The Annals of Statistics
2021, Vol. 49, No. 2, 745–768
https://doi.org/10.1214/20-AOS1976

NETWORK REPRESENTATION USING GRAPH ROOT DISTRIBUTIONS

BY JING LEI

Department of Statistics and Data Science, Carnegie Mellon University, jinglei@andrew.cmu.edu

Exchangeable random graphs serve as an important probabilistic framework for the statistical analysis of network data. In this work, we develop an alternative parameterization for a large class of exchangeable random graphs, where the nodes are independent random vectors in a linear space equipped with an indefinite inner product, and the edge probability between two nodes equals the inner product of the corresponding node vectors. Therefore, the distribution of exchangeable random graphs in this subclass can be represented by a node sampling distribution on this linear space, which we call the *graph root distribution*. We study existence and identifiability of such representations, the topological relationship between the graph root distribution and the exchangeable random graph sampling distribution and estimation of graph root distributions.

1. Introduction. In recent years, network analysis has been the focus of many theoretical and applied research efforts in the scientific community, due to the increasing popularity of relational data. Generally speaking, a network records the presence and absence of pairwise interactions among a group of individuals, and statistical network analysis aims at recovering properties of the underlying population of individuals from their pairwise interactions. There is a vast literature on network analysis, and we refer to [20, 30, 40] for more detailed review of this field from a statistical perspective.

Exchangeable random graphs [4, 24, 26] are an important class of probabilistic models for network data. The exchangeability requirement is quite natural: The individuals recorded in the network are somewhat like random sample points, and the data distribution remains unchanged under permutation of the nodes. Many popularly studied network models are special cases of exchangeable random graphs, including the stochastic block model [7, 23], the degree-corrected block model [27], the mixed-membership block model [2], the random dot-product graph model [5, 41, 46] and random geometric graphs [44].

A central piece of the theoretical foundation of exchangeable random graphs is the celebrated Aldous–Hoover theorem [4, 24, 26], which says that any exchangeable random graph of infinite size can be generated by first sampling independent node variables $(s_i : i \geq 1)$ uniformly on $[0, 1]$, and then connect each pair of nodes (i, j) independently with probability $W(s_i, s_j)$, for some symmetric $W : [0, 1]^2 \mapsto [0, 1]$. The representation theory also says that two functions W_1, W_2 lead to the same distribution of exchangeable random graphs if and only if there exist measure-preserving mappings h_1, h_2, both from $[0, 1]$ to $[0, 1]$, such that $W_1(h_1(s), h_1(t)) = W_2(h_2(s), h_2(t))$ almost everywhere over $(s, t) \in [0, 1]^2$. Other random graph models have been proposed, such as exchangeable random measures [10], edge exchangeability [16] and ideas using a Bayesian framework [43]. A notable difference is that these models can cover sparse networks while the Aldous–Hoover type exchangeable arrays must be dense. Nevertheless, when a finite sample is concerned, one can add a sparsity parameter to generate a sparse network using the Aldous–Hoover type exchangeable array. See Section 4.4 for further discussion on sparse networks.

Received March 2019; revised November 2019.
MSC2020 subject classifications. Primary 62E10; secondary 62G05, 62M15.
Key words and phrases. Network data, exchangeable random graph, Kreĭn space, spectral embedding.

The purpose of this work is to develop an alternative parametrization for a subclass of exchangeable random graphs and corresponding estimators. This new parameterization is based on a representation of nodes as independent random vectors in a separable Kreĭn space $\mathcal{K}$, and the edge probability is the inner product of the two node vectors. A Kreĭn space is isomorphic to a Hilbert space, but has an indefinite inner product. With such a Kreĭn space node embedding, we shift the information hidden in the function W to the probability measure on $\mathcal{K}$, which can exhibit salient structures in a transparent and geometric manner.

We highlight a few key contributions:

1. We provide a constructive proof for the correspondence between a subclass of exchangeable random graphs and probability distributions on the Kreĭn space $\mathcal{K}$. Our construction starts from viewing W as an integral operator and considering its spectral decomposition. The variable $s \sim \text{Unif}(0, 1)$ is treated as the input variable in an infinite-dimensional inverse transform sampling. The induced measure is thus invariant under measure-preserving transforms of s. The existence and identifiability of such a representation is established for a wide class of exchangeable random graphs. In our construction, it becomes apparent that the induced measure is closely related to the square root of the integral operator W, with appropriate treatment of negative eigenvalues. Thus we call this induced measure the *graph root distribution* (GRD).
2. We show that the Wasserstein distance between two graph root distributions on $\mathcal{K}$ provides an upper bound of the cut-distance between sampling distributions of the corresponding exchangeable random graphs. This result is further extended to a modified version of Wasserstein distance that is suitable for measuring the distance between two equivalence classes of graph root distributions.
3. We show that a truncated adjacency spectral embedding weighted by the square roots of the absolute eigenvalues can approximate the empirical distribution of the latent node vectors in $\mathcal{K}$ with vanishing Wasserstein distance error when the network size n goes to infinity, under suitable regularity conditions. This in turn implies that such a weighted truncated spectral embedding can also consistently estimate the underlying graph root distribution when the truncation dimension is chosen appropriately.
4. As demonstrated in our numerical examples, the new parameterization allows for a simple, theoretically justifiable estimator, which can reveal salient geometric features in the network data.

Related work. The GRD parameterization is closely related to latent space network models. The idea of modeling the edge probability between a pair of nodes by the inner product of the corresponding latent vectors is studied by [22] under the name of "latent eigenmodel," and further developed by [5, 46] under the name of "random dot-product graph." The GRD framework extends this idea to a population perspective and connects it with the graphon literature.

The GRD parameterization and estimation are also related to spectral methods in random graphs. The spectral approach has been used in graphon estimation by [12, 29, 51]. The GRD estimation method considered in Section 4 uses a similar singular value thresholding of the adjacency matrix, but takes a further step of weighted spectral embedding to recover the latent vectors corresponding to each node, with the target parameter being a distribution in the latent Kreĭn space. Besides spectral methods, graphon estimation has also been studied using histogram and smoothing methods, including theoretical analysis [17, 28, 50] and practical algorithms [3, 11, 53]. GRD estimation and graphon (or probability matrix) estimation are different, as the target parameters are in different spaces with different error metrics (see Sections 4.4 and 5.2 for further discussion and comparison). Some potential benefits of using GRD are discussed in Section 2.3, and visualized in some simulated and real data examples in Section 5.

2. Background.

2.1. *Exchangeable random graphs and the graphon parameterization.* Consider a random symmetric two-way binary array

$$\mathbf{A} = (A_{ij} : 1 \le i < j)$$

such that $A_{ii} = 0$ and satisfies the row-column joint exchangeability

$$(A_{ij} : i \ge 1, j \ge 1) \overset{d}{=} (A_{\sigma(i)\sigma(j)} : i \ge 1, j \ge 1)$$

for all finite index permutation mapping σ: for some $1 \le i_0 < j_0$,

$$\sigma(i) = \begin{cases} i & \text{if } i \notin \{i_0, j_0\}, \\ j_0 & \text{if } i = i_0, \\ i_0 & \text{if } i = j_0. \end{cases}$$

Here, "$\overset{d}{=}$" means that two random objects have the same distribution.

Analogous to the de Finetti theorem, the Aldous–Hoover theorem [4, 24, 26] says that any symmetric exchangeable binary array $\mathbf{A}$ of infinite size can be generated by sampling independent $(s_i : i \ge 1)$ from Unif(0, 1) (the uniform distribution on [0, 1]), and sampling A_{ij} independently from a Bernoulli distribution with probability $W(s_i, s_j)$ for a symmetric measurable function $W(\cdot, \cdot) : [0, 1]^2 \mapsto [0, 1]$. Here, W is a random object measurable in the doubly-exchangeable σ-field. For any given realization of $\mathbf{A}$, we can simply treat W as a nonrandom parameter.

