ sufficiently small. Under these conditions, it can be shown

that the number of incorrect block assignments is $o(n\gamma)$, which is negligible to block sizes. Our proof technique breaks down for $K_n = \Omega(n^{1/4})$, as Proposition 2 no longer implies a gap in the singular values of $\mathbf{A}$.

To avoid the model selection quagmire, we assumed in Theorem 1 that the number of blocks K and the latent feature dimension d are known. However, the proof of this theorem suggests that both K and d can be estimated consistently. Corollary 2 shows that all but d of the singular values of $\mathbf{A}$ are less than $3^{1/4}n^{3/4}\log^{1/4} n$ for n large enough. As discussed earlier, this shows that $\hat{d} = \max\{i : \sigma_i(\mathbf{A}) > 3^{1/4}n^{3/4}\log^{1/4} n\}$ will be a consistent estimator for d. Though this estimator is consistent, the required number of nodes for it to become accurate will depend highly on the sparsity of the graph, which controls the magnitude of the largest singular values of $\mathbf{A}$. Furthermore, our bounds suggest that the number of nodes required for this estimate to be accurate will increase exponentially as the expected graph density decreases.

Estimating K is more complicated, and we do not present a formal method to do so. We do note that the proof shows that most of the embedded vectors are concentrated around K separated points. An appropriate covering of the points by slowly vanishing balls would allow for a consistent estimate of K. More work is needed to provide model selection criteria that are practical to the practitioner.

Note that some practitioners may have estimates or bounds for the parameters $\mathbf{P}$ and æ, derived from some prior study. In this case, provided bounds on α, β, and γ can be determined, the proof can be used to derive high probability bounds on the number of nodes that have been assigned to the incorrect block. This may also enable the practitioner to choose n to optimize some mis-assignment and cost criteria.

The proofs above would remain valid if the diagonals of the adjacency matrix are modified provided that each modification is bounded. In fact, modifying the diagonals may improve the embedding to give lower numbers of mis-assignments. Marchette, Priebe, and Coppersmith (2011) suggested replacing the diagonal element $\mathbf{A}_{uu}$ with $\deg(u)/(n-1)$ for each node $u \in [n]$. Scheinerman and Tucker (2010) provided an iterative algorithm to impute the diagonal. An optimal choice of the diagonal is not known for general stochastic blockmodels.

Another practical concern is the possibility of missing data in the observed graph. One example may be that each edge in the true graph is only observed with probability p in the observed graph. Our theory will be unaffected by this type of error since the observed graph is also distributed according to a stochastic blockmodel with edge probabilities $\mathbf{P}' = p\mathbf{P}$. As a result, asymptotic consistency remains valid. We may also allow p to decrease slowly with n and still achieve asymptotically

negligible misassignments. However, typically the finite sample performance will suffer if p is small.

Overall, the theory and results presented suggest that this embedding procedure is worthy of further investigation. The problems of estimating K and d, choosing between scaled and unscaled embedding, and choosing between the adjacency and the Laplacian will all be considered in future work. This work is also being generalized to more other latent position models.

Finally, under the stochastic blockmodel, our method will be less computationally demanding than ones that depend on maximizing likelihood or modularity criterion. Fast methods to compute singular value decompositions are possible, especially for sparse matrices. There are a plethora of methods for efficiently clustering points in Euclidean space. Overall, this embedding method may be valuable to the practitioner to provide a rapid method to identify blocks in networks.

APPENDIX

A.1 Proofs of Technical Lemmas

In this appendix, we prove the technical results stated in Section 3.2.

Lemma 1. It almost always holds that $\alpha \gamma n \leq \sigma_d(\mathbf{X}\mathbf{Y}^T)$ and it always holds that $\sigma_{d+1}(\mathbf{X}\mathbf{Y}^T) = 0$ and $\sigma_1(\mathbf{X}\mathbf{Y^T}) \leq n$.

Proof. Since $\mathbf{X}\mathbf{Y}^T \in [0, 1]^{n \times n}$, the nonnegative matrix $\mathbf{X}\mathbf{Y}^T(\mathbf{X}\mathbf{Y}^T)^T$ has entries bounded by n. The row sums are bounded by n^2 giving that $\sigma_1^2(\mathbf{X}\mathbf{Y}^T) = \lambda_1(\mathbf{X}\mathbf{Y}^T(\mathbf{X}\mathbf{Y}^T)^T) \leq n^2$. Since $\mathbf{X}$ and $\mathbf{Y}$ are at most rank d, we have $\sigma_{d+1}(\mathbf{X}\mathbf{Y}) = 0$.

