Received: 6 August 2018 | Accepted: 27 February 2022

DOI: 10.1111/rssb.12509

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Series B Statistical Methodology

A statistical interpretation of spectral embedding: The generalised random dot product graph

Patrick Rubin-Delanchy[1] | **Joshua Cape**[2] | **Minh Tang**[3] | **Carey E. Priebe**[4]

[1]School of Mathematics, University of Bristol, Bristol, UK

[2]University of Pittsburgh, Pittsburgh, Pennsylvania, USA

[3]North Carolina State University, Raleigh, North Carolina, USA

[4]Johns Hopkins University, Baltimore, Maryland, USA

Correspondence
Patrick Rubin-Delanchy, School of Mathematics, University of Bristol, Bristol, UK.
Email: patrick.rubin-delanchy@bristol.ac.uk

Abstract

Spectral embedding is a procedure which can be used to obtain vector representations of the nodes of a graph. This paper proposes a generalisation of the latent position network model known as the random dot product graph, to allow interpretation of those vector representations as latent position estimates. The generalisation is needed to model heterophilic connectivity (e.g. 'opposites attract') and to cope with negative eigenvalues more generally. We show that, whether the adjacency or normalised Laplacian matrix is used, spectral embedding produces uniformly consistent latent position estimates with asymptotically Gaussian error (up to identifiability). The standard and mixed membership stochastic block models are special cases in which the latent positions take only K distinct vector values, representing communities, or live in the $(K-1)$-simplex with those vertices respectively. Under the stochastic block model, our theory suggests spectral clustering using a Gaussian mixture model (rather than K-means) and, under mixed membership, fitting the minimum volume enclosing simplex, existing recommendations previously only supported under non-negative-definite assumptions. Empirical improvements in link prediction (over the random dot product graph), and the potential to uncover richer latent structure (than posited under the standard or mixed membership stochastic block models) are demonstrated in a cyber-security example.

KEYWORDS

graph embedding, networks, spectral clustering, stochastic block model

1 | INTRODUCTION

While the study of graphs is well-established in Mathematics and Computer Science, it has only more recently become mainstream in Statistics, a shift driven at least in part by the advent of the Internet (Newman, 2018). Yet, despite its breadth, translating existing graph theory into principled statistical procedures has produced many new mathematical challenges.

An example pertinent to this paper is the spectral clustering procedure (Von Luxburg, 2007). This procedure, which aims to find network communities, generally proceeds along the following steps. Given an undirected graph, the corresponding adjacency or normalised Laplacian matrix is first constructed. Next, the graph is *spectrally embedded* into d dimensions by computing the d principal eigenvectors of the matrix—in our case scaled according to eigenvalue—to obtain a d-dimensional vector representation of each node. The first scaled eigenvector can be thought to provide the x-coordinate of each node, the second the y-coordinate, and so forth. Finally, these points are input into a clustering algorithm such as K-means (Lloyd, 1982; Steinhaus, 1956) to obtain communities. The most popular justification for this algorithm, put forward by Shi and Malik (2000) based on earlier work by Donath and Hoffman (1973) and Fiedler (1973), is its solving a convex relaxation of the normalised cut problem. A more principled statistical justification was finally found by Rohe et al. (2011), see also Lei and Rinaldo (2015), showing that the spectral clustering algorithm provides consistent identification of communities under the stochastic block model (Holland et al., 1983). Their analysis, however, demands that eigenvectors corresponding to the largest *magnitude* eigenvalues are used, countering earlier and contemporary papers which recommend using only the positive.

The random dot product graph is a model which allows statistical interpretation of spectral embedding as a standalone procedure, that is, without the subsequent clustering step. Through this broader view of spectral embedding, one finds that geometric analyses other than clustering are also productive. For example, simplex-fitting (Rubin-Delanchy et al., 2017) and spherical clustering (Lei & Rinaldo, 2015; Lyzinski et al., 2014; Passino et al., 2022; Qin & Rohe, 2013), respectively, are appropriate under the mixed membership (Airoldi et al., 2008) and degree-corrected (Karrer & Newman, 2011) stochastic block models, and manifold fitting is appropriate under several other random graph models (Athreya et al., 2021; Rubin-Delanchy, 2020; Trosset et al., 2020; Whiteley et al., 2021). However, the random dot product graph has an important shortcoming, addressed in this paper, which is to make a positive-definite assumption, consistent with the aforementioned practice of retaining only the positive eigenvalues, that is problematic for modelling several common types of graph connectivity structure.

The limitations of this positive-definite assumption are easy to identify using a stochastic block model. In standard form, this model posits that there is a partition of the nodes into K

communities, conditional upon which the edges occur independently according to a symmetric inter-community edge probability matrix $\mathbf{B} \in [0,1]^{K \times K}$, known as the block matrix. If the model is extended to include degree correction (Karrer & Newman, 2011), those probabilities are subject to nodewise scaling, so that for example a node i, of community 1, and a node j, of community 2, form an edge with probability $w_i w_j \mathbf{B}_{12}$. If, and only if, the block matrix is non-negative-definite, the random dot product graph can reproduce either model (Lyzinski et al., 2014). It will assign a latent position X_i to each node i which, in the standard case, is precisely one of K possible points, each representing a community. Under degree correction, the position instead lives on one of K rays emanating from the origin, with $\|X_i\|_2 \propto w_i$. In this way, nodes with larger magnitude positions tend to have larger degree.

A positive-definite block matrix is said to reflect homophilic connectivity, in which 'birds of a feather flock together'. But to encounter a block matrix that has negative eigenvalues is not unusual. With $K = 2$, negative eigenvalues will occur when $\mathbf{B}_{12} > \mathbf{B}_{11}, \mathbf{B}_{22}$, for example, under heterophilic connectivity, and may occur when $\mathbf{B}_{11} > \mathbf{B}_{12} > \mathbf{B}_{22}$, for example, under core-periphery connectivity—a densely connected core, community 1, with sparsely connected periphery, community 2 (Borgatti & Everett, 2000). In fact, when $K > 2$, negative eigenvalues may even occur when $\mathbf{B}_{ii} > \mathbf{B}_{ij}$ for each $i \neq j$, connectivity which could reasonably be considered homophilic. A reviewer gave the example

$$\mathbf{B} = 0.1 \times \begin{pmatrix} 9 & 0 & 8 \\ 0 & 6 & 5 \\ 8 & 5 & 9 \end{pmatrix},$$

which has one negative and two positive eigenvalues.

In a graph following a stochastic block model, the signs of the principal eigenvalues of the adjacency and normalised Laplacian matrices will correspond to those of $\mathbf{B}$, up to noise. In this way a graph from a two-community stochastic block model with $\mathbf{B}_{12} > \mathbf{B}_{11}, \mathbf{B}_{22}$ should present two large-magnitude eigenvalues, one positive and one negative, with the others being close to zero. To give a light-hearted real data example, consider the graph of enmities between Harry Potter characters, a publically available dataset (Evans et al., 2014). The same graph was previously studied by Mara et al. (2020), and more generally several literature studies involve analysis of character networks (Labatut & Bost, 2019). A plot of the eigenvalues of the graph adjacency matrix is shown in Figure 1. Two eigenvalues stand out in magnitude, one positive and one negative, and for the purpose of this example the remaining will be treated as noise.

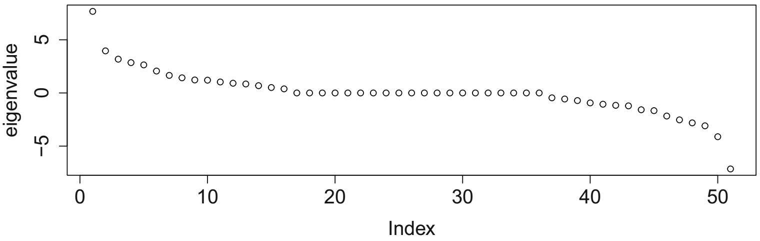

FIGURE 1 Eigenvalues of the adjacency matrix of the graph of enmities between Harry Potter characters. Of the two largest-magnitude eigenvalues, the first is positive and the second negative, and the corresponding eigenvectors are used for spectral embedding in Figure 2

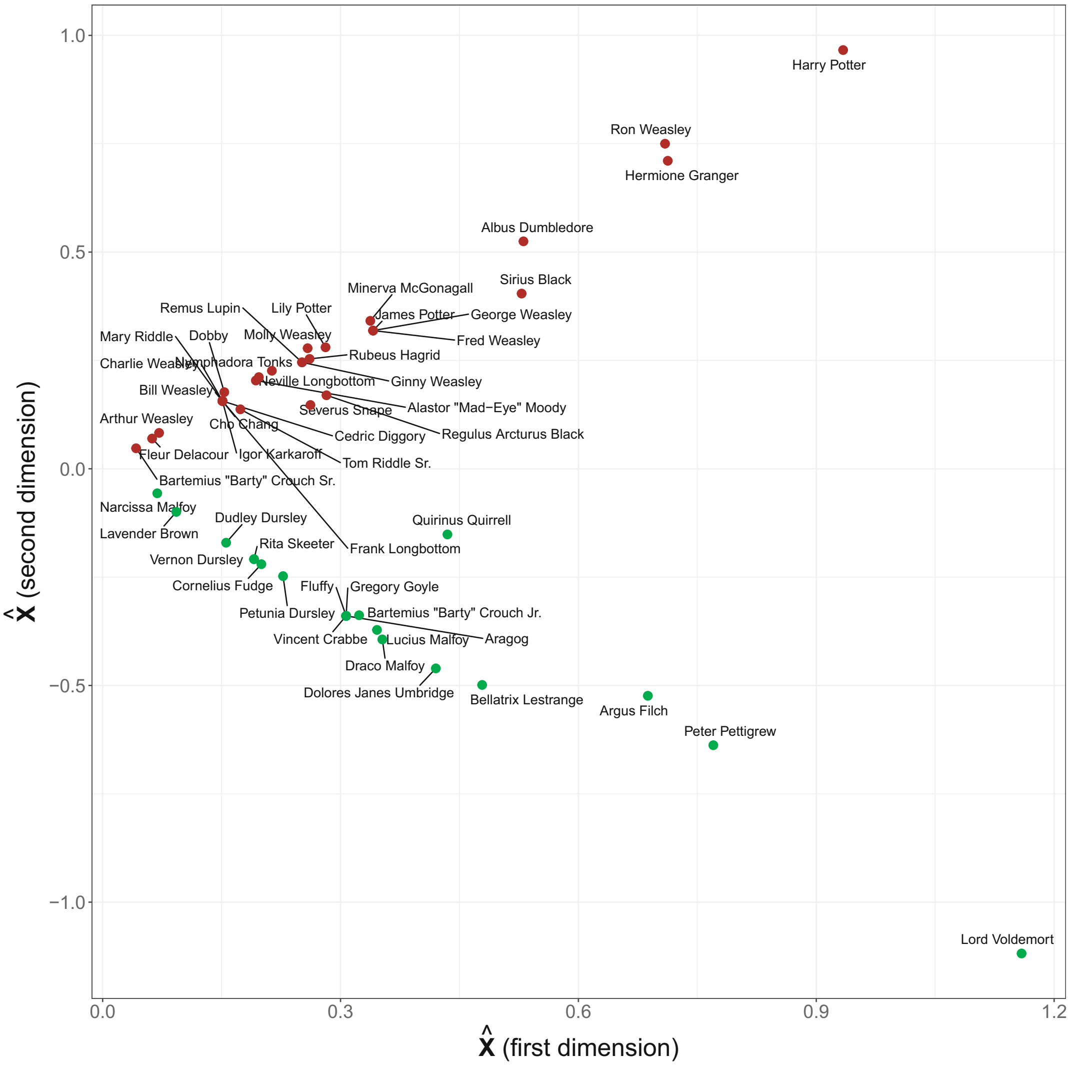

FIGURE 2 Adjacency spectral embedding into $\mathbb{R}^2$ of the graph of enmities between Harry Potter characters. This pattern of two rays is consistent with a two-community degree-corrected stochastic block model, and the points are coloured by their inferred community, estimated using spherical clustering. The embedding uses the eigenvectors corresponding one positive and one negative eigenvalue [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 2 shows the adjacency spectral embedding of the graph into two dimensions (a formal definition to follow, Definition 1), selecting eigenvectors corresponding to those eigenvalues. One can discern two rays from the origin which, as those familiar with the story will know, distinguish the good characters from the evil. This geometry is precisely what would be expected under a two-community degree-corrected stochastic block model, however, it falls outside the scope of the random dot product graph because the second eigenvector, which gives the y-axis, has a negative eigenvalue. Upon implementing spherical clustering (Lyzinski et al., 2014), we find

$$\hat{\mathbf{B}} \propto \begin{pmatrix} 0.05 & 1 \\ 1 & 0.09 \end{pmatrix},$$

which has one positive and one negative eigenvalue, and the colours of the points in the figure reflect the node partition obtained (community 1 in red, the 'good' characters; community 2 in green, the 'evil' characters). The block matrix $\mathbf{B}$ is not fully identifiable due to the presence of nodewise scaling, but the inter-to-intra community ratios are. In this way, two 'good' characters are estimated as 20 times less likely to be enemies than each would with someone 'evil', and a similarly low level of enmity between 'evil' characters is observed (the difference is not significant).