Once W is given, the distribution of $\mathbf{A}$ is completely determined. However, the converse is not true. Let $h : [0, 1] \mapsto [0, 1]$ be a measure-preserving mapping in the sense that

$$\mu\big(h^{-1}(B)\big) = \mu(B) \quad \forall B \in \mathcal{B}_{[0,1]},$$

where $\mu(\cdot)$ denotes the Lebesgue measure and $\mathcal{B}_{[0,1]}$ is the Borel σ-field. Two functions W_1 and W_2 generate the same distribution of exchangeable arrays if and only if there exist two measure-preserving mappings h_1, h_2 such that

$$W_1\big(h_1(s), h_1(s')\big) = W_2\big(h_2(s), h_2(s')\big) \quad \text{a.e.} \tag{1}$$

When (1) holds, we say W_1 and W_2 are *weakly isomorphic*, denoted as

$$W_1 \overset{w.i.}{=} W_2.$$

The notion "$\overset{w.i.}{=}$" defines an equivalence relation on the space of all symmetric functions that map $[0, 1]^2$ to [0, 1]. We use $\tilde{W}$ to denote the equivalence class containing W.

When W_1 and W_2 are not weakly isomorphic, then they lead to different distributions of exchangeable random graphs. In this case, the subgraph counts have different distributions under W_1 and W_2. Such a sampling distribution difference can be linked to the cut-distance, defined as

$$\delta_\square(W_1, W_2) = \inf_{h_1, h_2} \sup_{S \times S' : \subseteq [0,1]^2} \left| \int_{S \times S'} \big[W_1\big(h_1(s), h_1(s')\big) - W_2\big(h_2(s), h_2(s')\big) \big] \, ds \, ds' \right|,$$

where h_1, h_2 range over all measure-preserving mappings. It can be shown that $W_1 \overset{w.i.}{=} W_2$ if and only if $\delta_\square(W_1, W_2) = 0$. Therefore, the cut-distance $\delta_\square(\cdot, \cdot)$ can also be used to measure the distance between two equivalence classes $\tilde{W}_1$ and $\tilde{W}_2$. We will also adopt the terminology used in [37] to call the function W a *graphon* (abbreviation for "graph function"). In the rest of this paper, we will often use a graphon W to represent its equivalence class.

2.2. *Graph root distributions*. The focus of this paper is to develop an alternative characterization of a subclass of exchangeable random graphs. The construction involves probability distributions on a separable Kreĭn space, which we introduce first.

DEFINITION 1 (Kreĭn space). A Kreĭn space $\mathcal{K} = \mathcal{H}_+ \ominus \mathcal{H}_-$ is the direct sum of two Hilbert spaces $\mathcal{H}_+$ and $\mathcal{H}_-$. For each $(x, y), (x', y') \in \mathcal{K}$ with $x, x' \in \mathcal{H}_+$ and $y, y' \in \mathcal{H}_-$, the Kreĭn inner product is

$$\langle (x, y), (x', y') \rangle_{\mathcal{K}} = \langle x, x' \rangle_{\mathcal{H}_+} - \langle y, y' \rangle_{\mathcal{H}_-}. \tag{2}$$

The space $\mathcal{K}$ is also a linear normed space isomorphic to the Hilbert space $\mathcal{H}_+ \oplus \mathcal{H}_-$ equipped with norm

$$\|(x, y)\|_{\mathcal{K}} = \big(\|x\|_{\mathcal{H}_+}^2 + \|y\|_{\mathcal{H}_-}^2\big)^{1/2}.$$

The notation "$\ominus$" is used to emphasize the nonpositive definite inner product associated with the space $\mathcal{K}$, and $\mathcal{H}_+$, $\mathcal{H}_-$ represent the subspaces containing the positive and negative components, respectively. This notation has been used in existing machine learning literature involving Kreĭn spaces [42]. The traditional notation "$\oplus$" is saved for the direct sum of Hilbert spaces in the usual sense, where the direct sum is still a Hilbert space with a positive definite inner product.

Now we define graph root distributions.

DEFINITION 2 (Graph root distribution (GRD)). We call a probability measure F on $\mathcal{K}$ a *graph root distribution*, if for two independent samples Z_1 and Z_2 from F,

$$\mathbb{P}\big(\langle Z_1, Z_2 \rangle_{\mathcal{K}} \in [0, 1]\big) = 1.$$

Let F be a GRD on $\mathcal{K}$. We can generate an exchangeable random graph by first generating independent random vectors $(Z_i : i \geq 1)$ from F and then generating A_{ij} independently from a Bernoulli distribution with parameter $\langle Z_i, Z_j \rangle_{\mathcal{K}}$. We call this sampling procedure the graph root sampling with F. In contrast, the graphon based sampling scheme is called graphon sampling with W.

The embedding of network nodes in a Kreĭn space $\mathcal{K}$ has a clear interpretation. Suppose each node i $(1 \leq i \leq n)$ corresponds to a $Z_i = (X_i, Y_i) \in \mathcal{K}$, with X_i, Y_i being the positive and negative components, respectively. Then two nodes i and j are more likely to connect if $\langle X_i, X_j \rangle$ is large, or equivalently, $\|X_i\| \|X_j\| \langle \tilde{X}_i, \tilde{X}_j \rangle$ is large (where $\tilde{X}_i = X_i / \|X_i\|$). The quantities $\|X_i\|$, $\|X_j\|$ measure how "active" the individuals i, j are, respectively, while the normalized inner product $\langle \tilde{X}_i, \tilde{X}_j \rangle$ measures how well the two individuals match each other. Analogous interpretations can be given to the negative components Y_i, Y_j.

Now we list how some commonly considered network models fit in the framework of GRD. Further explanations of the correspondence are given in the Supplementary Material [36]. Simulations based on these models are reported in Section 5.1.

Stochastic block models: *Point mass mixture*. A stochastic block model (SBM, [23]) with k blocks is parameterized by (π, B), where π is in the $(k-1)$-dimensional simplex and B is a $k \times k$ symmetric matrix with entries in $[0, 1]$. The exchangeable random graph is generated by sampling $(e_i : i \geq 1)$ independently from a multinomial distribution with parameter π, and connecting nodes i, j independently with probability B_{e_i, e_j}. The corresponding GRD is a mixture of no more than k point masses in a k-dimensional space $\mathcal{K}$, with the point mass locations determined by B and the point mass weights determined by π.

For example, consider an SBM with $k = 3$, $\pi = (1/3, 1/3, 1/3)$, and

$$B = \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/6 \end{pmatrix}, \tag{3}$$

which corresponds to three blocks, but the rank is only 2, with one positive eigenvalue and one negative eigenvalue. Then a corresponding GRD is (after rounding)

$$F = (1/3)\delta_{(0.61;0.35)} + (1/3)\delta_{(0.61;-0.35)} + (1/3)\delta_{(0.41;0)},$$

where $\delta_z = \delta_{(x;y)}$ denotes the point mass at $z = (x; y)$ and the semicolon is used to delineate positive and negative components.

Degree corrected block models: 1*-D subspace mixture*. The degree-corrected block model (DCBM, [27]) is parameterized by (π, B, Θ) where π, B are the same as in SBM and Θ is a distribution on $(0, \infty)$. The random graph is generated similarly as in SBM, except that it also generates $(\theta_i : i \geq 1)$ independently from Θ and the connection probability of nodes i, j is $\theta_i \theta_j B_{e_i, e_j}$. In this case, the corresponding GRD is a mixture of distributions, each supported on a line connecting one of the SBM point masses and the origin.

For example, if we use the same k, π and B as in the SBM example above, and set Θ to be the uniform distribution on [0, 1], then a GRD for this DCBM is

$$F = (1/3)U\big(\mathbf{0}, (0.61; 0.35)\big) + (1/3)U\big(\mathbf{0}, (0.61; -0.35)\big) + (1/3)U\big(\mathbf{0}, (0.41; 0)\big),$$

where $U(z, z')$ denotes the uniform distribution on the line segment between z and z'. Other distributions Θ are allowed, and can be chosen differently for each mixture component. This will lead to different ending points of line segments and the distributions on them.