The nonzero eigenvalues of $\mathbf{X}\mathbf{Y}^T(\mathbf{X}\mathbf{Y}^T)^T = \mathbf{X}\mathbf{Y}^T\mathbf{Y}^T\mathbf{X}$ are the same as the nonzero eigenvalues of $\mathbf{Y}^T\mathbf{Y}\mathbf{X}^T\mathbf{X}$. It almost always holds that $n_i \geq \gamma n$ for all i so that

$$\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{K} n_i \boldsymbol{v}_i \boldsymbol{v}_i^T = \gamma n \boldsymbol{v}^T \boldsymbol{v} + \sum_{i=1}^{K} (n_i - \gamma n) \boldsymbol{v}_i \boldsymbol{v}_i^T \qquad (A.1)$$

is the sum of two positive semidefinite matrices, the first of which has eigenvalues all greater than $\alpha \gamma n$. This gives $\lambda_d(\mathbf{X}^T\mathbf{X}) \geq \alpha \gamma n$ and similarly $\lambda_d(\mathbf{Y}^T\mathbf{Y}) \geq \alpha \gamma n$. This gives that $\mathbf{Y}^T\mathbf{Y}\mathbf{X}^T\mathbf{X}$ is the product of positive definite matrices. We then use a bound on the smallest eigenvalues of the product of two positive semidefinite matrices so that $\lambda_d(\mathbf{Y}^T\mathbf{Y}\mathbf{X}^T\mathbf{X}) \geq \lambda_d(\mathbf{Y}^T\mathbf{Y})\lambda_d(\mathbf{X}^T\mathbf{X}) \geq (\alpha \gamma n)^2$ (Zhang and Zhang 2006, corollary 11). This establishes $\sigma_d^2(\mathbf{X}\mathbf{Y}^T) \geq (\alpha \gamma n)^2$. $\square$

Corollary 2. It almost always holds that $\alpha \gamma n \leq \sigma_d(\mathbf{A})$ and $\sigma_{d+1}(\mathbf{A}) \leq 3^{1/4} n^{3/4} \log^{1/4} n$ and it always holds that $\sigma_1(\mathbf{A}) \leq n$.

Proof. First, by the same arguments as in Lemma 1, we have $\sigma_1(\mathbf{A}) \leq n$. By Weyl's inequality (Horn and Johnson 1985, sec. 6.3), we have that

$$|\sigma_i^2(\mathbf{A}) - \sigma_i^2(\mathbf{X}\mathbf{Y}^T)| = |\lambda_i(\mathbf{A}\mathbf{A}^T) - \lambda_i(\mathbf{X}\mathbf{Y}^T(\mathbf{X}\mathbf{Y}^T)^T)|$$
$$\leq \|\mathbf{A}\mathbf{A}^T - \mathbf{X}\mathbf{Y}^T(\mathbf{X}\mathbf{Y}^T)^T\|_F. \qquad (A.2)$$

Together with Corollary 1 this shows that $\sigma_{d+1}(\mathbf{A}) \leq 3^{1/4} n^{3/4} \log^{1/4} n$ almost always. Since $\gamma < \rho_i$ for each i, Lemma 1 can be strengthened to show that there exists $\epsilon > 0$, not dependent on n, such that $(\alpha \gamma + \epsilon) n < \sigma_d(\mathbf{X}\mathbf{Y}^T)$. Thus, we have that $(\alpha \gamma + \epsilon)^2 n^2 < \sigma_d^2(\mathbf{X}\mathbf{Y}^T)$ so that $(\alpha \gamma)^2 n^2 \leq \sigma_d^2(\mathbf{A})$ since $\sqrt{3} n^{3/2} \sqrt{\log n} < \epsilon^2 n^2$ for n large enough. $\square$

The singular value decomposition of $\mathbf{X}\mathbf{Y}^T$ is given by $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$. The next result provides bounds for the gaps between the at most K distinct rows of $\mathbf{U}$ and $\mathbf{V}$. Recall that for a matrix $\mathbf{M}$, row u is given M_u^T for all u.

Lemma 3. It almost always holds that, for all u, v such that $\mathbf{X}_u \neq \mathbf{X}_v$, $\|\mathbf{U}_u - \mathbf{U}_v\| \geq \beta \sqrt{\alpha \gamma} n^{-1/2}$. Similarly, for all $\mathbf{Y}_u \neq \mathbf{Y}_v$, $\|\mathbf{V}_u - \mathbf{V}_v\| \geq \beta \sqrt{\alpha \gamma} n^{-1/2}$. As a result, $\|\mathbf{W}_u - \mathbf{W}_v\| \geq \beta \sqrt{\alpha \gamma} n^{-1/2}$ for all u, v such that $\tau(u) \neq \tau(v)$.