Moving beyond this toy example, a diversity of real-world graphs are surveyed in Section 5.1, finding that half present important negative eigenvalues. Following this, a closer study of computer network data is conducted in Section 5.2 giving physical reasons for heterophilic connectivity structure, and showing that including negative eigenvalues improves predictions. A model generalising the random dot product graph to allow for negative eigenvalues is therefore called for.

Our proposed model has the generic structure of a *latent position model* (Hoff et al., 2002), in that each node i is posited to have a latent position $X_i \in \mathbb{R}^d$, and two nodes i and j form an edge, conditionally independently, with probability $f(X_i, X_j)$, for some function f. A generalised random dot product graph (GRDPG) is a latent position model with $f(x, y) = x^\top \mathbf{I}_{p,q} y$, where $\mathbf{I}_{p,q} = \text{diag}(1, \ldots, 1, -1, \ldots, -1)$, with p ones followed by q minus ones on its diagonal, and where $p \geq 1$ and $q \geq 0$ are two integers satisfying $p + q = d$. When $q = 0$, the function f becomes the usual inner product on $\mathbb{R}^d$, and the model reduces to the standard random dot product graph (Athreya et al., 2017; Nickel, 2006; Young & Scheinerman, 2007).

The core asymptotic findings of this paper (Theorems 1–4) mirror existing results for the random dot product graph (Athreya et al., 2016; Cape et al., 2019a, 2019b; Lyzinski et al., 2014; Lyzinski et al., 2017; Sussman et al., 2012; Tang & Priebe, 2018): Whether the adjacency or normalised Laplacian matrix is used, the vector representations of nodes obtained by spectral embedding provide uniformly consistent and asymptotically Gaussian latent position estimates, up to identifiability, as the number of nodes goes to infinity. In this way, existing recommendations to fit a Gaussian mixture model (rather than K-means clustering) for spectral clustering under the stochastic block model (Athreya et al., 2016; Tang & Priebe, 2018), and simplex fitting under mixed membership (Rubin-Delanchy et al., 2017), previously justified only under non-negative-definite assumptions, now stand in general.

Other than the random dot product graph, there are several precursors to the GRDPG, most notably the eigenmodel of Hoff (2008) with kernel $f(x, y) = \Phi(\mu + x^\top \mathbf{\Lambda} y)$ (ignoring covariates), and the proposed random dot product graph generalisation with kernel $f(x, y) = x^\top \mathbf{\Lambda} y$ by Rohe et al. (2018), where $\mathbf{\Lambda} \in \mathbb{R}^{d \times d}$ is respectively diagonal or symmetric, μ is a scalar and Φ is the normal distribution function. Negative eigenvalues in $\mathbf{\Lambda}$ allow us to model negative eigenvalues in the graph adjacency matrix. To our best knowledge those models' connection to spectral embedding is unexplored. While the model by Rohe et al. (2018) evidently absorbs ours, the two are equivalent up to identifiability, and we will discuss how our asymptotic results can be adapted to this larger parameter space in Section 2.3. A contemporaneously written paper, Lei (2021), comes closer to our work, proposing the kernel $f(x, y) = \langle x_1, y_1 \rangle_1 - \langle x_2, y_2 \rangle_2$, where $x = (x_1, x_2)$ and $y = (y_1, y_2)$ live on the direct sum of two Hilbert spaces with respective inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$, and proving the consistency of adjacency spectral embedding in a form of Wasserstein distance. The GRDPG is a special case where the Hilbert spaces are $\mathbb{R}^p$ and $\mathbb{R}^q$, equipped with the Euclidean inner product. The advantage of Lei's analysis is to handle the infinite-dimensional case; on the other hand, our finite-dimensional results are stronger, so much that they lead to concrete methodological recommendations that could not be made based only on Lei's results. Those include to fit a Gaussian mixture model (rather than K-means clustering) for spectral clustering under the stochastic block model (calling on Theorems 2 and 4), and minimum volume simplex fitting under mixed membership (calling on Theorem 1). If the latent positions of the GRDPG are independent and identically distributed (i.i.d.), as will be assumed in our asymptotic study, the model also admits an Aldous–Hoover representation (Aldous, 1981; Hoover, 1979), wherein each node is instead independently assigned a latent position uniformly on the unit interval, and connections occur conditionally independently according to a kernel $g : [0,1]^2 \to [0,1]$, known as a graphon (Lovász, 2012). Conversely, an Aldous–Hoover graph follows a GRDPG if the integral operator associated with g has finite rank (Lei, 2021; Rubin-Delanchy, 2020). If this operator has negative eigenvalues, it cannot be reproduced by the random dot product graph or any other latent position model with positive-definite kernel.

The rest of this article is organised as follows. In Section 2, we formally describe the data envisaged, the spectral embedding procedure and the problem of finding a model-based rationale for this approach. Then we propose our solution, a generalisation of the random dot product graph model, discussing its identifiability and alternative parameterisations. Section 3 presents our asymptotic results. In Section 4, the implications of this theory for standard and mixed membership stochastic block model estimation are discussed, namely, the advantages of fitting a Gaussian mixture model over K-means for spectral clustering under the stochastic block model, and the consistency of minimum volume enclosing simplex fitting under mixed membership. In Section 5 we review a diversity of real-world graphs, showing that many exhibit important negative eigenvalues, before focussing on a cyber-security application. Section 6 concludes. All proofs are relegated to the Appendix.

2 | THE DATA, SPECTRAL EMBEDDING AND MODEL

This paper concerns statistical inference based on a single, observed graph on n nodes, labelled $1, \dots, n$. In conventional statistical terms, one may view the graph as ‘the data’, and its number of nodes as a loose substitute for ‘sample size’. The graph is represented by its adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$, where $\mathbf{A}_{ij} = 1$ if and only if there is an edge between the ith and jth node. The graph is assumed to be undirected with no-self loops or, equivalently, $\mathbf{A}$ is symmetric ($\mathbf{A} = \mathbf{A}^\top$) and hollow ($\mathbf{A}_{ii} = 0$ for all i).

To allow statistical analysis using mainstream methods (e.g. clustering), it is common to seek a vector representation of each node, and spectral embedding is a popular tool for this purpose.

Definition 1 (Adjacency and Laplacian spectral embedding into $\mathbb{R}^d$). Let $\hat{\mathbf{S}}$ be the $d \times d$ diagonal matrix containing the d largest eigenvalues of $\mathbf{A}$ *in magnitude* on its diagonal, arranged in decreasing order (based on their actual, signed, value), and let $\hat{\mathbf{U}} \in \mathbb{R}^{n\times d}$ be a matrix containing, as columns, corresponding orthonormal eigenvectors arranged in the same order. Define the adjacency spectral embedding of the graph into $\mathbb{R}^d$ as the matrix $\hat{\mathbf{X}} = [\hat{X}_1, \dots, \hat{X}_n]^\top = \hat{\mathbf{U}}|\hat{\mathbf{S}}|^{1/2} \in \mathbb{R}^{n\times d}$, that is, $\hat{\mathbf{X}}$ is a matrix whose ith row, transposed into a column vector, is $\hat{X}_i$. Similarly, let $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} \in \mathbb{R}^{n\times n}$ denote the normalised Laplacian of the graph, where $\mathbf{D} \in \mathbb{R}^{n\times n}$ is the degree matrix, a diagonal matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$ for all i, and let $\check{\mathbf{S}}, \check{\mathbf{U}}$ respectively denote the corresponding matrices of largest-magnitude eigenvalues and associated eigenvectors. Define the Laplacian spectral embedding of the graph into $\mathbb{R}^d$ by $\check{\mathbf{X}} = [\check{X}_1, \dots, \check{X}_n]^\top = \check{\mathbf{U}}|\check{\mathbf{S}}|^{1/2} \in \mathbb{R}^{n\times d}$.

In the above and hereafter, the operation $|\mathbf{M}|^{1/2}$, applied to a diagonal matrix $\mathbf{M}$, returns a diagonal matrix of the same dimension with diagonal elements $(|\mathbf{M}|^{1/2})_{ii} = |\mathbf{M}_{ii}|^{1/2}$.

The problem considered in this paper is finding a model-based rationale for spectral embedding. We seek a random graph model that defines true latent positions $X_1, \ldots, X_n \in \mathbb{R}^d$ such that $\hat{X}_i$ provides an estimate of X_i with quantifiable error. This search will also yield a suitable transformation of X_i that may be treated as the estimand of $\check{X}_i$.

As alluded to in the introduction, a relatively large body of work exists, comprehensively reviewed in Athreya et al. (2017), addressing the same problem with the eigenvalues in Definition 1 selected by largest (signed) value, in other words, leaving out negative eigenvalues and corresponding eigenvectors. To interpret such embeddings, a latent position model known as the random dot product graph (Nickel, 2006; Young & Scheinerman, 2007) is put forward and, in this model, an edge between two nodes occurs with probability given by the inner product of their latent positions. However, such a model must result in a non-negative-definite edge probability matrix $\mathbf{P}_{ij} = X_i^\top X_j$, and cannot explain significant negative eigenvalues in $\mathbf{A}$, because the matrices are related by $\mathrm{E}(\mathbf{A}|\mathbf{P}) = \mathbf{P}$, so that any difference between their spectra is due to noise.

Our solution is a model which generalises the random dot product graph, in specifying that the probability of an edge between two nodes is given by the *indefinite* inner product of their latent positions. For two vectors $x, y \in \mathbb{R}^d$, this product is $x^\top \mathbf{I}_{p,q} y$, where $\mathbf{I}_{p,q}$ is a diagonal matrix with p ones followed by q minus ones on its diagonal, and $p \geq 1$ and $q \geq 0$ are two integers satisfying $p + q = d$. A formal model definition is now given.

Definition 2 (Generalised random dot product graph model). Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ such that $x^\top \mathbf{I}_{p,q} y \in [0, 1]$ for all $x, y \in \mathcal{X}$, and $\mathcal{F}$ a joint distribution on $\mathcal{X}^n$. We say that $(\mathbf{X}, \mathbf{A}) \sim \mathrm{GRDPG}(\mathcal{F})$, with signature (p, q), if the following hold. First, let $(X_1, \ldots, X_n) \sim \mathcal{F}$, to form the latent position matrix $\mathbf{X} = [X_1, \ldots, X_n]^\top \in \mathbb{R}^{n \times d}$. Then, the graph adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$ is symmetric, hollow and, conditional on $X_1, \ldots, X_n$,

$$\mathbf{A}_{ij} \overset{\text{ind}}{\sim} \mathrm{Bernoulli}(X_i^\top \mathbf{I}_{p,q} X_j), \tag{1}$$

for all $i < j$.

2.1 | Special cases

2.1.1 | The stochastic block model

A graph follows a stochastic block model if there is a partition of the nodes into K communities, conditional upon which $\mathbf{A}_{ij} \overset{\text{ind}}{\sim} \mathrm{Bernoulli}(\mathbf{B}_{Z_i Z_j})$, for $i < j$, where $\mathbf{B} \in [0, 1]^{K \times K}$ is symmetric and $Z_i \in \{1, \ldots, K\}$ is an index denoting the community of the ith node.

Let $p \geq 1$, $q \geq 0$ denote the number of strictly positive and strictly negative eigenvalues of $\mathbf{B}$ respectively, put $d = p + q$, and choose $\mathrm{v}_1, \ldots, \mathrm{v}_K \in \mathbb{R}^d$ such that $\mathrm{v}_k^\top \mathbf{I}_{p,q} \mathrm{v}_l = \mathbf{B}_{kl}$, for $k, l \in \{1, \ldots, K\}$. We will take as a canonical choice the K rows of $\mathbf{U}_\mathbf{B} |\boldsymbol{\Sigma}_\mathbf{B}|^{1/2}$, where $\boldsymbol{\Sigma}_\mathbf{B}$ is diagonal containing the d non-zero eigenvalues of $\mathbf{B}$, and $\mathbf{B}$ has spectral decomposition $\mathbf{B} = \mathbf{U}_\mathbf{B} \boldsymbol{\Sigma}_\mathbf{B} \mathbf{U}_\mathbf{B}^\top$. It may help to remember that $p + q = d = \mathrm{rank}(\mathbf{B}) \leq K$. By letting $X_i = \mathrm{v}_{Z_i}$, we find that the graph is a GRDPG, and can set $\mathcal{X} = \{\mathrm{v}_1, \ldots, \mathrm{v}_K\}$.

2.1.2 | The mixed membership stochastic block model

Now, assign to the ith node a random probability vector $\pi_i \in \mathbb{S}^{K-1}$ where $\mathbb{S}^m$ denotes the standard m-simplex. Conditional on this assignment, let

$$\mathbf{A}_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\mathbf{B}_{Z_{i\to j}Z_{j\to i}}),$$

where

$$Z_{i\to j} \overset{\text{ind}}{\sim} \text{categorical}(\pi_i) \quad \text{and} \quad Z_{j\to i} \overset{\text{ind}}{\sim} \text{categorical}(\pi_j),$$

for $i < j$, and the distribution categorical(p), for $p \in \mathbb{S}^{K-1}$, assigns probabilities $p_1, \dots, p_K$ to the values $1, \dots, K$ respectively. The resulting graph is said to follow a mixed membership stochastic block model (Airoldi et al., 2008).