Mixed membership block models: *Convex polytope*. The mixed membership block model (MMBM, [2]) allows each node to have a mixture of memberships. Given a matrix B as in the SBM, and a Dirichlet distribution $\mathrm{Dir}(a)$ with parameter $a \in (0, \infty)^k$, the random graph is generated by sampling $(g_i : i \geq 1)$ from $\mathrm{Dir}(a)$ independently, and connecting nodes (i, j) with probability $g_i^T B g_j$. In this case, the corresponding GRD is a distribution supported on the convex polytope with extreme points given by the SBM point masses determined by B.

For example, using the same B matrix as in the previous examples for SBM and DCBM, and choosing $a = (1, 1, 1)$ so that the Dirichlet distribution is uniform on the simplex, a GRD for this MMBM is

$$F = U\big((0.61; 0.35), (0.61; -0.35), (0.41; 0)\big),$$

where $U(z_1, z_2, z_3)$ denotes the uniform distribution on the convex hull of $\{z_1, z_2, z_3\}$.

Random dot-product graphs: *Finite-dimensional subspace*. The random dot-product graph [41] and generalized random dot-product graph [46] generate the random graph by connecting nodes (i, j) independently with probability $\langle X_i, X_j \rangle$, where $(X_i : 1 \leq i \leq n)$ are node covariate vectors in an Euclidean space. The original random dot-product graph only considers positive semidefinite inner products, while the generalized model allows for indefinite inner products in a similar fashion as we have defined for Kreĭn spaces. In principle, a generalized random dot-product graph can be viewed as a finite-sample realization of a GRD supported on a finite-dimensional space.

2.3. *Potential features and benefits of the graph root parameterization*. As we will see in the following section, GRDs can be used to parameterize exchangeable random graphs under some mild regularity conditions. Here, we list some features and potential benefits of the GRD parameterization.

GRDs are identifiable up to orthogonal transforms. Roughly speaking, two GRDs that differ by a pair of orthogonal transforms can lead to the same distribution of exchangeable random graphs. The choice of orthogonal transforms used in GRD estimation is straightforward: It diagonalizes the covariance operator of the embedded node vectors. Moreover, many important geometric features of the distribution, such as clusters, pairwise distances, and the shape of support, are invariant under orthogonal transforms. In contrast, graphon estimators that do not attempt to recover the ordering of the nodes cannot reveal the same geometric structures in the data. This difference is illustrated in our numerical examples in Sections 5.1, 5.3 and 5.4.

GRD provides new tools for some inference problems. The embedding of network nodes as independent realizations of a common latent distribution makes it possible to apply the methods and tools developed for i.i.d. data to network related problems. For example, suppose we observe an exchangeable random graph with n nodes, where each node i is associated with a covariate vector $U_i \in \mathbb{R}^d$. Here, the nodes can be students in a school, edges represent friendship and covariates are demographic information. A question of interest is to test whether the covariate and the network are independent. In the GRD parameterization, the problem can be formulated as testing independence of two random vectors (U_i, Z_i) in a paired sample, where Z_i is the Kreĭn space embedding of node i. In a second example, suppose we have two exchangeable random graphs on two disjoint set of individuals. Here, the two networks can be student friendship networks from different middle schools. Then one may want to test whether these two networks have the same distribution. In GRD parameterization, this reduces to testing equality (up to orthogonal transform) of two distributions using independent samples. Moreover, if the underlying GRDs are the same, it is even possible to aggregate two independently estimated GRDs, as well as to predict edge probabilities between nodes in different samples.

GRD provides connection between the graphon, spectral embedding and latent space perspectives. Spectral embedding of network vertices has been an active research topic related to exchangeable random graphs, especially stochastic block models [15, 25, 38, 45, 48]. Examples of spectral embeddings beyond stochastic block models include the random dot product graph [5, 46] and the latent eigenmodel [22]. The GRD parameterization shows that such embeddings exist in an infinite-dimensional space for all trace-class graphons, partially reconciling the graphon, spectral embedding and latent space model literature.

3. Existence, identifiability and topology of GRD. From now on, we only consider separable Kreĭn spaces, where $\mathcal{K} = \mathcal{H}_+ \ominus \mathcal{H}_-$, $\mathcal{H}_+ = \mathcal{H}_- = \{x \in \mathbb{R}^\infty : \sum_j x_j^2 < \infty\}$ with inner product $\langle x, x' \rangle_{\mathcal{H}_\pm} = \sum_{j \geq 1} x_j x'_j$. These spaces are associated with the Borel σ-field.

3.1. *Existence of GRD for exchangeable random graphs*. To find an underlying GRD for an exchangeable random graph, we consider the spectral decomposition of the corresponding graphon. We will show that such GRDs exist for graphons whose spectral series converge in a strong sense.

Recall that a graphon W is a symmetric function from $[0,1]^2$ to $[0,1]$. We can view W as an integral operator on $L^2([0,1])$:

$$(Wf)(\cdot) = \int_{[0,1]} W(\cdot, s) f(s)\, ds \quad \forall f \in L^2\big([0,1]\big).$$

Since $\int_{[0,1]^2} W(s,s')^2\, ds\, ds' \leq 1$, W is a Hilbert–Schmidt operator, and hence admits a spectral decomposition

$$W(s,s') = \sum_{j=1}^{\infty} \lambda_j \phi_j(s) \phi_j(s') - \sum_{j=1}^{\infty} \gamma_j \psi_j(s) \psi_j(s'), \tag{4}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots > 0$, $\gamma_1 \geq \gamma_2 \geq \ldots > 0$, and $(\phi_j : j \geq 1) \cup (\psi_j : j \geq 1)$ are orthonormal functions in $L^2([0,1])$. The convergence in (4) shall be interpreted as L^2-convergence. In general, almost everywhere convergence does not hold without further assumptions.

DEFINITION 3 (Strong spectral decomposition). We say a graphon W admits *strong spectral decomposition* if the eigencomponents $(\lambda_j, \phi_j)_{j\geq 1}$, $(\gamma_j, \psi_j)_{j\geq 1}$ in (4) satisfy

$$\sum_{j\geq 1} [\lambda_j \phi_j^2(s) + \gamma_j \psi_j^2(s)] < \infty \quad \text{a.e.} \tag{5}$$

Strong spectral decomposition implies, among other things, that the sum in (4) converges almost everywhere.

The spectral decomposition of a graphon has been considered in the mathematical side of the literature, such as in [9, 26, 37], and recently in graphon estimation in [51]. Here, we use the spectral decomposition to define a mapping from $[0,1]$ to a pair of infinite sequences: $[\sqrt{\lambda_j}\phi_j(s), j \geq 1]$ and $[\sqrt{\gamma_j}\psi_j(s), j \geq 1]$, and our key object, the graph root distribution (GRD), is the corresponding induced probability measure on the infinite-dimensional space. Such an induced probability measure carries all the information about the corresponding exchangeable random graph, and removes the ambiguity caused by measure preserving transforms.

THEOREM 3.1 (Graph root representation). *Any exchangeable random graph generated by a graphon that admits strong spectral decomposition can be generated by a GRD on $\mathcal{K}$.*

PROOF OF THEOREM 3.1. For a graphon W, consider its spectral decomposition (4), and define $Z(s) = (X(s), Y(s))$ as

$$\begin{aligned} X_j(s) &= \lambda_j^{1/2} \phi_j(s) \quad \forall j \geq 1, \\ Y_j(s) &= \gamma_j^{1/2} \psi_j(s) \quad \forall j \geq 1. \end{aligned} \tag{6}$$

If $s \sim \text{Unif}(0,1)$, the resulting $Z(s) = (X(s), Y(s))$ is a random object. By the strong spectral decomposition assumption, $\|X\|_{\mathcal{H}_+}$ and $\|Y\|_{\mathcal{H}_-}$ are finite with probability one, so Z is a well-defined random vector in $\mathcal{K}$. Moreover,

$$\langle Z(s), Z(s') \rangle_{\mathcal{K}} = \sum_j [\lambda_j \phi_j(s)\phi_j(s') - \gamma_j \psi_j(s)\psi_j(s')]$$

converges almost everywhere since for s, s' we have

$$\begin{aligned} |\lambda_j \phi_j(s)\phi_j(s')| &\leq \frac{1}{2}(\lambda_j \phi_j^2(s) + \lambda_j \phi_j^2(s')), \\ |\gamma_j \psi_j(s)\psi_j(s')| &\leq \frac{1}{2}(\gamma_j \psi_j^2(s) + \gamma_j \psi_j^2(s')), \end{aligned}$$

and the summability is ensured for all s and s' satisfying (5).