Proof. Let $\mathbf{Y}^T\mathbf{Y} = \mathbf{E}\mathbf{D}^2\mathbf{E}^T$ for $\mathbf{E} \in \mathbb{R}^{d \times d}$ orthogonal, $\mathbf{D} \in \mathbb{R}^{d \times d}$ diagonal. Define $\mathbf{G} = \mathbf{X}\mathbf{E}$, $\mathbf{G}' = \mathbf{G}\mathbf{D}$, and $\mathbf{U}' = \mathbf{U}\boldsymbol{\Sigma}$. Let u, v be such that $\mathbf{X}_u \neq \mathbf{X}_v$. From Lemma 1 and its proof, diagonals of $\mathbf{D}$ are almost always at least $\sqrt{\alpha \gamma n}$ and the diagonals of $\boldsymbol{\Sigma}$ are at most n.

Now,

$$\mathbf{G}'\mathbf{G}'^T = \mathbf{G}\mathbf{D}^2\mathbf{G}^T = \mathbf{X}\mathbf{E}\mathbf{D}^2\mathbf{E}^T\mathbf{X}^T = \mathbf{X}\mathbf{Y}^T\mathbf{Y}\mathbf{X}^T$$
$$= \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^T = \mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^T = \mathbf{U}'\mathbf{U}'^T. \qquad (A.3)$$

Let $e \in \mathbb{R}^n$ denote the vector with all zeros except 1 in the uth coordinate and -1 in the vth coordinate. By the above, we have $\|\mathbf{G}'_u - \mathbf{G}'_v\|^2 = e^T\mathbf{G}'\mathbf{G}'^T e = e^T\mathbf{U}'\mathbf{U}'^T e = \|\mathbf{U}'_u - \mathbf{U}'_v\|^2$. Therefore we obtain that $\beta \leq \|\mathbf{X}_u - \mathbf{X}_v\| = \|\mathbf{G}_u - \mathbf{G}_v\| \leq \frac{1}{\sqrt{\alpha \gamma n}} \|\mathbf{G}'_u - \mathbf{G}'_v\| = \frac{1}{\sqrt{\alpha \gamma n}} \|\mathbf{U}'_u - \mathbf{U}'_v\| \leq \frac{1}{\sqrt{\alpha \gamma n}} n \|\mathbf{U}_u - \mathbf{U}_v\|$, as desired.

A symmetric argument holds for $\|\mathbf{V}_u - \mathbf{V}_v\|$. For $\|\mathbf{W}_u - \mathbf{W}_v\|$, note that if $\tau(u) \neq \tau(v)$, then either $\mathbf{U}_u \neq \mathbf{U}_v$ or $\mathbf{V}_u \neq \mathbf{V}_v$. $\square$

Lemma 2. It almost always holds that there exists an orthogonal matrix $\mathbf{R} \in \mathbb{R}^{2d \times 2d}$ such that $\|\mathbf{W}\mathbf{R} - \widetilde{\mathbf{W}}\| \leq \sqrt{2} \frac{\sqrt{6}}{\alpha^2 \gamma^2} \sqrt{\frac{\log n}{n}}$.

Proof. Let $\mathcal{S} = (\frac{1}{2}\alpha^2\gamma^2 n^2, \infty)$. By Lemma 1 and Corollary 2, it almost always holds that exactly d eigenvalues of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{X}\mathbf{Y}^T(\mathbf{X}\mathbf{Y}^T)^T$ are in $\mathcal{S}$. Additionally, Lemma 1 shows that the gap $\delta > \alpha^2\gamma^2 n^2$. Together with Corollary 1, we have that

$$\sqrt{2} \frac{\|\mathbf{A}\mathbf{A}^T - \mathbf{X}\mathbf{Y}^T(\mathbf{X}\mathbf{Y}^T)^T\|_F}{\delta} \leq \sqrt{2} \frac{\sqrt{3}n^{3/2}\sqrt{\log n}}{\alpha^2\gamma^2 n^2}. \qquad (A.4)$$

This shows that there exists an $\mathbf{R}_1 \in \mathbb{R}^{d \times d}$ such that $\|\mathbf{U}\mathbf{R}_1 - \widetilde{\mathbf{U}}\|_F \leq \frac{\sqrt{6}}{\alpha^2\gamma^2}\sqrt{\frac{\log n}{n}}$.