Averaging over $Z_{i\to j}$ and $Z_{j\to i}$, we can equivalently write that, conditional on $\pi_1, \dots, \pi_n$,

$$\mathbf{A}_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\pi_i^\top \mathbf{B}\pi_j).$$

But if p, q, d and $\mathrm{v}_1, \dots, \mathrm{v}_K$ are as defined previously, then $\pi_i^\top \mathbf{B}\pi_j = \left(\sum_{k=1}^K \pi_{ik}\mathrm{v}_k^\top\right)\mathbf{I}_{p,q}\left(\sum_{k=1}^K \pi_{jk}\mathrm{v}_k\right) = X_i^\top \mathbf{I}_{p,q}X_j$, where $X_i = \sum \pi_{ik}\mathrm{v}_k$. Therefore, conditional on $X_1, \dots, X_n$, Equation (1) holds, and the graph is a GRDPG with latent positions $X_1, \dots, X_n$. These live in the convex hull of $\mathrm{v}_1, \dots, \mathrm{v}_K$, a $(K-1)$-simplex if $\mathbf{B}$ has full rank ($d = K$).

The GRDPG model therefore gives the standard and mixed membership stochastic block models a natural spatial representation in which $\mathrm{v}_1, \dots, \mathrm{v}_K$ represent communities, and latent positions in between them represent nodes with mixed membership. This is illustrated in Figure 3.

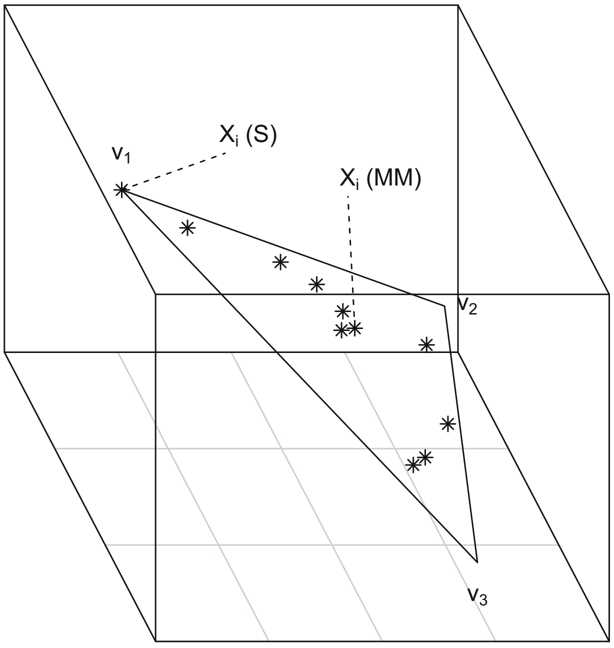

FIGURE 3 Illustration of standard (S) and mixed membership (MM) stochastic block models as special cases of the GRDPG model (Definition 2). The models have $K = 3$ communities and so the corresponding GRPDG will require $d = 3$ dimensions (or fewer, if the block matrix has low rank). The points $\mathrm{v}_1, \dots, \mathrm{v}_K$ represent communities. Under the standard stochastic block model, the ith node is assigned to a single community so that $X_i \in \{\mathrm{v}_1, \dots, \mathrm{v}_K\}$. Under mixed membership, if the ith node has a community membership probability vector π_i, then its latent position, X_i, is the corresponding convex combination of $\mathrm{v}_1, \dots, \mathrm{v}_K$

Airoldi et al. (2008) set $\pi_1, \dots, \pi_n \overset{i.i.d.}{\sim} \text{Dirichlet}(\alpha)$ for some $\alpha \in \mathbb{R}_+^K$. The corresponding latent positions $X_1, \dots, X_n$ are then (a) also i.i.d., and (b) fully supported on the convex hull of $\mathrm{v}_1, \dots, \mathrm{v}_K$. Consistency of simplex fitting, Algorithm 2 (Section 3) and illustrated in Figure 5e, relies on these two points, without requiring a Dirichlet distribution assumption.

2.1.3 | The degree-corrected stochastic block model

Instead, assign to the ith node a random weight $w_i \in [0, 1]$. Conditional on this assignment, let

$$\mathbf{A}_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(w_i w_j \mathbf{B}_{Z_i Z_j}\right),$$

for $i < j$. The resulting graph is said to follow a degree-corrected stochastic block model (Karrer & Newman, 2011).

With p, q, d and $\mathrm{v}_1, \dots, \mathrm{v}_K$ as defined previously, a corresponding GRDPG is constructed by letting $X_i = w_i \mathrm{v}_{Z_i}$, which lives on one of K rays emanating from the origin.

2.2 | Identifiability

In the definition of the GRDPG, it is clear that the conditional distribution of $\mathbf{A}$ given $X_1, \dots, X_n$ would be unchanged if $X_1, \dots, X_n$ were replaced by $\mathbf{Q}X_1, \dots, \mathbf{Q}X_n$, for any matrix $\mathbf{Q} \in \mathbb{O}(p, q) = \{\mathbf{M} \in \mathbb{R}^{d \times d} \ : \ \mathbf{M}^\top \mathbf{I}_{p,q} \mathbf{M} = \mathbf{I}_{p,q}\}$, known as the indefinite orthogonal group. The vectors $X_1, \dots, X_n$ are therefore identifiable from $\mathbf{A}$ only up to such transformation.

The property of identifiability up to *orthogonal* transformation, that is, by a matrix $\mathbf{W} \in \mathbb{O}(d) = \{\mathbf{M} \in \mathbb{R}^{d \times d} \ : \ \mathbf{M}^\top \mathbf{M} = \mathbf{I}\}$ is encountered in many statistical applications and occurs when $q = 0$. This identifiability property often turns out to be moot since inter-point distances are invariant under the action of a common orthogonal transformation, and many statistical analyses (such as K-means clustering) depend only on distance. When $q > 0$, the transformation is *indefinite orthogonal* and can affect inter-point distances. This is illustrated in Figure 4 with a GRDPG of signature (1, 2). The group $\mathbb{O}(1, 2)$ contains rotation matrices

$$r_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{bmatrix},$$

but also hyperbolic rotations

$$\rho_\theta = \begin{bmatrix} \cosh\theta & \sinh\theta & 0 \\ \sinh\theta & \cosh\theta & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

as can be verified analytically. A rotation $r_{\pi/3}$ is applied to the three GRDPG latent positions to get from the top-left to the top-right panel in Figure 4. Hyperbolic rotations ρ_θ ($\theta = 1.3$, chosen arbitrarily) and $\rho_{-\theta}$ take the positions from the top-left to the bottom-left and from the top-right

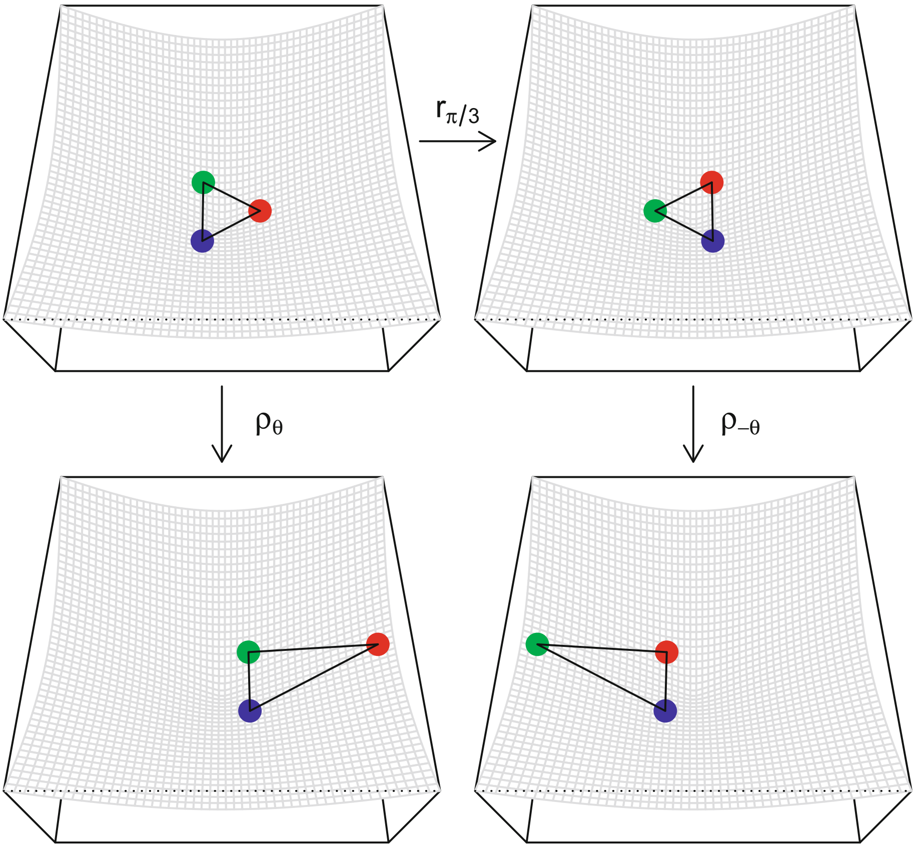

FIGURE 4 Identifiability of the latent positions of a GRDPG with signature (1, 2). In each panel, the three coloured points represent latent positions X_1, X_2 and X_3, which live inside the cone $\{x \in \mathbb{R}^3 : x^\top \mathbf{I}_{1,2} x = 0\}$ (grey mesh). The positions are only identifiable up to transformation by a matrix in the indefinite orthogonal group $\mathbb{O}(1, 2)$, including certain rotations (e.g. that used to go from the top-left to top-right panel), but also hyperbolic rotations (e.g. going from top-left to bottom-left and top-right to bottom-right). The observed graph is equally likely under those four latent position configurations, and so the configurations cannot be distinguished despite their inter-point distances varying [Colour figure can be viewed at wileyonlinelibrary.com]

to the bottom-right panels, respectively. These transformations alter inter-point distances: in the bottom row, the blue position is closer to the green on the left and closer to the red on the right; the three positions are equidistant in the top row.

2.3 | Alternative parameterisations

In this section we discuss alternative, equivalent parameterisations, explaining why we opted for the GRDPG without claiming objective superiority.

It may be observed that the only ambiguity in computing the spectral embedding $\hat{\mathbf{X}}$ is how the principal eigenvectors for $\mathbf{A}$ are chosen. If we assume repeated non-zero eigenvalues are rare in real data, this choice in practice is typically limited to the option of reversing any eigenvector. A model for interpreting spectral embedding might have been expected to reflect only this kind of ambiguity in its unidentifiability.

The model structure can indeed be brought closer to the spectral decomposition of $\mathbf{A}$ by defining an alternative, 'spectral' estimand, $\tilde{\mathbf{X}} = [\tilde{X}_1, \dots, \tilde{X}_n]^\top = \mathbf{U}|\mathbf{S}|^{1/2} \in \mathbb{R}^{n \times d}$, where $\mathbf{S} \in \mathbb{R}^{d \times d}$ is a diagonal matrix containing the non-zero eigenvalues of $\mathbf{P} = \mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^\top$, in decreasing order, and $\mathbf{U} \in \mathbb{R}^{n \times d}$ contains corresponding orthonormal eigenvectors as columns. With the help of a follow-on paper (Agterberg et al., 2020), the vector $\hat{X}_i$ will be found to estimate $\tilde{X}_i$ *up to orthogonal transformation*, uniformly and with asymptotically Gaussian error, or, under a distinct eigenvalue assumption, *up to reflection of the axes* (see Section 3.1).

The object $\tilde{\mathbf{X}}$ has special structure which, for example, precludes its rows from being i.i.d.; under a stochastic block model, the K unique vector values taken by $\tilde{X}_1, \dots, \tilde{X}_n$, representing the communities, cannot be determined from only $\mathbf{B}$, because they depend on $Z_1, \dots, Z_n$—so, for example, those K vectors may change as n grows. These are some reasons why we choose to model $\tilde{\mathbf{X}}$ through $\mathbf{X}$ (whence $\mathbf{P}$), rather than as a standalone object. However, a reader only interested in estimating $\tilde{X}_i$ can ignore indefinite orthogonal transformations, which only appear when we try to relate $\tilde{X}_i$ to X_i.

A convention could be imposed on $\mathcal{F}$ to make X_i and $\tilde{X}_i$ converge to each other, up to orthogonal transformation, and proposals to this effect were made in a follow-on paper (Agterberg et al., 2020). However, convergence is not fast enough, when it comes to the central limit theorem, to substitute $\tilde{X}_i$ by X_i and avoid indefinite orthogonal transformations (see Section 3.2). Moreover, under such a convention, the construction of vector representatives of the K communities under the stochastic block model and extensions is more involved, compared to our canonical construction in Section 2.1.1, and will vary depending on the distributions of $Z_1, \dots, Z_n$, degree-correction weights $w_1, \dots, w_n$ and community membership probabilities $\pi_1, \dots, \pi_n$.