Let F be the probability measure induced by $Z(s) : [0,1] \mapsto \mathcal{K}$ with $s \sim \text{Unif}(0,1)$. By construction, $W(s,s') = \langle Z(s), Z(s') \rangle_{\mathcal{K}}$ almost everywhere, so that the graphon W and GRD F lead to the same sampling distribution of exchangeable random graph. □

The examples given in Section 2.2 are special cases of Theorem 3.1. We provide more detailed explanation of the correspondence in the Supplementary Material [36].

3.2. *How stringent is strong spectral decomposition?*. The requirement of strong spectral decomposition is, indeed, quite mild. The following proposition states that trace-class integral operators admit strong spectral decomposition.

PROPOSITION 3.2. *If W is trace-class, in the sense that $\sum_{j\geq 1}(\lambda_j + \gamma_j) < \infty$ with λ_j, γ_j defined in (4), then W admits strong spectral decomposition, and the GRD constructed from the spectral decomposition of W is square-integrable.*

Trace-class integral operators are well studied in functional analysis [19, 32]. Some important subclasses are the following:

1. Finite rank graphons. If W has finite rank, then it only has finitely many nonzero eigenvalues, and hence strong spectral decomposition holds trivially. Important examples covered by this case include the stochastic block models, the degree corrected block models, and the mixed membership block models.
2. Smooth graphons. If a graphon W, or an element in its equivalence class, is in α-Hölder class:

$$\left|W(x, y) - W(x, y')\right| \leq C\left|y - y'\right|^{\alpha}$$

for constants $C > 0$, $\alpha > 1/2$ and all $x, y, y' \in [0, 1]$, then it is trace-class, and hence admits strong spectral decomposition.
3. Continuous positive graphons. If W is positive semidefinite and continuous, then W is trace-class, and hence admits strong spectral decomposition. This is the famous Mercer's theorem. One can relax the requirement of positivity by instead requiring $W = W_+ - W_-$ with W_+, W_- both being positive semidefinite and continuous. An example covered in this case is

$$W(x, y) = \frac{1}{\log(e/x)\log(e/y)} \quad \text{if } (x, y) \in (0, 1]^2$$

and $W(x, y) = 0$ if $x = 0$ or $y = 0$. This graphon is not in any Hölder class but is trace-class.

The second case in the list above has a useful consequence. Even though not all continuous graphons are trace-class, one can approximate any continuous graphon arbitrarily well using trace-class graphons.

PROPOSITION 3.3. *Let W be a continuous graphon. For any $\epsilon > 0$, there exists a trace-class graphon W' such that*

$$\delta_{\square}(W, W') \leq \epsilon.$$

Moreover, the set of trace-class continuous graphons is a dense subset of continuous graphons.

If a continuous graphon W does not satisfy strong spectral decomposition, the approximation W' given in Proposition 3.3 with a small approximation error ϵ may have a very large trace, and its eigenvalues may decay very slowly. This can pose challenges in estimation. We make further discussion in Section 4.2.

3.3. *Identifiability of GRD*. When do two graph root distributions F_1 and F_2 on $\mathcal{K}$ lead to the same exchangeable random graph distribution? We first exclude some trivial sources of ambiguity.

Ambiguity by concatenation. Let (X, Y) be a random vector on $\mathcal{K}$, and let R be any random variable in an Euclidean space or separable Hilbert space, the random vector (X', Y') in the augmented space with $X' = (X, R)$, $Y' = (Y, R)$ leads to the same random graph sampling distribution as (X, Y).

Ambiguity by rotation. Let Q be an inner product preserving mapping from $\mathcal{K}$ to $\mathcal{K}$:

$$\langle z, z' \rangle_{\mathcal{K}} = \langle Qz, Qz' \rangle_{\mathcal{K}} \quad \forall z, z' \in \mathcal{K}.$$

Then, in terms of the generated exchangeable random graph, a GRD F is indistinguishable from F_Q, the measure induced by transforming $Z \sim F \mapsto QZ$. An obvious example of Q is the direct sum of two orthogonal transforms Q_+, Q_- on $\mathcal{H}_+$, $\mathcal{H}_-$, respectively, such that $Q(x, y) = (Q_+x, Q_-y)$. Such a Q preserves the inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ because it preserves the inner products in both the positive and negative components. However, due to the indefinite inner product, this is not the only type of inner product preserving transforms on $\mathcal{K}$. Other transforms, such as hyperbolic rotations, can also preserve the indefinite inner product. See [46] for some examples of hyperbolic rotations under the context of random dot-product graphs.

To resolve the identifiability issue, a key observation is that both concatenation and hyperbolic rotation necessarily mix up the positive and negative components. So these ambiguities can be precluded by the requiring uncorrelated positive and negative components. In this subsection, we show that the direct sum of a pair of orthogonal transforms is the only possible ambiguity in identifying a square-integrable GRD with uncorrelated positive and negative components.

DEFINITION 4 (Equivalence up to orthogonal transforms). We say two distributions F_1, F_2 on $\mathcal{K}$ are equivalent up to orthogonal transform, written as $F_1 \overset{o.t.}{=} F_2$, if there exist orthogonal transforms Q_+ on $\mathcal{H}_+$ and Q_- on $\mathcal{H}_-$, such that $(X, Y) \sim F_1 \Leftrightarrow (Q_+X, Q_-Y) \sim F_2$.

THEOREM 3.4 (Identifiability of GRD). *Two square-integrable GRDs F_1, F_2 with uncorrelated positive and negative components give the same exchangeable random graph sampling distribution if and only if $F_1 \overset{o.t.}{=} F_2$.*

The main idea of the proof is to establish a direct connection between a GRD F and its corresponding graphon W. Now F is a probability measure on $\mathcal{K}$, while W is a function from $[0, 1]^2 \to [0, 1]$. Our idea is to use an *inverse transform sampling mapping* to relate the distribution F to a measurable function on $[0, 1]$.

DEFINITION 5 (Inverse transform sampling (ITS)). Let F be a distribution on $\mathcal{K}$. A measurable function $Z : [0, 1] \mapsto \mathcal{K}$ is called an *inverse transform sampling mapping* of F if

$$s \sim \text{Unif}(0, 1) \quad \Rightarrow \quad Z(s) \sim F.$$

In other words, an ITS induces the Lebesgue measure on $[0, 1]$ to F on $\mathcal{K}$. The mapping $Z(\cdot)$ given by (6) in the proof of Theorem 3.1 is an example of an ITS of the GRD F. If $\mathcal{K}$ is one-dimensional, then a well-known example of ITS is the inverse cumulative distribution function. It is also straightforward to see that ITS's are not unique since if $Z(\cdot)$ is an ITS of F and $h(\cdot)$ is measure-preserving then $Z(h(\cdot))$ is also an ITS of F. The following result ensures that ITS's always exist for distributions on a separable Hilbert space.

PROPOSITION 3.5 (Existence of ITS). *Let F be a distribution on a separable Hilbert space, then there exists an ITS of F.*

Here, we give a sketch of proof of Theorem 3.4. For $i = 1, 2$, let $Z_i(s) = (X_i(s), Y_i(s))$ be an ITS of F_i and define graphon

$$W_i(s, s') = \langle Z_i(s), Z_i(s') \rangle_{\mathcal{K}}. \tag{7}$$

By assumption that F_1 and F_2 lead to the same exchangeable random graph sampling distribution, we have $W_1 \overset{w.i.}{=} W_2$. Also, by choosing appropriate orthogonal rotations in the positive and negative components we can make the covariance of Z_i diagonal so that $W_i(s, s') = \langle Z_i(s), Z_i(s') \rangle_{\mathcal{K}}$ indeed corresponds to the spectral decomposition of W_i. Then the desired result follows by invoking an exchangeable array representation theorem in the form of spectral decompositions due to Kallenberg [26].