Now note that all of the above could be repeated for $\mathbf{A}^T\mathbf{A}$ and $(\mathbf{X}\mathbf{Y}^T)^T\mathbf{X}\mathbf{Y}^T$, to find $\mathbf{R}_2 \in \mathbb{R}^{d \times d}$ such that $\|\mathbf{V}\mathbf{R}_2 - \widetilde{\mathbf{V}}\|_F \leq \frac{\sqrt{6}}{\alpha^2\gamma^2}\sqrt{\frac{\log n}{n}}$. Taking $\mathbf{R}$ as the direct sum of $\mathbf{R}_1$ and $\mathbf{R}_2$ gives the result. $\square$

A.2 Davis–Kahan Theorem

We now state and provide a brief discussion of the Davis–Kahan theorem (Davis and Kahan 1970; Rohe, Chatterjee, and Yu 2010). First, we consider some general results from the theory of Grassman spaces (Qiu, Zhang, and Li 2005). Let $\mathcal{G}_{d,n}$ denote the set of d-dimensional subspaces of $\mathbb{R}^n$. Two important metrics on $\mathcal{G}_{d,n}$ are the *gap metric* d_g and the *Hausdorff metric* d_h, which are defined as follows. For all $\mathcal{W}, \mathcal{W}' \in \mathcal{G}_{d,n}$,

$$d_g(\mathcal{W}, \mathcal{W}') = \sqrt{\sum_{i=1}^{d} \sin^2 \theta_i(\mathcal{W}, \mathcal{W}')}$$

and

$$d_h(\mathcal{W}, \mathcal{W}') = \sqrt{\sum_{i=1}^{d} \left(2 \sin \frac{\theta_i(\mathcal{W}, \mathcal{W}')}{2}\right)^2}, \qquad (A.5)$$

where $\theta_1(\mathcal{W}, \mathcal{W}')$, $\theta_2(\mathcal{W}, \mathcal{W}')$, ... $\theta_d(\mathcal{W}, \mathcal{W}')$ denote the *principal angles* between $\mathcal{W}$ and $\mathcal{W}'$. By simple trigonometry $d_h(\mathcal{W}, \mathcal{W}') \leq \sqrt{2} \cdot d_g(\mathcal{W}, \mathcal{W}')$. Suppose $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{n,d}$ have columns that are orthonormal bases for $\mathcal{W}$ and $\mathcal{W}'$, respectively. It is well known that

$d_h(\mathcal{W}, \mathcal{W}') = \min_{\mathbf{R}} \|\mathbf{WR} - \mathbf{W}'\|_F$ where the minimum is over all orthogonal matrices $\mathbf{R} \in \mathbb{R}^{d \times d}$.

The next theorem states the original form of the theorem from Davis and Kahan (1970) followed by the version proved in Rohe, Chatterjee, and Yu (2010).

Theorem 2 (Davis and Kahan). Let $\mathbf{H}, \mathbf{H}' \in \mathbb{R}^{n \times n}$ be symmetric, suppose $\mathcal{S} \subset \mathbb{R}$ is an interval, and suppose for some positive integer d that $\mathcal{W} \in \mathcal{G}_{d,n}$ is the sum of the eigenspaces of $\mathbf{H}$ associated with the eigenvalues of $\mathbf{H}$ in $\mathcal{S}$, and that $\mathcal{W}' \in \mathcal{G}_{d,n}$ is the sum of the eigenspaces of $\mathbf{H}'$ associated with the eigenvalues of $\mathbf{H}'$ in $\mathcal{S}$. If δ is the minimum distance between any eigenvalue of $\mathbf{H}$ in $\mathcal{S}$ and any eigenvalue of $\mathbf{H}$ not in $\mathcal{S}$, then $\delta \cdot d_g(\mathcal{W}, \mathcal{W}') \leq \|\mathbf{H} - \mathbf{H}'\|_F$.

Furthermore, suppose $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{n \times d}$ are such that the columns of $\mathbf{W}$ form an orthonormal basis for $\mathcal{W}$ and that the columns $\mathbf{W}'$ form an orthonormal basis for $\mathcal{W}'$. Then there exists an orthogonal matrix $\mathbf{R} \in \mathbb{R}^{d \times d}$ such that $\|\mathbf{WR} - \mathbf{W}'\|_F \leq \frac{\sqrt{2}}{\delta} \|\mathbf{H} - \mathbf{H}'\|_F$.