In the context of graph simulation, Rohe et al. (2018) proposed to generalise the random dot product graph via a latent position model with kernel $f(x, y) = x^\top \boldsymbol{\Lambda} y$, where $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$ is a symmetric matrix. Alternatively, in the spirit of the eigenmodel of Hoff (2008), one might consider enforcing $\boldsymbol{\Lambda}$ to be diagonal. In either case, the model's latent positions, Y_i say, can be transformed into the latent positions of an equivalent GRDPG with signature (p, q), via $X_i = \mathbf{L}Y_i$, given a decomposition $\boldsymbol{\Lambda} = \mathbf{L}^\top \mathbf{I}_{p,q}\mathbf{L}$ for some p,q (such as that proposed for $\mathbf{B}$ in Section 2.1.1). If $\boldsymbol{\Lambda}$ is only assumed symmetric and the Y_i restricted only to give valid probabilities (i.e. to belong to a set $\mathcal{Y}$ in which $x^\top \boldsymbol{\Lambda} y \in [0, 1]$ for all $x, y \in \mathcal{Y}$), then Y_i are identifiable only up to invertible linear transformation, since we can replace $Y_1, \dots, Y_n$ with $\mathbf{M}Y_1, \dots, \mathbf{M}Y_n$ and $\boldsymbol{\Lambda}$ with $\mathbf{M}^{-\top}\boldsymbol{\Lambda}\mathbf{M}^{-1}$, for any invertible matrix $\mathbf{M}$, without changing the conditional distribution of $\mathbf{A}$. If the model of Rohe et al. (2018) is preferred, the results of Section 3.1 can be re-interpreted to say that $\hat{X}_i$ estimates Y_i *up to invertible linear transformation*, uniformly and with asymptotically Gaussian error: in the theorems of Section 3.1 we would replace $\mathbf{Q}\hat{X}_i$ with $\mathbf{L}^{-1}\mathbf{Q}\hat{X}_i$ (assuming $\boldsymbol{\Lambda}$ has full rank), X_i with Y_i, updating the covariance matrices accordingly.

3 | ASYMPTOTICS

This section describes the statistical properties of GRDPG latent position estimates obtained by spectral embedding, in an asymptotic regime where the number of nodes $n \to \infty$.

3.1 | Results for adjacency spectral embedding

In this section, the spectral estimates $\hat{X}_1, \dots, \hat{X}_n$ are shown to converge to $X_1, \dots, X_n$, in two standard statistical senses: uniformly, and with asymptotically Gaussian error. Analogous results for Laplacian spectral embedding are given in Section 3.3.

For a given n, the latent positions are assumed to be independent and identically distributed. As $n \to \infty$, their distribution is either fixed or, to produce a regime in which the average node degree grows less than linearly in n, it is made to shrink. This is done by letting $X_i = \rho_n^{1/2}\xi_i$, where $\xi_i \overset{i.i.d.}{\sim} F$, for some distribution F on $\mathbb{R}^d$, and allowing the cases $\rho_n = 1$ or $\rho_n \to 0$ sufficiently slowly. The generic joint distribution $\mathcal{F}$ occurring in Definition 2 is therefore assumed to factorise into a product of n identical marginal distributions that are equal to F up to scaling. The dimension of the model, d, is assumed to have been chosen 'economically' in the sense that, for $\xi \sim F$, the second moment matrix $\mathbf{\Delta} = \mathbb{E}(\xi\xi^\top) \in \mathbb{R}^{d\times d}$ has full rank. Here d is viewed as fixed and known, so for simplicity we suppress d-dependent factors in the statements of our theorems. Our proofs, however, keep track of d.

Since the average node degree grows as $n\rho_n$, the cases $\rho_n = 1$ and $\rho_n \to 0$ can be thought to respectively produce dense and sparse regimes and ρ_n is called a sparsity factor. No algorithm can produce uniformly consistent estimates of $X_1, \dots, X_n$ if the average node degree grows less than logarithmically. Indeed, if one did, it could be used to break the information-theoretic limit for perfect community recovery under the stochastic block model (Abbe, 2017). Our results hold under polylogarithmic growth.

Recall that we have also defined a 'spectral' estimand, $\tilde{X}_i$, identifiable up to orthogonal rather than indefinite orthogonal transformation (see Section 2.3). To move between $\hat{X}_i$, $\tilde{X}_i$ and X_i, we introduce the transformations:

1. $\mathbf{W}_\star \in \mathbb{O}(d) \bigcap \mathbb{O}(p,q)$: a block orthogonal matrix of which the first $p \times p$ block (respectively second $q \times q$ block) aligns the first p (respectively second q) columns of $\mathbf{U}$ with the first p (respectively second q) columns of $\hat{\mathbf{U}}$, solving the two orthogonal Procrustes problems independently (explicit construction in the Appendix).
2. $\mathbf{Q}_\mathbf{X} \in \mathbb{O}(p,q)$: an indefinite orthogonal matrix solving $\mathbf{X} = \tilde{\mathbf{X}}\mathbf{Q}_\mathbf{X}$.

We will find that $\mathbf{W}_\star\hat{X}_i$ converges to $\tilde{X}_i$ and that $\mathbf{Q}\hat{X}_i$ converges to X_i, where $\mathbf{Q} = \mathbf{Q}_\mathbf{X}^\top\mathbf{W}_\star$.

Theorem 1 (Uniform consistency of adjacency spectral embedding). *There exists a universal constant $c > 1$ such that, provided the sparsity factor satisfies $n\rho_n = \omega\{\log^{4c} n\}$,*

$$\max_{i\in\{1,\dots,n\}} \|\mathbf{W}_\star\hat{X}_i - \tilde{X}_i\| = O_\mathbb{P}\left(\frac{\log^c n}{n^{1/2}}\right); \quad \max_{i\in\{1,\dots,n\}} \|\mathbf{Q}\hat{X}_i - X_i\| = O_\mathbb{P}\left(\frac{\log^c n}{n^{1/2}}\right).$$

We say that a random variable Y is $O_\mathbb{P}(f(n))$ if, for any positive constant $a > 0$ there exists an integer n_0 and a constant $C > 0$ (both of which possibly depend on a) such that for all $n \geq n_0$, $|Y| \leq Cf(n)$ with probability at least $1 - n^{-a}$. We write that sequences $a_n = \omega(b_n)$ when there exist a positive constant C and an integer n_0 such that $a_n \geq Cb_n$ for all $n \geq n_0$ and $a_n/b_n \to \infty$.

We will now look at a fixed, finite subset of the nodes, indexed $1, \dots, m$ without loss of generality, to obtain a central limit theorem on the corresponding errors.

Theorem 2 (Adjacency spectral embedding central limit theorem). *Assume the same sparsity conditions as Theorem 1. Conditional on X_i, for $i = 1, \dots, m$, the random vectors $\sqrt{n}(\mathbf{Q}\hat{X}_i - X_i) = \sqrt{n}\mathbf{Q}_\mathbf{X}^\top(\mathbf{W}_\star\hat{X}_i - \tilde{X}_i)$ converge in distribution, as $n \to \infty$, to independent zero-mean Gaussian random vectors with covariance matrix $\mathbf{\Sigma}(\xi_i)$ respectively, where*

$$\mathbf{\Sigma}(x) = \begin{cases} \mathbf{I}_{p,q}\mathbf{\Delta}^{-1}\mathbb{E}[(x^\top \mathbf{I}_{p,q}\xi)(1 - x^\top \mathbf{I}_{p,q}\xi)\xi\xi^\top]\mathbf{\Delta}^{-1}\mathbf{I}_{p,q} & \textit{if } \rho_n = 1, \\ \mathbf{I}_{p,q}\mathbf{\Delta}^{-1}\mathbb{E}[(x^\top \mathbf{I}_{p,q}\xi)\xi\xi^\top]\mathbf{\Delta}^{-1}\mathbf{I}_{p,q} & \textit{if } \rho_n \to 0. \end{cases}$$

The matrix $\mathbf{Q}_\mathbf{X}$ has, since the first edition of this paper, been shown to converge (Agterberg et al., 2020), in the sense that for each n one can construct an orthogonal matrix $\mathbf{W}_{\star\star}$ such that $\mathbf{W}_{\star\star}\mathbf{Q}_\mathbf{X} \to \mathbf{Q}_0$ almost surely, where $\mathbf{Q}_0 \in \mathbb{O}(p,q)$ is a fixed matrix, made explicit in Agterberg et al. (2020). The latter observed that this made a central limit theorem 'up to orthogonal transformation' possible: in the above, we can replace $\sqrt{n}(\mathbf{Q}\hat{X}_i - X_i)$ with $\sqrt{n}\mathbf{W}_{\star\star}(\mathbf{W}_\star\hat{X}_i - \tilde{X}_i)$ and $\mathbf{\Sigma}(\xi_i)$ with $\mathbf{Q}_0^{-\top}\mathbf{\Sigma}(\xi_i)\mathbf{Q}_0^{-1}$. Moreover, if the eigenvalues of $\Delta\mathbf{I}_{p,q}$ are distinct, the matrices $\mathbf{W}_\star$ and $\mathbf{W}_{\star\star}$ can be taken to be diagonal with entries 1 or -1, reflecting that any eigenvector can be reversed in the spectral decompositions of $\mathbf{P}$ and $\mathbf{A}$.

Remark 1 (Proof overview). Theorems 1 and 2 are proved in succession within a unified framework. The proof begins with a collection of matrix perturbation decompositions which eventually yield the relation

$$\hat{\mathbf{U}}|\hat{\mathbf{S}}|^{1/2} = \mathbf{U}|\mathbf{S}|^{1/2}\mathbf{W}_\star + (\mathbf{A} - \mathbf{P})\mathbf{U}|\mathbf{S}|^{-1/2}\mathbf{W}_\star\mathbf{I}_{p,q} + \mathbf{R}$$

for some residual matrix $\mathbf{R} \in \mathbb{R}^{n\times d}$. Appropriately manipulating the above display equation subsequently yields the important identity

$$n^{1/2}(\hat{\mathbf{X}}\mathbf{W}_\star^\top\mathbf{Q}_\mathbf{X} - \mathbf{X}) = n^{1/2}(\mathbf{A} - \mathbf{P})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{I}_{p,q} + n^{1/2}\mathbf{R}\mathbf{W}_\star^\top\mathbf{Q}_\mathbf{X},$$

Theorem 1 is then established by bounding the maximum Euclidean row norm (equivalently, the two-to-infinity norm (Cape et al., 2019b)) of the right-hand side of the above display equation sufficiently tightly. Theorem 2 is established with respect to the same transformation $\mathbf{Q}$ by showing that, conditional on the ith latent position, that is, ith row of $\mathbf{X}$, the classical multivariate central limit theorem can be invoked for the ith row of the matrix $n^{1/2}(\mathbf{A} - \mathbf{P})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{I}_{p,q}$, whereas the remaining residual term has vanishing two-to-infinity norm. The technical tools involved include a careful matrix perturbation analysis involving an infinite matrix series expansion of $\hat{\mathbf{U}}$, probabilistic concentration bounds for $(\mathbf{A} - \mathbf{P})^k\mathbf{U}, 1 \le k \le \log n$, delicately passing between norms, and indefinite orthogonal matrix group considerations.

The joint proof of Theorems 1 and 2 captures the novel techniques and necessary additional considerations for moving beyond random dot product graphs considered in previous work to generalised random dot product graphs. The proofs of Theorem 3 and 4 (for Laplacian spectral embedding), while laborious, follow *mutatis mutandis* by applying the aforementioned proof considerations within the earlier work and context of the Laplacian spectral embedding limit theorems proven in Tang and Priebe (2018). For this reason, we elect to state those theorems without proof.

3.2 | Discussion

Could we remove all notion of indefinite orthogonal transformation from our results? The answer is yes if we consider only the 'spectral' estimand, $\tilde{X}_i$, but we have not found a way of describing $\hat{X}_i$ as asymptotically 'Gaussian with mean X_i', where X_i are i.i.d., regardless of sparsity, without invoking indefinite orthogonal transformations. Among equivalent latent position distributions—equal up to indefinite orthogonal transformation (by push-forward)—we can choose F such that $\mathbf{W}_{\star\star}\tilde{X}_i$ and X_i are asymptotically equal (Agterberg et al., 2020). However, the matrix $\mathbf{Q}_\mathbf{X}$ does not appear to converge faster than $n^{-1/2}$. As a result, however we choose F, the error between $\mathbf{W}_{\star\star}\tilde{X}_i$ and X_i may not vanish when scaled by $\sqrt{n}$.

It is perhaps remarkable how often the presence of indefinite orthogonal transformation will turn out not to matter. First, a follow-on inference procedure, for example, for cluster analysis, may happen to be invariant to such a transformation of its input data. A key example is fitting a Gaussian mixture model, which we will shortly discuss in more detail. Second, even if the follow-on procedure is not invariant, it may still be consistent as $n \to \infty$. Indeed, there is nothing in our results disputing the consistency of spectral clustering using K-means clustering (Rohe et al., 2011). Our uniform consistency result allows us to reprove this, and in the same movement prove the consistency of simplex fitting (Section 4.2) under the mixed membership stochastic block model, given some control on the behaviour of $\mathbf{Q}$, which we now provide.

Lemma 1 *The matrix* $\mathbf{Q}_\mathbf{X}$ *has bounded spectral norm almost surely.*

The same can be said of $\mathbf{Q}$, since $\mathbf{W}_*$ is orthogonal, and of $\mathbf{Q}^{-1} = \mathbf{I}_{p,q}\mathbf{Q}^\top\mathbf{I}_{p,q}$.