We summarize our representation results in the following corollary.

COROLLARY 3.6 (Correspondence between graphon and GRD). *There exists a one-to-one correspondence between trace-class graphons* (*under the equivalence relation "$\overset{w.i.}{=}$"*) *and square-integrable GRD's with uncorrelated positive and negative components* (*under the equivalence relation "$\overset{o.t.}{=}$"*).

Canonical GRD. Since any square-integrable GRD with uncorrelated positive and negative components is identifiable up to a pair of orthogonal transforms, we can choose appropriate orthogonal transforms so that the covariance of the GRD is diagonalized. Such a choice can be used as a canonical representation. If all eigenvalues of the covariance operator have multiplicity one, then the canonical GRD F is determined up to the sign of each coordinate. As we will see in Section 4 below, our estimator recovers one of the canonical GRD's.

3.4. *Topology of the GRD space*: *Orthogonal Wasserstein distance*. Having established the GRD representation of exchangeable random graphs, we can study the closeness of graph sampling distributions by looking at the closeness of GRD's. To this end, we consider a metric on the quotient space of square-integrable distributions on $\mathcal{K}$ with respect to the equivalence relation "$\overset{o.t.}{=}$", which we call the *orthogonal Wasserstein metric*. We will show that convergence of a sequence of GRD's in this metric implies convergence of corresponding graphons in cut-distance.

We start by recalling the Wasserstein distance. We will only use a special case of the Wasserstein distance suitable for our purpose. Given two probability distributions F_1, F_2 on $\mathcal{K}$, the Wasserstein distance between F_1, F_2 is

$$d_{\mathrm{w}}(F_1, F_2) := \inf_{\nu \in \mathcal{V}(F_1, F_2)} \mathbb{E}_{(Z_1, Z_2) \sim \nu} \|Z_1 - Z_2\|,$$

where $\mathcal{V}(F_1, F_2)$ is the collection of all distributions on $\mathcal{K} \times \mathcal{K}$ with F_1 and F_2 being its two marginal distributions.

The following lemma says that if two square-integrable GRD's are close in Wasserstein distance, then the corresponding graphons are close in cut-distance.

LEMMA 3.7 (Wasserstein and cut distances). *Let F_1 and F_2 be two square-integrable GRD's on $\mathcal{K}$, with corresponding graphons W_1, W_2 defined using ITS as in* (7). *Then*

$$\delta_{\square}(W_1, W_2) \le \big(\mathbb{E}_{Z \sim F_1} \|Z\| + \mathbb{E}_{Z \sim F_2} \|Z\|\big) d_{\mathrm{w}}(F_1, F_2).$$

Since we do not distinguish two GRD's differing only by orthogonal transforms on positive and negative components, we consider the *orthogonal Wasserstein distance*

$$d_{\mathrm{ow}}(F_1, F_2) := \inf_{\nu \in \mathcal{V}(F_1, F_2)} \inf_{Q_+, Q_-} \mathbb{E}_{(Z_1, Z_2) \sim \nu} \big\| Z_1 - (Q_+ \oplus Q_-) Z_2 \big\|, \tag{8}$$

where Q_+, Q_- range over all orthogonal transforms on $\mathcal{H}_+$, $\mathcal{H}_-$, respectively, and $(Q_+ \oplus Q_-)$ denotes the orthogonal transform as the direct sum of Q_+ and Q_-: $(Q_+ \oplus Q_-)(x, y) = (Q_+x, Q_-y)$.

We can improve Lemma 3.7 to the orthogonal Wasserstein distance, which says that orthogonal Wasserstein distance induces a stronger topology than cut-distance.

THEOREM 3.8. *Let F_1, F be two square-integrable GRD's on $\mathcal{K}$, with corresponding graphons W_1, W as obtained by ITS in (7). Then*

$$\delta_\square(W_1, W) \le \big(\mathbb{E}_{Z\sim F_1}\|Z\| + \mathbb{E}_{Z\sim F}\|Z\|\big) d_{\mathrm{ow}}(F_1, F).$$

As a consequence, if $(F_N : N \ge 1)$ are square-integrable GRD's on $\mathcal{K}$ with corresponding graphons $(W_N : N \ge 1)$, then

$$d_{\mathrm{ow}}(F_N, F) \to 0 \quad \Rightarrow \quad \delta_\square(W_N, W) \to 0.$$

4. Estimation of graph root distributions. Given $n \ge 1$, suppose we have observed an $n \times n$ block of $\mathbf{A}$: $\mathbf{A}_n = (A_{ij} : 1 \le i, j \le n)$, where $\mathbf{A}$ is generated from a GRD F. In such finite sample scenarios, GRD and graphon are used as modeling tools and are no longer linked to the infinite exchangeability, since the Aldous–Hoover theorem is only applicable to the infinite case. We consider the following two inference questions:

1. Node embedding: Can we recover the realized sample of node vectors $Z_1, .., Z_n$ in $\mathcal{K}$?
2. Distribution estimation: Can we recover the GRD F with small orthogonal Wasserstein distance?

Notation. For an infinite vector x, $x^{(p)}$ denotes the first p elements of x. For a matrix $\mathbf{M}$ with countably infinite number of columns, $\mathbf{M}^{(p)}$ denotes the submatrix consisting of the first p columns. For a matrix $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ with n rows and each row taking value in $\mathcal{K}$, $\mathbf{Z}^{(p_1,p_2)} = (\mathbf{X}^{(p_1)}, \mathbf{Y}^{(p_2)})$.

4.1. *Truncated weighted spectral embedding*. Write $\mathbf{A}_n$ in its eigendecomposition

$$\mathbf{A}_n = \sum_{j=1}^{n_1} \hat\lambda_{j,A} \hat a_j \hat a_j^T - \sum_{j=1}^{n-n_1} \hat\gamma_{j,A} \hat b_j \hat b_j^T,$$

where $\hat\lambda_{1,A} \ge \hat\lambda_{2,A} \ge \cdots \ge \hat\lambda_{n_1,A} \ge 0$ are the nonnegative eigenvalues of A, and $\hat\gamma_{1,A} \ge \cdots \ge \hat\gamma_{n-n_1,A} > 0$ are the absolute negative eigenvalues of A.

Let $p_1, p_2 < n$ be nonnegative integers to be specified later. We consider the weighted $(p_1 + p_2)$-dimensional spectral embedding of the nodes

$$\begin{aligned} \hat{\mathbf{Z}}_A &= [\hat{\mathbf{X}}_A^{(p_1)}, \hat{\mathbf{Y}}_A^{(p_2)}], \\ \hat{\mathbf{X}}_A^{(p_1)} &= [\hat\lambda_{1,A}^{1/2} \hat a_1, \ldots, \hat\lambda_{p_1,A}^{1/2} \hat a_{p_1}] = [\hat a_1, \ldots, \hat a_{p_1}] \hat\Lambda_{p_1,A}^{1/2}, \\ \hat{\mathbf{Y}}_A^{(p_2)} &= [\hat\gamma_{1,A}^{1/2} \hat b_1, \ldots, \hat\gamma_{p_2,A}^{1/2} \hat b_{p_2}] = [\hat b_1, \ldots, \hat b_{p_2}] \hat\Gamma_{p_2,A}^{1/2}, \end{aligned} \tag{9}$$

where $\hat\Lambda_{p_1,A}$ is the $p_1 \times p_1$ diagonal matrix with diagonal entries being $(\hat\lambda_{1,A}, \ldots, \hat\lambda_{p_1,A})$, and $\hat\Gamma_{p_2,A}$ is defined similarly.