From the preceding analysis, we see that the version from Rohe, Chatterjee, and Yu (2010) follows from the original theorem; indeed, we have for some orthogonal $\mathbf{R} \in \mathbb{R}^{d \times d}$ that $\|\mathbf{WR} - \mathbf{R}'\|_F = d_h(\mathcal{W}, \mathcal{W}') \leq \sqrt{2} d_g(\mathcal{W}, \mathcal{W}') \leq \frac{\sqrt{2}}{\delta} \|\mathbf{H} - \mathbf{H}'\|_F$.

REFERENCES

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Membership Stochastic Blockmodels," *The Journal of Machine Learning Research*, 9, 1981–2014. [1119]

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(50), 21068–21073. [1120]

Choi, D. S., Wolfe, P. J., and Airoldi, E. M. (2010), "Stochastic Blockmodels With Growing Number of Classes," *Biometrika*, 99, 273–284. [1120]

Condon, A., and Karp, R. M. (2001), "Algorithms for Graph Partitioning on the Planted Partition Model," *Random Structures and Algorithms*, 18(2), 116–140. [1120]

Davis, C., and Kahan, W. (1970), "The Rotation of Eigenvectors by a Pertubation. III," *SIAM Journal on Numerical Analysis*, 7, 1–46. [1122,1127,1128]

Eckart, C., and Young, G. (1936), "The Approximation of One Matrix by Another of Lower Rank," *Psychometika*, 1(3), 211–218. [1121]

Fjallstrom, P. (1998), "Algorithms for Graph Partitioning: A Survey," *Computer and Information Science*, 3(10). Available at *http://www.ep.liu.se/ea/cis/1998/010/index.html*. [1119]

Fortunato, S. (2010), "Community Detection in Graphs," *Physics Reports*, 486(3–5), 75–174. [1119]

Goldenberg, A. (2009), "A Survey of Statistical Network Models," *Foundations and Trends in Machine Learning*, 2(2), 129–233. [1119]

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007), "Model-Based Clustering for Social Networks," *Journal of the Royal Statistical Society,* Series A, 170(2), 301–354. [1119]

Hoff, P., Rafferty, A., and Handcock, M. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, 1090–1098. [1119]

Holland, P. W., Laskey, K., and Leinhardt, S. (1983), "Stochastic Blockmodels: First Steps," *Social Networks*, 5(2), 109–137. [1119,1120]

Horn, R., and Johnson, C. (1985), *Matrix Analysis*, Cambridge: Cambridge University Press. [1127]

Hubert, L., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2(1), 193–218. [1125]

Marchette, D., Priebe, C., and Coppersmith, G. (2011), "Vertex Nomination via Attributed Random Dot Product Graphs," in *Proceedings of the 57th ISI World Statistics Congress*. [1121,1126]

McSherry, F. (2001), "Spectral Partitioning of Random Graphs," in *Proceedings 2001 IEEE International Conference on Cluster Computing*, pp. 529–537. [1120]

Newman, M., and Girvan, M. (2004), "Finding and Evaluating Community Structure in Networks," *Physical Review*, 69(2), 1–15. [1120]

Nowicki, K., and Snijders, T. A. B. (2001), "Estimation and Prediction for Stochastic Blockstructures," *Journal of the American Statistical Association*, 96(455), 1077–1087. [1119]

Qiu, L., Zhang, Y., and Li, C.-K. (2005), "Unitarily Invariant Metrics on the Grassmann Space," *SIAM Journal on Matrix Analysis and Applications*, 27(2), ISSN: 08954798, doi: 10.1137/040607605. [1127]

Rohe, K., Chatterjee, S., and Yu, B. (2010), "Spectral Clustering and the High-Dimensional Stochastical Block Model," Technical Report 791, University of California, Berkeley. [1120,1122,1124,1126,1127,1128]

Scheinerman, E., and Tucker, K. (2010), "Modeling Graphs Using Dot Product Representations," *Computational Statistics*, 25(1), 1–16. [1126]

Snijders, T., and Nowicki, K. (1997), "Estimation and Prediction for Stochastic Blockmodels for Graphs With Latent Block Structure," *Journal of Classification*, 14, 75–100. [1119,1120]

Wang, Y. J., and Wong, G. Y. (1987), "Stochastic Blockmodels for Directed Graphs," *Journal of the American Statistical Association*, 82(397), 8–19. [1119,1120]

Young, S., and Scheinerman, E. (2007), "Random Dot Product Models for Social Networks," in *Proceedings of the 5th International Conference on Algorithms and Models for the Web-Graph*, pp. 138–149. [1119]

Zhang, F., and Zhang, Q. (2006), "Eigenvalue Inequalities for Matrix Product," *IEEE Transactions on Automatic Control*, 51(9), 1506–1509. [1127]