Proof of Lemma 1. The matrices $\mathbf{S}$ and $\mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^\top$ have common spectrum by definition which is further equivalent to the spectrum of $\mathbf{X}^\top\mathbf{X}\mathbf{I}_{p,q}$, since for any conformable matrices $\mathbf{M}_1, \mathbf{M}_2$, $\text{spec}(\mathbf{M}_1\mathbf{M}_2) = \text{spec}(\mathbf{M}_2\mathbf{M}_1)$, excluding zero-valued eigenvalues. By the law of large numbers, $(n\rho_n)^{-1}(\mathbf{X}^\top\mathbf{X}) \to \mathbb{E}(\xi\xi^\top)$ almost surely, and so $(n\rho_n)^{-1}(\mathbf{X}^\top\mathbf{X}\mathbf{I}_{p,q}) \to \mathbb{E}(\xi\xi^\top)\mathbf{I}_{p,q}$. It follows that both $(n\rho_n)^{-1}\|\mathbf{X}^\top\mathbf{X}\|$ and $(n\rho_n)^{-1}\min_i|\mathbf{S}_{ii}|$ converge to positive constants almost surely as $n \to \infty$.

Now for $\mathbf{Q}_\mathbf{X}$ as in the hypothesis, with respect to Loewner order $\mathbf{Q}_\mathbf{X}^\top(\min_i|\mathbf{S}_{ii}|\mathbf{I})\mathbf{Q}_\mathbf{X} \leq \mathbf{Q}_\mathbf{X}^\top|\mathbf{S}|\mathbf{Q}_\mathbf{X}$, where $\mathbf{Q}_\mathbf{X}^\top|\mathbf{S}|\mathbf{Q}_\mathbf{X} = \mathbf{X}^\top\mathbf{X}$. Hence, $\min_i|\mathbf{S}_{ii}|\|\mathbf{Q}_\mathbf{X}\|^2 = \|\mathbf{Q}_\mathbf{X}^\top(\min_i|\mathbf{S}_{ii}|)\mathbf{Q}_\mathbf{X}\| \leq \|\mathbf{X}^\top\mathbf{X}\|$, from which the claim follows.

Apart from converging slowly, the limiting $\mathbf{Q}_\mathbf{X}$ depends on F. Thus, a typical situation where the presence of indefinite orthogonal transformation *does* matter is when comparing two graphs with latent position distributions $F_1 \neq F_2$. In the Appendix, we provide two, two-graph examples, following standard and degree-corrected stochastic block models respectively, where the block matrices are equal, but the community proportions or degree distributions differ.

3.3 | Results for Laplacian spectral embedding

Analogous results are now given for the case of Laplacian spectral embedding. Here, the estimand is defined as

$$\frac{X_i}{\sqrt{\sum_j X_i^\top \mathbf{I}_{p,q} X_j}},$$

a latent position normalised according to its expected degree. As before, the estimate $\breve{X}_i$ will only resemble its estimand after indefinite orthogonal transformation by a matrix $\breve{\mathbf{Q}} \in \mathbb{O}(p, q)$, constructed in an analogous fashion to $\mathbf{Q}$. To avoid more definitions and notation, we will forego defining a 'spectral' estimand, as we did with adjacency spectral embedding.

Theorem 3 (Uniform consistency of Laplacian spectral embedding). *Assume the same sparsity conditions as Theorem 1. Then,*

$$\max_{i \in \{1,\dots,n\}} \left\| \breve{\mathbf{Q}}\breve{X}_i - \frac{X_i}{\sqrt{\sum_j X_i^\top \mathbf{I}_{p,q} X_j}} \right\| = O_{\mathbb{P}}\left(\frac{\log^c n}{n\rho_n^{1/2}} \right).$$

Theorem 4 (Laplacian spectral embedding central limit theorem). *Assume the same sparsity conditions as Theorem 1. Conditional on X_i, for $i = 1, \dots, m$, the random vectors*

$$n\rho_n^{1/2} \left(\breve{\mathbf{Q}}\breve{X}_i - \frac{X_i}{\sqrt{\sum_j X_i^\top \mathbf{I}_{p,q} X_j}} \right),$$

converge in distribution, as $n \to \infty$, to independent zero-mean Gaussian random vectors with covariance matrix $\breve{\boldsymbol{\Sigma}}(\xi_i)$ respectively, where

$$\breve{\boldsymbol{\Sigma}}(x) = \begin{cases} \mathbf{I}_{p,q}\breve{\boldsymbol{\Delta}}^{-1}\mathbb{E}\left\{ \left(\frac{x^\top \mathbf{I}_{p,q}\xi(1 - x^\top \mathbf{I}_{p,q}\xi)}{x^\top \mathbf{I}_{p,q}\mu} \right) \left(\frac{\xi}{\mu^\top \mathbf{I}_{p,q}\xi} - \frac{\breve{\boldsymbol{\Delta}}\mathbf{I}_{p,q}x}{2\mu^\top \mathbf{I}_{p,q}x} \right) \left(\frac{\xi}{\mu^\top \mathbf{I}_{p,q}\xi} - \frac{\breve{\boldsymbol{\Delta}}\mathbf{I}_{p,q}x}{2\mu^\top \mathbf{I}_{p,q}x} \right)^\top \right\} \breve{\boldsymbol{\Delta}}^{-1}\mathbf{I}_{p,q}, & \text{if } \rho_n = 1, \\ \mathbf{I}_{p,q}\breve{\boldsymbol{\Delta}}^{-1}\mathbb{E}\left\{ \left(\frac{x^\top \mathbf{I}_{p,q}\xi}{x^\top \mathbf{I}_{p,q}\mu} \right) \left(\frac{\xi}{\mu^\top \mathbf{I}_{p,q}\xi} - \frac{\breve{\boldsymbol{\Delta}}\mathbf{I}_{p,q}x}{2\mu^\top \mathbf{I}_{p,q}x} \right) \left(\frac{\xi}{\mu^\top \mathbf{I}_{p,q}\xi} - \frac{\breve{\boldsymbol{\Delta}}\mathbf{I}_{p,q}x}{2\mu^\top \mathbf{I}_{p,q}x} \right)^\top \right\} \breve{\boldsymbol{\Delta}}^{-1}\mathbf{I}_{p,q} & \text{if } \rho_n \to 0, \end{cases}$$

and $\mu = \mathbb{E}(\xi)$, $\breve{\boldsymbol{\Delta}} = \mathbb{E}\left(\frac{\xi\xi^\top}{\mu^\top \mathbf{I}_{p,q}\xi} \right)$.

4 | IMPLICATIONS FOR STOCHASTIC BLOCK MODEL ESTIMATION

The asymptotic results of Section 3 suggest the use of the following high-level algorithms (Athreya et al., 2016; Rubin-Delanchy et al., 2017; Tang & Priebe, 2018) for standard and mixed membership stochastic block model estimation, previously only supported under non-negative-definite assumptions on the block matrix $\mathbf{B}$.

Algorithm 1 Fitting a stochastic block model (spectral clustering)

```
   input adjacency matrix $\mathbf{A}$, dimension $d$, number of communities $K \geq d$
1: compute spectral embedding $\hat{X}_1, \dots, \hat{X}_n$ of the graph into $\mathbb{R}^d$
2: fit a Gaussian mixture model (varying volume, shape, and orientation) with $K$ components
   return cluster centres $\hat{\mathrm{v}}_1, \dots, \hat{\mathrm{v}}_K$ and community memberships $\hat{Z}_1, \dots, \hat{Z}_n$
```

Where this algorithm differs most significantly from Rohe et al. (2011) is in the use of a Gaussian mixture model over K-means clustering. In Section 4.1, we show why Theorems 2 and 4 would recommend this modification, with a pedagogical example.

To accomplish step 2 we have been employing the mclust algorithm (Fraley & Raftery, 1999), which has a user-friendly R package. In step 1, either adjacency or Laplacian spectral embedding can be used (see Definition 1). If the latter, the resulting node memberships can be interpreted as alternative estimates of $Z_1, \dots, Z_n$ but the output cluster centres should be treated as estimating degree-normalised versions of $\mathrm{v}_1, \dots, \mathrm{v}_K$ (see Section 3.3).

In real data, neither d nor K would typically be known, and yet they are required input for Algorithm 1. Estimating the rank (d) of a matrix ($\mathbf{P}$) observed ($\mathbf{A}$) under noise, and the number (K) of components in a Gaussian mixture model, are two problems with no undisputed, 'best', solution. We defer to the discussion of both by Priebe et al. (2019), which ultimately suggests the method of Zhu and Ghodsi (2006) for finding the elbow in the scree plot of $\mathbf{A}$ to select d (implemented in the R package igraph) and selecting K according to the Bayesian information criterion (implemented in the R package mclust). Alternatively, it is common to assume that $\mathbf{B}$ has full rank, as in Rohe et al. (2011), so that $d = K$, in which case the elbow method can be used, once, for both.

Algorithm 2 Fitting a mixed membership stochastic block model

```
   input adjacency matrix $\mathbf{A}$, dimension $d$, number of communities $K = d$
1: compute adjacency spectral embedding $\hat{X}_1, \dots, \hat{X}_n$ of the graph into $\mathbb{R}^d$
2: project the data onto the $(d-1)$-dimensional principal hyperplane, to obtain points $\hat{Y}_1, \dots, \hat{Y}_n$,
   and fit the minimum volume enclosing $K-1$-simplex, with vertices $\hat{\mathrm{v}}_1, \dots, \hat{\mathrm{v}}_K$
3: obtain barycentric coordinates $\hat{Y}_i = \sum_{k=1}^{K} \hat{\pi}_{il} \hat{\mathrm{v}}_k$, for $i = 1, \dots, n$
   return vertices $\hat{\mathrm{v}}_1, \dots, \hat{\mathrm{v}}_K$ of the simplex, and estimated community membership probability
   vectors $\hat{\pi}_1, \dots, \hat{\pi}_n$
```

This algorithm is unchanged from Rubin-Delanchy et al. (2017), but to prove it is consistent when $\mathbf{B}$ has negative eigenvalues requires Theorem 1 and Lemma 1. We provide a pedagogical example in Section 4.2.

To fit the minimum volume enclosing simplex in step 2, we have been using the algorithm by Lin et al. (2016) and are grateful to the authors for providing code. Algorithm 2 can be extended to the case $K \geq d$ by fitting a minimum volume enclosing convex K-polytope, but for identifiability one must then assume $\hat{\mathrm{v}}_1, \dots, \hat{\mathrm{v}}_K$ are in convex position.

4.1 | A two-community stochastic block model

In this section, we show why the central limit theorems (Theorems 2 and 4) would recommend fitting a Gaussian mixture model, rather than using K-means, for spectral clustering. To illustrate ideas, we consider a two-community stochastic block model under which every node is independently assigned to the first or second community, with respective probabilities 0.2 and 0.8, and the block matrix is

$$\mathbf{B}^{(1)} = \begin{bmatrix} 0.02 & 0.03 \\ 0.03 & 0.01 \end{bmatrix},$$

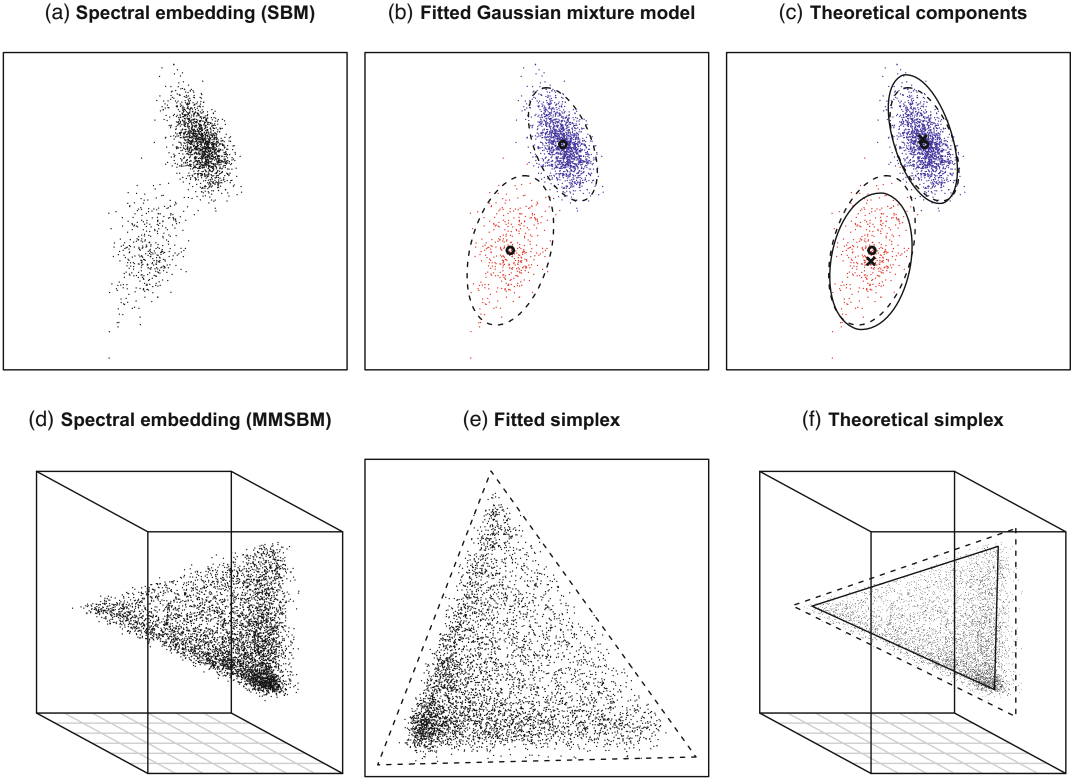

FIGURE 5 Spectral embedding and analysis of simulated graphs from the standard (SBM—top) and mixed membership (MMSBM—bottom) stochastic block models. (a) Adjacency spectral embedding into $\mathbb{R}^2$ of a simulated graph with two communities ($n = 2000$ nodes); (b) two-component Gaussian mixture model fit using the R package mclust, coloured by estimated component assignment, with estimated component centres as circles and 95% Gaussian contours as dashed ellipses; (c) two-component Gaussian mixture model predicted by the asymptotic theory, with component centres shown as crosses and 95% Gaussian contours shown as solid ellipses, overlaid onto the empirical centres (circles) and contours (dashed ellipses) shown in (b); (d) adjacency spectral embedding into $\mathbb{R}^3$ of a simulated graph with three communities ($n = 5000$ nodes); (e) minimum volume enclosing simplex (dashed triangle) enclosing the two principal components (points); (f) theoretical simplex supporting the latent positions (solid triangle) for comparison with the minimum volume enclosing simplex (dashed triangle). Detailed discussion in Section 4 [Colour figure can be viewed at wileyonlinelibrary.com]

which has one positive and one negative eigenvalue. The two-dimensional adjacency spectral embedding of a simulated graph on $n = 2000$ nodes is shown in Figure 5a. As we will see, Theorem 2 provides a formal sense in which this embedding resembles data from a two-component Gaussian mixture distribution. Figure 5b shows how this model fits the embedding, based on the approximate maximum likelihood parameters found by the mclust algorithm (Fraley & Raftery, 1999). The estimated component assignment of each point is indicated by colouring, the empirical component centres are shown as small circles and corresponding empirical 95% Gaussian contours in dashed lines.