We use the rows of $\hat{\mathbf{Z}}_A$ to estimate the realized sample points Z_1, ..., Z_n as follows:

$$\begin{aligned} \hat X_{i,A} &= (\hat\lambda_{1,A}^{1/2} \hat a_{1i}, \ldots, \hat\lambda_{p_1,A}^{1/2} \hat a_{p_1,i}, 0, \ldots) \in \mathcal{H}_+, \\ \hat Y_{i,A} &= (\hat\gamma_{1,A}^{1/2} \hat b_{1i}, \ldots, \hat\gamma_{p_2,A}^{1/2} \hat b_{p_2,i}, 0, \ldots) \in \mathcal{H}_-. \end{aligned} \tag{10}$$

In other words, $\hat{X}_{i,A}$, $\hat{Y}_{i,A}$ are the ith row of $\hat{\mathbf{X}}_A^{(p_1)}$, $\hat{\mathbf{Y}}_A^{(p_2)}$, padded with zeros in the tails. The same weighted spectral embedding has been considered in random dot-product graphs in finite-dimensional spaces and at the sample level [5, 41, 46]. Here, we focus more on the infinite-dimensional case, where p_1, p_2 need to grow with n, and study the statistical properties of the embeddings at a population level with a goal of estimating the GRD F.

4.2. *Reconstruction error of sample points.* In order to show that the estimated node vectors $(\hat{X}_{i,A}, \hat{Y}_{i,A})$ are close to the true but hidden realized node vectors (X_i, Y_i), it is necessary to identify a particular orthogonal transform $Q = Q_+ \oplus Q_-$ to work with. To this end, we make the following assumption to clear the identifiability issue:

(A1) For all $j, j' \geq 1$, $\mathbb{E}_{(X,Y)\sim F}(X_j X_{j'}) = \lambda_j \mathbf{1}(j = j')$, $\mathbb{E}_{(X,Y)\sim F}(Y_j Y_{j'}) = \gamma_j \mathbf{1}(j = j')$, $\mathbb{E}_{(X,Y)\sim F}(X_j Y_{j'}) = 0$.

This assumption is nontechnical, it merely says that we pick a canonical element among all possible orthogonal transforms on the positive and negative spaces.

Our next assumption is a polynomial eigendecay and eigengap condition.

(A2) There exist positive numbers $c_1 \leq c_2$, $1 < \alpha \leq \beta$ such that for all $j \geq 1$,

$$c_1 j^{-\alpha} \leq (\lambda_j \wedge \gamma_j) \leq (\lambda_j \vee \gamma_j) \leq c_2 j^{-\alpha},$$

$$(\lambda_j - \lambda_{j+1}) \wedge (\gamma_j - \gamma_{j+1}) \geq c_1 j^{-\beta},$$

with λ_j, γ_j defined in assumption (A1).

Assumption (A2) is often used in the literature of functional data analysis, where one needs to control the estimation error of individual eigenvectors for random variables in Hilbert spaces, using a truncated empirical eigen decomposition [21, 33, 39]. When $\lambda_j \propto j^{-\alpha}$, the eigengap condition usually holds with $\beta = \alpha + 1$. The random vector $Z = (X, Y) \sim F$ is square-integrable if $\alpha > 1$.

Assumption (A2) may seem a bit too stringent. Indeed we only need to consider the first $p_1 + 1$ ($p_2 + 1$) positive (negative) eigenvalues. The estimation error bound can be given as a function of all individual gaps between these eigenvalues, where equal or nearly equal eigenvalues can be treated by considering the corresponding principal subspace. This will make the presentation too cumbersome and will not change much of the nature of our argument. Another simplification made in Assumption (A2) is that the positive eigenvalues $(\lambda_j : j \geq 1)$ and negative eigenvalues $(\gamma_j : j \geq 1)$ decay at the same speed. Our analysis does allow for different decay speeds for the positive and negative eigenvalues. That will require choosing p_1 and p_2 separately, which involves a heavier notation. We choose to work with the version of Assumption (A2) stated above for presentation simplicity.

As mentioned in the discussion after Proposition 3.3, if we use a trace-class graphon W' to approximate a continuous graphon W that does not satisfy strong spectral decomposition, the eigenvalues of W' may decay slowly, which corresponds to a small value of α in Assumption (A2), and leads to a slower rate of convergence of the estimation error bound.

Finally, our procedure requires accurate estimation of the eigenvectors, which in turn requires accurate estimation of the covariance operator. We assume that the GRD has finite fourth moment.

(A3) $\mathbb{E}_{Z\sim F}\|Z\|^4 < \infty$.

THEOREM 4.1 (Sample points recovery). *Let $\mathbf{A}_n$ be generated from a GRD F satisfying (A1-A3). Let $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ be the hidden node data matrix with the ith row being $(X_i, Y_i) \in \mathcal{K}$ $(1 \leq i \leq n)$. If*

$$p_1 = p_2 = p = o\big(n^{\frac{1}{2\beta+\alpha}}\big),$$

then the estimator $\hat{\mathbf{Z}}_A$ *given in* (9) *satisfies*

$$n^{-1}\|\hat{\mathbf{Z}}_A - \mathbf{Z}^{(p,p)}\|_F^2 = O_P\big(n^{-\frac{\alpha-1}{2\beta}} + p^{2\beta+1}n^{-1}\big),$$

where $\|\cdot\|_F$ *denotes the Frobenius norm. As a consequence, let* $\hat{F}_A^{(p)}$ *be the empirical distribution putting* $1/n$ *probability mass at each row of* $\hat{\mathbf{Z}}_A$, *and* $\hat{F}^{(p)}$ *putting* $1/n$ *mass at each row of* $\mathbf{Z}^{(p,p)}$, *then*

$$d_{\mathrm{w}}\big(\hat{F}_A^{(p)}, \hat{F}^{(p)}\big) = O_P\big(n^{-\frac{\alpha-1}{4\beta}} + p^{\beta+1/2}n^{-1/2}\big).$$

The main technical task in the proof is to control the difference between the true realized random vectors (X_i, Y_i) and their projections on principal subspaces obtained from several approximations of the gram matrix. Thus the tools used are similar to those in functional data analysis [21, 39]. However, the additional challenge here is that we do not observe any of the empirical covariance matrices, and the adjacency matrix we observe is actually a noisy version of the indefinite gram matrix, which is the difference of two positive semidefinite gram matrices, one in the positive space and one in the negative space. This issue does not exist in the ordinary functional data analysis literature and requires more delicate spectral perturbation analysis.

4.3. *Estimating the GRD.* The second part of Theorem 4.1 gives the possibility of estimating the GRD F using $\hat{F}_A^{(p)}$. According to Theorem 4.1, we only need to show that $\hat{F}^{(p)}$ is close to F. To this end, we consider an intermediate object $F^{(p)}$, the distribution of truncated vector $(X^{(p)}, Y^{(p)})$ with $(X, Y) \sim F$. The argument proceeds in two steps.

The first step is to compare F and $F^{(p)}$, which is straightforward.

LEMMA 4.2. *Under assumptions* (A1)–(A2),

$$d_{\mathrm{w}}\big(F, F^{(p)}\big) \le cp^{-(\alpha-1)/2}.$$

The second part is comparing the population truncated distribution $F^{(p)}$ and its empirical version $\hat{F}^{(p)}$. We apply the result of [35] which provides Wasserstein error bounds for empirical distributions. Here, we state a special case, which is suitable for our purpose.

LEMMA 4.3 (Adapted from [35]). *Under assumptions* (A1)–(A3), *there exists a constant* c *independent of* n, p, *such that*

$$\mathbb{E}d_{\mathrm{w}}\big(\hat{F}^{(p)}, F^{(p)}\big) \le cn^{-1/(p\vee 2)}\big[1 + (\log n)\mathbf{1}(p=2)\big],$$

where $\mathbf{1}(\cdot)$ *is the indicator function.*

Combining the above two lemmas with Theorem 4.1, we have the following result on estimating the GRD.

THEOREM 4.4 (GRD estimation error). *Under assumptions* (A1)–(A3), *we have, when* $p \ge 3$ *satisfies the conditions in Theorem* 4.1,

$$d_{\mathrm{w}}\big(\hat{F}_A^{(p)}, F\big) = O_P\big[n^{-\frac{\alpha-1}{4\beta}} + p^{\beta+\frac{1}{2}}n^{-1/2} + p^{-(\alpha-1)/2} + n^{-1/p}\big].$$

The right-hand side is $o_P(1)$ *if* $p \to \infty$ *and* $p = o(\log n)$.