Following the construction of Section 2.1.1, the graph is a GRDPG with signature (1, 1), and its latent positions $X_i \in \mathbb{R}^2$ are i.i.d. from a distribution F which places all its mass on two distinct points, v_1 and v_2, with respective probabilities 0.2 and 0.8. The implication of Theorem 2 for this example is that, conditional on Z_i, the transformed embedding $\mathbf{Q}\hat{X}_i$ is approximately distributed as an independent Gaussian vector with centre v_{Z_i} and covariance $n^{-1/2}\boldsymbol{\Sigma}(\mathrm{v}_{Z_i})$, where $\mathbf{Q} \in \mathbb{O}(1, 1)$; or, together, the vectors $\mathbf{Q}\hat{X}_1, \dots, \mathbf{Q}\hat{X}_n$ approximately follow a two-component Gaussian mixture model.

As a result, the points $\hat{X}_1, \dots, \hat{X}_n$ resemble data from a two-component Gaussian mixture model which have been put through a random, data-dependent, linear transformation $\mathbf{Q}^{-1}$. Given $\mathbf{A}$, the cluster assignments, v_1 and v_2, we can compute $\mathbf{Q}$ (see Section 3.1). For the graph that we simulated,

$$\mathbf{Q} \approx \begin{bmatrix} 1.05 & -0.32 \\ 0.32 & -1.05 \end{bmatrix}.$$

Figure 5c shows the embedding, the transformed centres $\mathbf{Q}^{-1}\mathrm{v}_1, \mathbf{Q}^{-1}\mathrm{v}_2$ (crosses), and correspondingly transformed 95% Gaussian contours (solid ellipses) predicted by Theorem 2 for comparison with the empirical versions (circles and dashed ellipses) obtained from the Gaussian mixture model fit.

In practice, we typically observe only $\mathbf{A}$ and do not have access to $\mathbf{Q}$, and indeed Algorithm 1 is suggesting we fit a Gaussian mixture model to $\hat{X}_1, \dots, \hat{X}_n$, not $\mathbf{Q}\hat{X}_1, \dots, \mathbf{Q}\hat{X}_n$. Remarkably, by doing the former, we are effectively accomplishing the latter.

This is because, under a Gaussian mixture model, the value of the likelihood is unchanged if, while the component weights are held fixed, the data, component means and covariances are respectively transformed as $X \to \mathbf{M}X$, $\mu \to \mathbf{M}\mu$, $\Gamma \to \mathbf{M}\Gamma\mathbf{M}^\top$, where $\mathbf{M}$ is any indefinite orthogonal matrix, for the simple reason that $|\det(\mathbf{M})| = 1$. For the maximum likelihood parameters obtained from $\hat{X}_1, \dots, \hat{X}_n$, the maximum *a posteriori* probability assignment of the data indices to mixture components, $\hat{Z}_1, \dots, \hat{Z}_n$, is identical to that which would have been obtained from $\mathbf{Q}\hat{X}_1, \dots, \mathbf{Q}\hat{X}_n$. The cluster centres which would have been obtained from $\mathbf{Q}\hat{X}_1, \dots, \mathbf{Q}\hat{X}_n$ are just $\mathbf{Q}\hat{\mathrm{v}}_1$ and $\mathbf{Q}\hat{\mathrm{v}}_2$, where $\hat{\mathrm{v}}_1, \hat{\mathrm{v}}_2$ are those actually obtained from $\hat{X}_1, \dots, \hat{X}_n$. The pairs provide identical estimates of $\mathbf{B}^{(1)}$ through $\hat{\mathbf{B}}^{(1)}_{kl} = (\mathbf{Q}\hat{\mathrm{v}}_k)^\top \mathbf{I}_{1,1}(\mathbf{Q}\hat{\mathrm{v}}_l) = \hat{\mathrm{v}}_k^\top \mathbf{I}_{1,1}\hat{\mathrm{v}}_l$ for $k, l \in \{1, 2\}$. In practice regularisation parameters in the clustering method may yield results that are not invariant to indefinite transformation, but such effects should be small for large n, especially taken alongside the additional result, in Lemma 1, that the spectral norm of $\mathbf{Q}$ is almost surely bounded.

As can be seen in Figure 5c, the clusters are elliptical rather than spherical, in theory and in practice. For this reason, a clustering algorithm which favours spherical solutions might produce inaccurate results. This is one issue with K-means clustering, often said to be implicitly fitting a Gaussian mixture model with equal-volume spherical components (Fraley & Raftery, 2002), and indeed numerical studies (Athreya et al., 2016; Tang & Priebe, 2018) show it has higher misclassification rate on the task of community separation. Another issue, perhaps more a limitation of our own framework, is that Euclidean distance is not identifiable under the GRDPG, and so the empirical distances $\|\hat{X}_i - \hat{X}_j\|$ are not easily understood through our approach. While our results say enough to prove K-means clustering is consistent (a fact already established by other methods), the algorithm is not invariant: the clusterings obtained from $\hat{X}_1, \dots, \hat{X}_n$ and from $\mathbf{Q}\hat{X}_1, \dots, \mathbf{Q}\hat{X}_n$ are different (although they agree asymptotically).

4.2 | A three-community mixed membership stochastic block model

In this section, we show how uniform consistency (Theorem 1) guarantees the consistency of simplex fitting for mixed membership stochastic block model estimation. To illustrate ideas, we consider a three-community mixed membership stochastic block model under which every node is first independently assigned a three-dimensional probability vector $\pi_i \sim \text{Dirichlet}(1, 0.5, 0.5)$, for $i = 1, \dots, n$, and the block matrix is

$$\mathbf{B}^{(2)} = \begin{pmatrix} 0.6 & 0.9 & 0.9 \\ 0.9 & 0.6 & 0.9 \\ 0.9 & 0.9 & 0.3 \end{pmatrix},$$

which has one positive and two negative eigenvalues. The three-dimensional adjacency spectral embedding of a simulated graph on $n = 5000$ nodes is shown Figure 5d; this point cloud resembles a 'noisy simplex'. Figure 5e shows the minimum volume 2-simplex (dashed lines) enclosing the two principal components of the points. As we will see, Theorem 1 provides a formal sense in which the point cloud becomes denser *and* sharper as n grows, so that this simplex converges.

Following the construction of Section 2.1.2, the graph is a GRDPG with signature (1,2) and its latent positions $X_i \in \mathbb{R}^3$ are i.i.d. from a distribution F supported on the simplex with vertices $\mathrm{v}_1, \mathrm{v}_2, \mathrm{v}_3 \in \mathbb{R}^3$.

Since $\log^c n/n^{1/2} \to 0$, Theorem 1 states that the *maximum error* between any $\mathbf{Q}\hat{X}_i$ and X_i vanishes, across $i \in \{1, \dots, n\}$, and this has wide-reaching implications for geometric analysis since, as sets, $\mathbf{Q}\hat{X}_1, \dots, \mathbf{Q}\hat{X}_n$ and $X_1, \dots, X_n$ are asymptotically equal in Hausdorff distance. In particular, for our example, where $\mathbf{Q} \in \mathbb{O}(1,2)$, the point set $\mathbf{Q}\hat{X}_1, \dots, \mathbf{Q}\hat{X}_n$, like $X_1, \dots, X_n$, converges in Hausdorff distance to the simplex with vertices $\mathrm{v}_1, \mathrm{v}_2, \mathrm{v}_3$. This simplex would be consistently estimated by the minimum volume simplex enclosing $\mathbf{Q}\hat{X}_1, \dots, \mathbf{Q}\hat{X}_n$, projected onto their two-dimensional principal hyperplane, by the argument presented in Rubin-Delanchy et al. (2017).

We must just verify that the *actual* procedure, that is, applied to $\hat{X}_1, \dots, \hat{X}_n$, is consistent despite indefinite orthogonal transformation. The bounded spectral norm of $\mathbf{Q}^{-1}$ (Lemma 1 and discussion) guarantees that

$$\max_{i \in \{1, \dots, n\}} \|\hat{X}_i - \mathbf{Q}^{-1} X_i\| \le \left\|\mathbf{Q}^{-1}\right\| \max_{i \in \{1, \dots, n\}} \|\mathbf{Q}\hat{X}_i - X_i\| = O_{\mathbb{P}}\left(\frac{\log^c n}{n^{1/2}}\right),$$

through which one can replace X_i by $\mathbf{Q}^{-1}X_i$ in the argument of Rubin-Delanchy et al. (2017). The minimum volume simplex enclosing $\hat{Y}_1, \dots, \hat{Y}_n$ (the points $\hat{X}_1, \dots, \hat{X}_n$ projected onto their two-dimensional principal hyperplane, see Algorithm 2), is then found to converge to the simplex with vertices $\mathbf{Q}^{-1}\mathrm{v}_1, \mathbf{Q}^{-1}\mathrm{v}_2, \mathbf{Q}^{-1}\mathrm{v}_3$. This simplex is shown in Figure 5f (solid lines).

As with the stochastic block model, the presence of indefinite orthogonal transformation in the estimated vertices $\hat{\mathrm{v}}_1, \hat{\mathrm{v}}_2, \hat{\mathrm{v}}_3$ is immaterial for estimating $\mathbf{B}$ through $\hat{\mathbf{B}}^{(2)}_{kl} = \hat{\mathrm{v}}_k^\top \mathbf{I}_{1,2} \hat{\mathrm{v}}_l$ for $k, l \in \{1, 2, 3\}$. Moreover, the community membership probability vectors $\hat{\pi}_1, \dots, \hat{\pi}_n$, obtained as the barycentric coordinates of $\hat{Y}_1, \dots, \hat{Y}_n$ with respect to $\hat{\mathrm{v}}_1, \hat{\mathrm{v}}_2, \hat{\mathrm{v}}_3$, are the same as those which would be obtained from $\mathbf{Q}\hat{Y}_1, \dots, \mathbf{Q}\hat{Y}_n$ with respect to $\mathbf{Q}\hat{\mathrm{v}}_1, \mathbf{Q}\hat{\mathrm{v}}_2, \mathbf{Q}\hat{\mathrm{v}}_3$.

5 | REAL DATA

5.1 | A collection of real-world graphs

Allowing a principled treatment of negative eigenvalues increases the scope of application of spectral embedding. To gain an impression of the significance of our extension, we conduct a survey of graphs from a variety of application domains. Graphs with about 5000 nodes were chosen from each of the domain categories of a comprehensive online network repository (https://networkrepository.com), selecting the largest if all graphs in a category were smaller, and rejecting the category if all graphs were much larger.

For each of the resulting 24 graphs, an estimated embedding dimension $\hat{d}$ was obtained using the elbow method of Zhu and Ghodsi (2006), and the estimates $\hat{p}$ (respectively, $\hat{q}$) correspond to the number of positive (respectively, negative) eigenvalues among the largest $\hat{d}$ in magnitude. Results are shown in Table 1, and it happens that precisely half of the graphs have $\hat{q} > 0$. Moreover, the smallest negative eigenvalue often ranks among the largest in magnitude.

TABLE 1 A collection of real-world graphs. $\hat{d}$: the estimated dimension; $\hat{p}$ (respectively, $\hat{q}$): the number of positive (respectively, negative) eigenvalues of the adjacency matrix among the first $\hat{d}$. The estimate $\hat{q}$ is non-zero for half of these graphs

Graph category	Nodes	Edges	$\hat{d}$	$\hat{p}$	$\hat{q}$
Animal social	1686	5324	9	9	0
Benchmark (BHOSLIB)	4000	7,425,226	2	1	1
Benchmark (DIMACS 10)	4096	12, 264	6	6	0
Benchmark (DIMACS)	4000	4,000,268	3	2	1
Biological	4413	108,818	23	11	12
Brain	1781	33,641	6	6	0
Cheminformatics	125	282	12	12	0
Collaboration	4158	13,422	1	1	0
Communication	1899	61,734	54	27	27
Ecology	128	2106	47	47	0
Economic	4008	8188	4	2	2
Email	1133	5451	28	25	3
Infrastructure	4941	6594	2	1	1
Interaction	1266	6451	1	1	0
Molecular	5110	10,532	16	16	0
Power	5300	13,571	22	22	0
Proximity	410	2765	17	17	0
Retweet	5248	6394	25	20	5
Road	2642	3303	11	9	2
Router	2113	6632	13	13	0
Social (advogato)	6551	51,332	46	46	0
Social (Facebook)	5372	279,191	6	5	1
Structural mechanics	5489	143,300	12	6	6
Web	4767	37,375	12	10	2

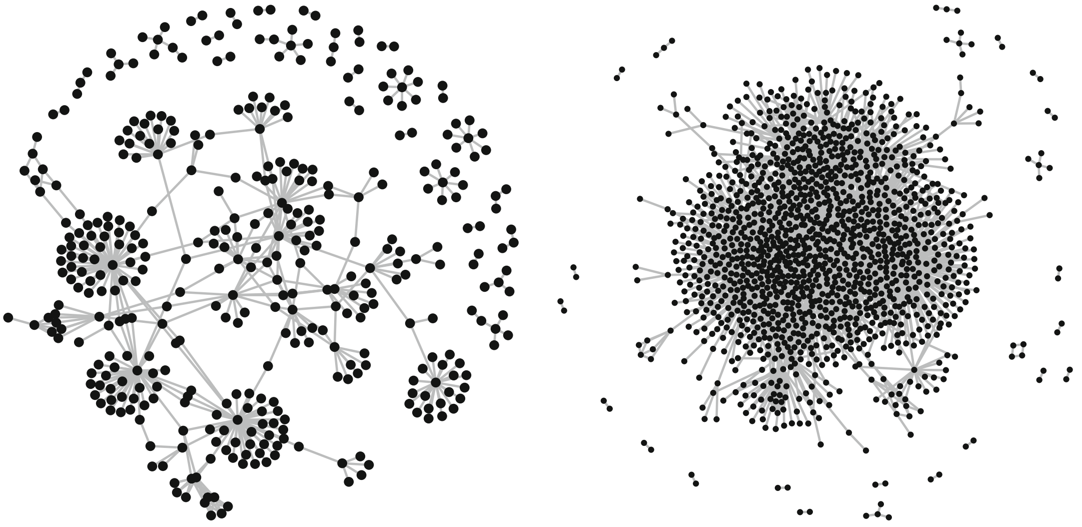

FIGURE 6 Los Alamos National Laboratory computer network. Graphs of the connections made between different computers (IP addresses) over the first minute (left) and first 5 min (right) of the 'network flow events' dataset (Kent, 2016a). Neither graph contains a single triangle, a motif which would be expected in abundance under homophily ('a friend of my friend is my friend'), suggesting the need to relax this modelling assumption

5.2 | Detailed example: Link prediction on a computer network

Cyber-security applications often involve data with a network structure, for example, data relating to computer network traffic (Neil et al., 2013a), the underground economy (Li & Chen, 2014), and the internet-of-things (Hewlett Packard Enterprise research study, 2015). In the first example, a concrete reason to seek to develop an accurate network model is to help identify intrusions on the basis of anomalous links (Heard & Rubin-Delanchy, 2016; Neil et al., 2013b).

Figure 6 shows, side by side, graphs of the communications made between computers on the Los Alamos National Laboratory network (Kent, 2016a, 2016b) over a single minute on the left, and 5 min on the right. The graphs were extracted from the 'network flow events' dataset, by mapping each IP address to a node, and recording an edge if the corresponding two IP addresses are observed to communicate at least once over the specified period. The extracted graphs are available as supplementary material.

Neither graph contains a single triangle, that is, three nodes all connecting to each other. This is a symptom of a broader property, known as heterophily or disassortivity (Khor, 2010), that similar nodes are relatively *unlikely* to connect. In computer networks, such behaviour might be expected for a number of reasons, including the common server/client networking model and the physical location of routers (where collection happens) (Rubin-Delanchy et al., 2016). The random dot product graph is unsuited to modelling heterophilic connectivity patterns. For example, any two-community stochastic block model with lower on- than off-diagonal elements is out of scope. The eigenvalues of the adjacency matrix of the 5-min graph are plotted in Figure 7, showing an abundance of negative eigenvalues which, again, cannot be modelled by the random dot product graph (apart from as noise).

The modelling improvement offered by the GRDPG over the random dot product graph is now demonstrated empirically, through out-of-sample link prediction. For the observed 5-min graph,

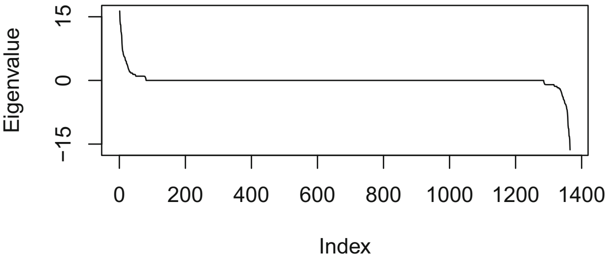

FIGURE 7 Eigenvalues of the adjacency matrix of the 5-min connection graph of computers on the Los Alamos National Laboratory network, showing roughly equal contribution, in magnitude, from the positive and negative eigenvalues

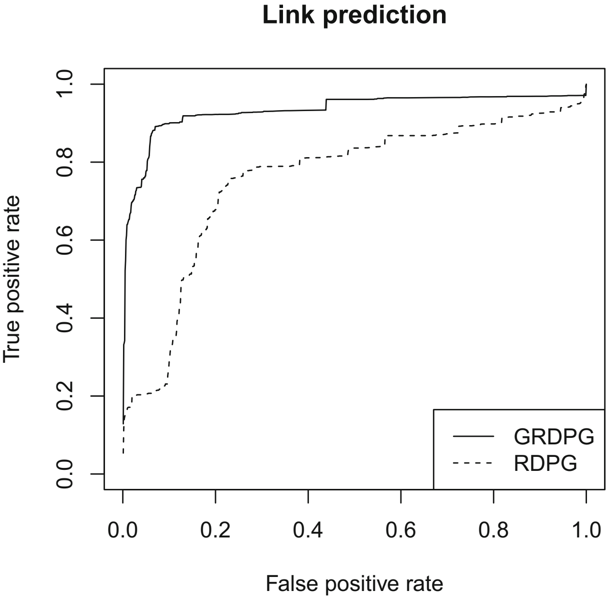

FIGURE 8 Receiver Operating Characteristic curves comparing the random dot product graph (RDPG) and GRDPG at the task of link prediction on the Los Alamos National Laboratory computer network. The nodes are embedded based on their graph of connections over the first 5-min window, using either the largest positive (RDPG) or largest-magnitude (GRDPG) eigenvalues. The performance of these embeddings is then evaluated for predicting new edges over the next 5-min window, using the matrix of pairwise inner products (RDPG) or pairwise indefinite inner products (GRDPG) as estimated edge probabilities. The ROC curves for each embedding is presented, showing superior performance for the GRDPG

we estimate the GRDPG latent positions via adjacency spectral embedding, as in Definition 1, and the random dot product graph latent positions using an analogous procedure that retains instead only the largest eigenvalues and corresponding eigenvectors. In both cases we choose $\hat{d} = 6$ as the embedding dimension, using the elbow method of Zhu and Ghodsi (2006), following Priebe et al. (2019), and $\hat{p}$ (respectively, $\hat{q}$) as the number of positive (respectively, negative) eigenvalues among the largest $\hat{d}$ in magnitude.

To compare the models, we then attempt to predict which *new* edges will occur in the next 5-min window, a task known as link prediction, disregarding those involving new nodes. Figure 8 shows the receiver operating characteristic (ROC) curves for each model, treating the prediction task as a classification problem where the presence or absence of an edge is encoded as an instance of the positive or negative class, respectively, and predicted by thresholding the inner product or indefinite inner product of the relevant pair of estimated latent positions. By presenting estimated classification performance (true positive vs. false positive rate) at all possible thresholds (which give different points along the curve), the ROC allows a direct comparison that is independent of

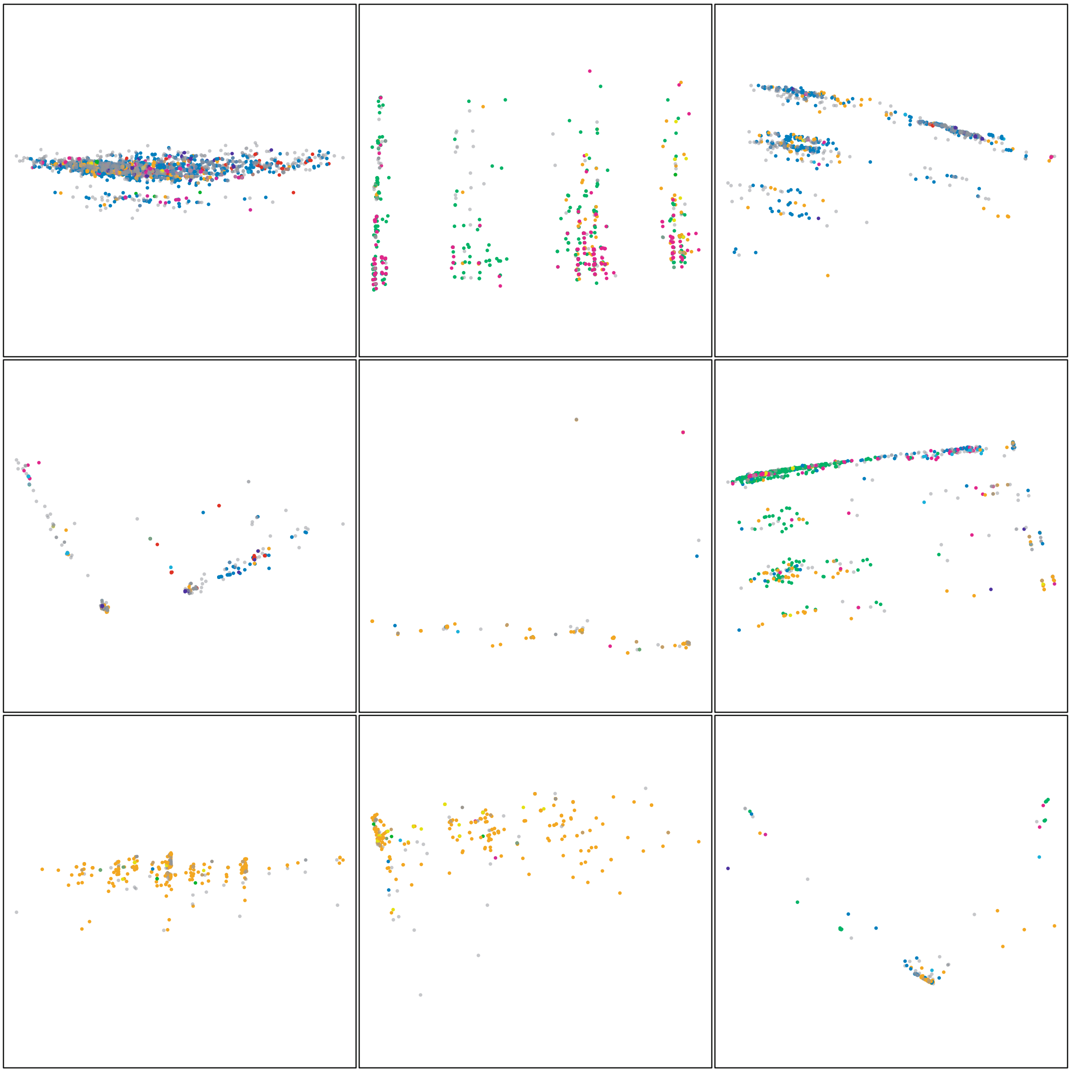

FIGURE 9 Visualisation of the full graph of connections between computers on the Los Alamos National Laboratory network, using adjacency spectral embedding into $\mathbb{R}^6$, followed by fitting a Gaussian mixture model with nine components. Each of nine clusters obtained is shown in a panel (two principal components) and the colour of each point represents the corresponding node's most commonly employed port (loosely representing the connection purpose, e.g. web, email), showing association with the structure observed in the embedding. The structure is richer than the stochastic block model or extensions would predict [Colour figure can be viewed at wileyonlinelibrary.com]

the potentially different ranges and scales of the two inner products. For this prediction problem, the GRDPG model is far superior.