This result seems to suggest that one must have $p \to \infty$ to have a vanishing error. The term $p^{-(\alpha-1)/2}$ comes from Lemma 4.2 which corresponds to the truncation error. It is necessary only because we assumed a particular eigenvalue sequence in Assumption (A2). What really matters here is the sum of absolute eigenvalues beyond p: $(\lambda_j, \gamma_j : j > p)$. When the graphon is nearly low rank or the GRD is supported close to a finite-dimensional space, there is no need to use a large value of p.

4.4. *Estimation for sparse graphs*. One limitation of the theoretical framework of exchangeable random graphs is that they can only model dense graphs. The total number of edges in $\mathbf{A}_n$ will concentrate around $n^2 \int_{[0,1]^2} W$. In reality, the number of edges in a network rarely grows as the squared number of nodes. Therefore, sparse networks are of greater practical interest. To this end, for a given graphon W and a node sample size n one can consider adding a "sparsity parameter" to the network sampling scheme [7, 8, 28, 50, 52]:

$$\mathbf{A}_{n,i,j} \sim \text{Bernoulli}\big(\rho_n W(s_i, s_j)\big) \quad \forall 1 \le i < j \le n.$$

This sparsity parameter can be carried over to the graph root sampling scheme. Let F be a GRD. For a node sample size n and sparsity parameter ρ_n, the corresponding sparse graph root sampling scheme is equivalent to generating node sample points from a scaled distribution:

$$\mathbf{A}_{n,i,j} \sim \text{Bernoulli}\big(\langle \rho_n^{1/2} Z_i, \rho_n^{1/2} Z_j \rangle_{\mathcal{K}}\big), \tag{11}$$

where $Z_i \overset{iid}{\sim} F$. For notational simplicity, for scalar a and distribution F we use aF to denote the distribution obtained by scaling the distribution F by a factor of a: $Z \sim F \Leftrightarrow aZ \sim aF$.

In the SBM and DCBM literature, it is well known that consistent estimation of network communities is possible only if $n\rho_n \to \infty$. Our estimation theory developed in the previous subsections can be extended to cover sparse sampling schemes. One technical challenge is that when $n\rho_n = o(\log n)$, spectral methods tend to be sensitive to overly large node degrees. In the spectral clustering literature [14, 15], a common approach is to zero out rows and columns of $\mathbf{A}_n$ for which the degrees are too high. Some data-driven degree thresholding rules are developed, for example, in [6, 18]. In the following, we consider an adaptive trimmed spectral embedding method.

Let $d_i = \sum_{1 \le j \le n} \mathbf{A}_{n,i,j}$ be the degree of node i. Let I_n be the set of nodes whose degrees are among the $\lfloor \frac{n(n-1)}{\sum_{i=1}^n d_i} \rfloor$ largest, and $\tilde{\mathbf{A}}_n$ be the adjacency matrix obtained by zeroing out the columns and rows in I_n. Let $\tilde{\mathbf{Z}}_A^{(p,p)}$ and $\tilde{F}_A$ be the corresponding embeddings and GRD estimate defined in Section 4.1 and Theorem 4.1, respectively, with $\mathbf{A}_n$ replaced by $\tilde{\mathbf{A}}_n$.

THEOREM 4.5. *Under assumptions* (A1)–(A3), *assuming sparse sampling scheme* (11), *the following hold*:

1. *If* $n\rho_n \to \infty$ *and* $p_1 = p_2 = p = o[n^{1/(2\beta+\alpha)} \wedge (n\rho_n)^{1/(2\beta)}]$, *then*

$$n^{-1} \big\| \rho_n^{-1/2} \tilde{\mathbf{Z}}_A^{(p,p)} - \mathbf{Z}^{(p,p)} \big\|_F^2 = O_P\big(p^{2\beta-\alpha+1}(n\rho_n)^{-1} + n^{-\frac{\alpha-1}{2\beta}} + p^{2\beta+1} n^{-1}\big).$$

2. *If in addition we assume* $p \ge 3$, *then*

$$d_{\mathrm{w}}\big(\rho_n^{-1/2} \tilde{F}_A, F\big) = O_P\big(p^{\beta-(\alpha-1)/2}(n\rho_n)^{-1/2} + n^{-\frac{\alpha-1}{4\beta}} + p^{\beta+1/2} n^{-1/2} + p^{-\frac{\alpha-1}{2}} + n^{-1/p}\big),$$

where the error bound is $o_P(1)$ *if* $p \to \infty$ *and* $p = o(\log n \wedge (n\rho_n)^{1/(2\beta)})$.

Comparison to graphon estimation using spectral methods. Graphon estimation using singular value thresholding has been considered in [12, 29, 51]. For specificity of discussion, we focus on [51]. The method first performs a singular value decomposition of the adjacency matrix $\mathbf{A}_n$, and keeps only the components whose singular values exceed a threshold τ. Then the remaining low-rank approximation is multiplied by ρ_n^{-1} and entrywise trimmed to $[0, 1]$

to obtain $\hat{\mathbf{G}}_n$ as an approximation of the probability matrix $\mathbf{G}_n = \rho_n^{-1}\mathbb{E}\mathbf{A}_n$, which can be further used as a piecewise constant approximation to the graphon. The error rates reported in [51] and discussed here refer to the probability matrix estimation error.

We first explain the difference between the GRD estimation problem and probability matrix estimation problem. In GRD estimation, an intermediate error quantity is the empirical distribution approximation error $n^{-1}\sum_{i=1}^n \|\hat{Z}_i - Z_i\|^2$, where Z_i's are the true latent vectors sampled from the underlying GRD and $\hat{Z}_i$'s are the estimated versions. The total GRD estimation error bound needs to add another term due to the empirical distribution approximation. On the other hand, the probability matrix estimation is concerned with the error metric $n^{-2}\|\hat{\mathbf{G}}_n - \mathbf{G}_n\|_F^2$. Assuming that the GRD estimation and spectral probability matrix estimation use the same rules to select significant eigen components, and ignoring the trimming step in obtaining $\hat{\mathbf{G}}_n$, we have $\hat{\mathbf{G}}_{n,i,j} = \langle \hat{Z}_i, \hat{Z}_j \rangle_{\mathcal{K}}$. Using Cauchy–Schwarz, we obtain

$$
\begin{aligned}
&n^{-2}\|\hat{\mathbf{G}}_n - \mathbf{G}_n\|_F^2 \\
&\quad \le 2\left[n^{-1}\sum_{i=1}^n \|\hat{Z}_i - Z_i\|^2\right]\left[n^{-1}\sum_{i=1}^n \|Z_i\|^2 + n^{-1}\sum_{i=1}^n \|\hat{Z}_i\|^2\right].
\end{aligned} \tag{12}
$$

If we accept the assumption that $n^{-1}\sum_{i=1}^n \|\hat{Z}_i\|^2$ is close to $n^{-1}\sum_{i=1}^n \|Z_i\|^2 \approx \mathbb{E}\|Z_1\|^2$, then the GRD empirical distribution approximation error provides an upper bound of the probability matrix estimation error, up to a multiplicative factor.

Our requirement of polynomially decaying eigenvalues (first part of Assumption A2) implies the tail sum condition of eigenvalue sequence in [51]. In addition, we require a lower bound of the eigenvalue gap (second part of Assumption A2), which is not required in [51]. This is because in GRD estimation we need to recover the eigenvectors, and the use of subspace perturbation theory (Davis–Kahan $\sin\Theta$ theorem) involves a multiplicative factor of the inverse of eigenvalue gap near the threshold. This multiplicative factor of inverse eigenvalue gap leads to a slower convergence rate. When $\lambda_j = cj^{-\alpha}$ for some $\alpha > 1$, Assumption A2 holds with $\beta = \alpha + 1$. In the moderately sparse case, $n\rho_n$ is only a polynomial of $\log n$, then Theorem 4.5 implies a GRD empirical distribution estimation error rate of $(n\rho_n)^{-\frac{a-1}{2(a+1)}}$, while the probability matrix estimation error rate in [51] is $(n\rho_n)^{-\frac{2a-1}{2a}}$. In Section 5.2, we will empirically observe that when the underlying graphon is low rank (e.g., a stochastic block model with a small number of blocks), then the GRD empirical distribution approximation error and probability matrix estimation error roughly differ by a multiplicative factor; and when the underlying graphon is high rank, such as a graphon whose eigenvalues decay polynomially, then the GRD empirical distribution approximation error exhibits a slower rate of convergence than the corresponding probability matrix estimation error.