What does the GRDPG model add over the stochastic block model and extensions? For large real-world networks, the latter models are often too simplistic, whereas the GRDPG model and the statistical investigation thereof, as presented in this paper, provide a more broadly applicable, principled starting point for analyses when low-dimensional latent structure is supposed. To illustrate this, we construct the full graph of connections between computers on the Los Alamos National Laboratory network, comprising roughly 12,000 nodes and one hundred thousand edges. As before, the nodes are spectrally embedded into $\mathbb{R}^6$, but these are now visualised in two ways. First we fit a Gaussian mixture model to the data, as would be consistent with a stochastic block model assumption. Following Priebe et al. (2019), we pick $\hat{K} = 9$ components using BIC. Each of the nine panels in Figure 9 shows the two principal components of one of the inferred clusters in a faithful aspect ratio. Because every communication has an associated port number indicating the type of service being used, for example port 80 corresponds to web activity and port 25 to email, this information can be used to colour the nodes according to their most commonly employed port. The embedding, obtained using only connectivity data, is clearly highly associated with port activity, so that the geometry that can be distinguished appears to be somehow predictive of nodes' behaviour. At the same time the clusters, for the most part, do not appear

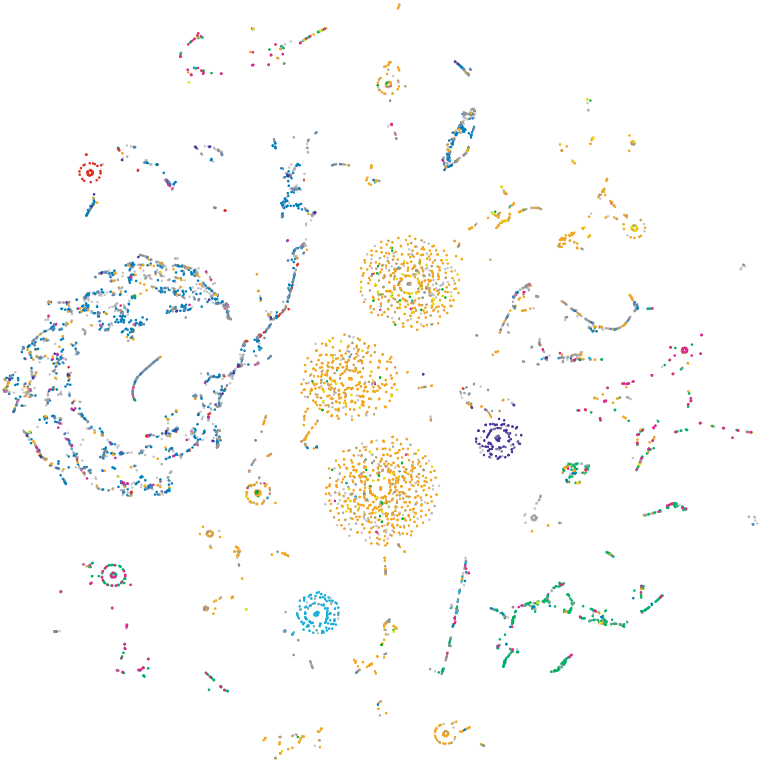

FIGURE 10 Alternative visualisation the full graph of connections between computers on the Los Alamos National Laboratory network, using adjacency spectral embedding into $\mathbb{R}^6$ followed by t-distributed stochastic neighbour embedding. The colour of each point represents the corresponding node's most commonly employed port (loosely representing the connection purpose, e.g. web, email), showing association with the structure observed in the embedding [Colour figure can be viewed at wileyonlinelibrary.com]

to follow Gaussian distributions. Similarly, the data do not have the appearance of a simplex (as predicted under mixed membership) or a pattern of rays (predicted under degree correction). A different view of the data is obtained using *t*-distributed stochastic neighbour embedding (Maaten & Hinton, 2008) in Figure 10, again showing high association with port activity. Taken together, these views of the data reveal complex structure in low-dimensional pseudo-Euclidean latent space, for which the GRDPG provides a preferable starting point for statistical analysis to the stochastic block model.

6 | CONCLUSION

This paper presents the *generalised random dot product graph*, a latent position model which includes the stochastic block model, its extensions, and the random dot product graph as special cases. The key feature that is added by the generalisation is the possibility of modelling non-homophilic connectivity behaviour, for example, where 'opposites attract'.

This model provides an appropriate statistical framework for interpreting spectral embedding. This is substantiated in several theoretical results that together show that the vector representations of nodes obtained by spectral embedding provide uniformly consistent latent position estimates with asymptotically Gaussian error. A by-product of this theory is to add insight and methodological improvements to the estimation of community structure in networks, and practical applications are demonstrated in a cyber-security example.

ORCID

Patrick Rubin-Delanchy https://orcid.org/0000-0003-0577-0795

REFERENCES

Abbe, E. (2017) Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(1), 6446–6531.

Agterberg, J., Tang, M. & Priebe, C.E. (2020) On two distinct sources of nonidentifiability in latent position random graph models. *arXiv preprint arXiv:2003.14250*.

Airoldi, E.M., Blei, D.M., Fienberg, S.E. & Xing, E.P. (2008) Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep), 1981–2014.

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

Athreya, A., Priebe, C.E., Tang, M., Lyzinski, V., Marchette, D.J. & Sussman, D.L. (2016) A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78(1), 1–18.

Athreya, A., Fishkind, D.E., Tang, M., Priebe, C.E., Park, Y., Vogelstein, J.T. et al. (2017) Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(1), 8393–8484.

Athreya, A., Tang, M., Park, Y. & Priebe, C.E. (2021) On estimation and inference in latent structure random graphs. *Statistical Science*, 36(1), 68–88.

Bhatia, R. (1997) *Matrix analysis*. Berlin: Springer.

Borgatti, S.P. & Everett, M.G. (2000) Models of core/periphery structures. *Social Networks*, 21(4), 375–395.

Cape, J., Tang, M. & Priebe, C.E. (2019a) Signal-plus-noise matrix models: eigenvector deviations and fluctuations. *Biometrika*, 106(1), 243–250.

Cape, J., Tang, M. & Priebe, C.E. (2019b) The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47(5), 2405–2439.

Donath, W.E. & Hoffman, A.J. (1973) Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5), 420–425.

Erdős, L., Knowles, A., Yau, H.-T. & Yin, J. (2013) Spectral statistics of Erdős-Rényi' graphs I: local semicircle law. *The Annals of Probability*, 41, 2279–2375.

Evans, C., Friedman, J., Karakus, E. & Pandey, J. (2014) Potterverse. Available from: https://github.com/efekarakus/potter-network/

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

Fraley, C. & Raftery, A.E. (1999) MCLUST: software for model-based cluster analysis. *Journal of Classification*, 16(2), 297–306.

Fraley, C. & Raftery, A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 611–631.

Gallier, J.H. (2000) *Curves and surfaces in geometric modeling: theory and algorithms*. Burlington, MA: Morgan Kaufmann.

Heard, N.A. & Rubin-Delanchy, P. (2016) Network-wide anomaly detection via the Dirichlet process. In: *Proceedings of IEEE workshop on big data analytics for cyber-security computing*.

Hewlett Packard Enterprise research study. (2015) Internet of things: research study. Available from: http://h20195.www2.hpe.com/V4/getpdf.aspx/4aa5-4759enw

Hoff, P. (2008) Modeling homophily and stochastic equivalence in symmetric relational data. In: Platt, J., Koller, D., Singer, Y. & Roweis, S. (Eds.) *Advances in neural information processing systems*, volume 20, Red Hook, NY: Curran Associates, Inc.

Hoff, P.D., Raftery, A.E. & Handcock, M.S. (2002) Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460), 1090–1098.

Holland, P.W., Laskey, K.B. & Leinhardt, S. (1983) Stochastic blockmodels: first steps. *Social Networks*, 5(2), 109–137.

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

Karrer, B. & Newman, M.E. (2011) Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1), 16107.

Kent, A.D. (2016a) Cybersecurity data sources for dynamic network research. In: Adams, N. & Heard, N. (Eds.) *Dynamic networks and cyber-security*. London: World Scientific, pp. 37–65.

Kent, A.D. (2016b) Cybersecurity data sources for dynamic network research. Available from: https://csr.lanl.gov/data/cyber1/ [Accessed 25th January 2022].

Khor, S. (2010) Concurrency and network disassortativity. *Artificial Life*, 16(3), 225–232.

Labatut, V. & Bost, X. (2019) Extraction and analysis of fictional character networks: a survey. *ACM Computing Surveys (CSUR)*, 52(5), 1–40.

Lei, J. (2021) Network representation using graph root distributions. *The Annals of Statistics*, 49(2), 745–768.

Lei, J. & Rinaldo, A. (2015) Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1), 215–237.

Li, W. & Chen, H. (2014) Identifying top sellers in underground economy using deep learningbased sentiment analysis. In: *2014 IEEE joint intelligence and security informatics conference (JISIC), 2014 IEEE Joint*, IEEE, pp. 64–67.

Lin, C.-H., Chi, C.-Y., Wang, Y.-H. & Chan, T.-H. (2016) A fast hyperplane-based minimumvolume enclosing simplex algorithm for blind hyperspectral unmixing. *IEEE Transactions on Signal Processing*, 64(8), 1946–1961.

Lloyd, S. (1982) Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137.

Lovász, L. (2012) *Large networks and graph limits. American Mathematical Society Colloquium Publications*, volume 60, Providence, RI: American Mathematical Society.

Lu, L. & Peng, X. (2013) Spectra of edge-independent random graphs. *Electronic Journal of Combinatorics*, 20(4), 27–45.

Lyzinski, V., Sussman, D.L., Tang, M., Athreya, A. & Priebe, C.E. (2014) Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8(2), 2905–2922.

Lyzinski, V., Tang, M., Athreya, A., Park, Y. & Priebe, C.E. (2017) Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions on Network Science and Engineering*, 4(1), 13–26.

Maaten, L.V.D. & Hinton, G. (2008) Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov), 2579–2605.

Mao, X., Sarkar, P. & Chakrabarti, D. (2021) Estimating mixed memberships with sharp eigenvector deviations. *Journal of the American Statistical Association*, 116(536), 1928–1940.

Mara, A., Mashayekhi, Y., Lijffijt, J. & De Bie, T. (2020) CSNE: conditional signed network embedding. *arXiv preprint arXiv:2005.10701*.

Neil, J.C., Hash, C., Brugh, A., Fisk, M. & Storlie, C.B. (2013a) Scan statistics for the online detection of locally anomalous subgraphs. *Technometrics*, 55(4), 403–414.

Neil, J.C., Uphoff, B., Hash, C. & Storlie, C. (2013b) Towards improved detection of attackers in computer networks: new edges, fast updating, and host agents. In: *6th International symposium on resilient control systems (ISRCS)*, IEEE, pp. 218–224.

Newman, M. (2018) *Networks: an introduction*. Oxford: Oxford University Press.

Nickel, C. (2006) *Random dot product graphs: a model for social networks*. PhD thesis, Johns Hopkins University.

Passino, F.S., Heard, N.A. & Rubin-Delanchy, P. (2022) Spectral clustering on spherical coordinates under the degree-corrected stochastic blockmodel. *Technometrics*, 1–12.

Priebe, C.E., Park, Y., Vogelstein, J.T., Conroy, J.M., Lyzinski, V., Tang, M. et al. (2019) On a two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences*, 116(13), 5995–6000.

Qin, T. & Rohe, K. (2013) Regularized spectral clustering under the degree-corrected stochastic blockmodel. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K.Q. (Eds.) *Advances in neural information processing systems*, volume 26, Red Hook, NY: Curran Associates, Inc.

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

Rohe, K., Tao, J., Han, X. & Binkiewicz, N. (2018) A note on quickly sampling a sparse matrix with low rank expectation. *Journal of Machine Learning Research*, 19(1), 3040–3052.

Rubin-Delanchy, P. (2020) Manifold structure in graph embeddings. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F. & Lin, H. (Eds.) *Advances in neural information processing systems*, volume 33, Red Hook, NY: Curran Associates, Inc., pp. 11687–11699.

Rubin-Delanchy, P., Adams, N.M. & Heard, N.A. (2016) Disassortivity of computer networks. In: *Proceedings of IEEE workshop on big data analytics for cyber-security computing*.

Rubin-Delanchy, P., Priebe, C.E. & Tang, M. (2017) Consistency of adjacency spectral embedding for the mixed membership stochastic blockmodel. *arXiv preprint arXiv:1705.04518*.

Sarkar, P. & Bickel, P.J. (2015) Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics*, 43(3), 962–990.

Shi, J. & Malik, J. (2000) Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888–905.

Steinhaus, H. (1956) Sur la division des corp matériels en parties. *Bulletin L'Académie Polonaise des Sciences*, 1(804), 801.

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

Tang, M. & Priebe, C.E. (2018) Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *The Annals of Statistics*, 46(5), 2360–2415.

Tang, M., Cape, J. & Priebe, C.E. (2017) Asymptotically efficient estimators for stochastic blockmodels: the naive MLE, the rank-constrained MLE, and the spectral. *arXiv preprint* at http://arxiv.org/abs/1710.10936

Trosset, M.W., Gao, M., Tang, M. & Priebe, C.E. (2020) Learning 1-dimensional submanifolds for subsequent inference on random dot product graphs. *arXiv preprint arXiv:2004.07348*.

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

Whiteley, N., Gray, A. & Rubin-Delanchy, P. (2021) Matrix factorisation and the interpretation of geodesic distance. *arXiv preprint arXiv:2106.01260*.

Young, S.J. & Scheinerman, E.R. (2007) Random dot product graph models for social networks. In: *International workshop on algorithms and models for the web-graph*, Springer, pp. 138–149.

Yu, Y., Wang, T. & Samworth, R.J. (2015) A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102(2), 315–323.

Zhu, M. & Ghodsi, A. (2006) Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2), 918–930.

How to cite this article: Rubin-Delanchy, P., Cape, J., Tang, M. & Priebe, C.E. (2022) A statistical interpretation of spectral embedding: The generalised random dot product graph. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(4), 1446–1473. Available from: https://doi.org/10.1111/rssb.12509