4.5. *Choice of embedding dimensions.* In practice, the value of p can affect the quality of the estimated GRD. If p is too small, the estimate may not have sufficient dimensionality to carry all useful structures in the GRD. If p is too large, the estimation becomes less stable and there would be a waste on computing and storage resources. Moreover, in many applications it may make sense to use different values of p_1 and p_2, since the effective dimensionality can be different for the positive and negative components as seen in the numerical examples in Section 5.

One potential way of choosing (p_1, p_2) is to follow a common practice in functional data analysis, where one chooses the leading principal subspace that explain a certain fraction (such as 90%) of the total variance. In network data, this approach has limited success due to the low-rank and high-noise nature of the adjacency matrix. Real-world network data often have low rank structures, but are observed with an entrywise Bernoulli noise.

Another way to choose p_1, p_2 is singular value thresholding [12], where p_1 and p_2 are the number of positive and negative eigenvalues whose absolute value exceeds a threshold, respectively. In particular, [12] suggests the threshold $2.01\sqrt{n\bar{\sigma}^2}$, where $\bar{\sigma}^2 \geq \max_{i,j} \mathrm{Var}(A_{ij})$ is an upper bound of the maximum entrywise variance of the adjacency matrix. When $\bar{\sigma}^2$ is unavailable, one may use the conservative bound $\mathrm{Var}(A_{ij}) \leq 1/4 \equiv \bar{\sigma}^2$, which results in the conservative singular value threshold $1.005\sqrt{n} \approx \sqrt{n}$. We find this simple rule of singular value thresholding working quite well for some real data sets.

5. Numerical examples.

5.1. *Simulation* 1: *Dense SBM, DCBM and MMBM.* We apply the truncated weighted spectral embedding for SBM, DCBM and MMBM. For comparison, we also apply the graphon estimation method based on stochastic block model approximation (SBA) [3].

The GRD estimator $\hat{F}$ and the graphon estimator $\hat{W}$ are in different spaces, where $\hat{F}$ is a probability measure on an Euclidean space and $\hat{W}$ is a symmetric function defined on the unit square. Thus they may reveal different underlying structures about the network data. We shall see that the GRD estimator is able to reveal the clustering and subspace clustering of the network nodes, while the graphon estimator is less visually informative without a correct ordering of the nodes. Estimation errors and convergence rates in more general settings, including sparse networks and infinite-dimensional GRDs, are considered in Section 5.2.

Following the notation in Section 2.2, we set $k = 3$ and B given in (3). The remaining parameters are set as follows:

- $\pi = (1/3, 1/3, 1/3)$ for the SBM and DCBM.
- $\Theta = \mathrm{Unif}(0.7, 1.4)$ for the DCBM, so the effects of node activeness parameter θ_i's range from halving to doubling the corresponding SBM edge probabilities.
- $a = (0.5, 0.5, 0.5)$ for the MMBM, so that the mixed memberships are not too close to the extreme points.

For each model, we generate a random graph with $n = 1000$ nodes, and apply the truncated weighted spectral embedding. The number of eigencomponents is determined by the singular value thresholding rule as described in Section 4.5 with threshold $\sqrt{n}$, which chooses top two absolute eigenvalues in all three cases, with one positive component and one negative component. The smooth graphon estimation method requires two independent realizations of the adjacency matrix on the same set of nodes. To make it a fair comparison, we generate two adjacency matrices of size 708×708 to use in the smooth graphon estimation algorithm, so that the number of independent observations is the same as a 1000×1000 adjacency matrix.

A typical output of GRD estimation and the SBA algorithm are visualized in Figure 1, Figure 2 and Figure 3 for the SBM, DCBM and MMBM, respectively. The SBA algorithm requires a tuning parameter δ for grouping similar nodes. Here, we use $\delta = 0.1$ since it achieves the smallest probability matrix estimation error. In each figure, the left plot shows the truncated and weighted spectral embedding of network nodes. The red dots, line segments and triangles are corresponding supports of the true GRD as theoretically predicted in Section 2.2. The embedded empirical distributions exhibit reasonable approximations to the underlying graph root distributions. On the other hand, the graphon estimation method SBA outputs an estimated probability matrix.

5.2. *Simulation* 2: *Sparse and infinite-dimensional graphons.* In this simulation study, we demonstrate GRD estimation in sparse and infinite-dimensional settings, and compare with the simulation results in corresponding graphon estimation using singular value thresholding (USVT, [12, 51]). We adopt two simulation settings in [51]: a stochastic block model with four communities and a smooth graphon.

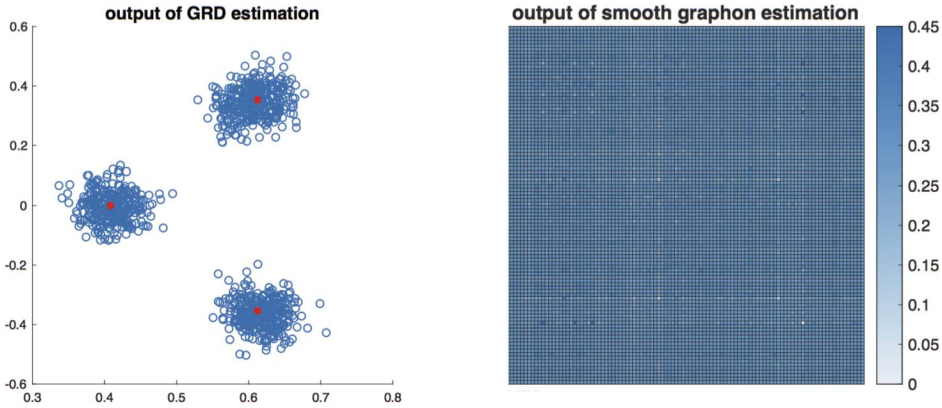

FIG. 1. *Simulation* 1, *SBM*. *Left*: *truncated and weighted spectral embedding output by the GRD estimation algorithm. The red dots are the point masses theoretically predicted in Section* 2.2. *Right*: *heatmap of estimated probability matrix output by the smooth graphon estimation algorithm with original output node ordering. The heatmap is shown at a lower resolution* (1 : 7) *for better visibility.*

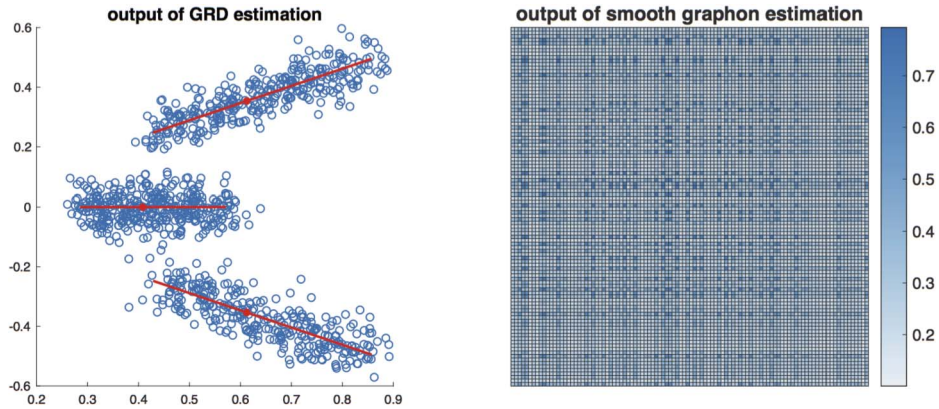

FIG. 2. *Simulation* 1, *DCBM*. *Left*: *truncated and weighted spectral embedding output by the GRD estimation algorithm. The red line segments are the subspace clusters theoretically predicted in Section* 2.2. *Right*: *heatmap of estimated probability matrix output by the smooth graphon estimation algorithm with original output node ordering. The heatmap is shown at a lower resolution* (1 : 7) *for better visibility.*

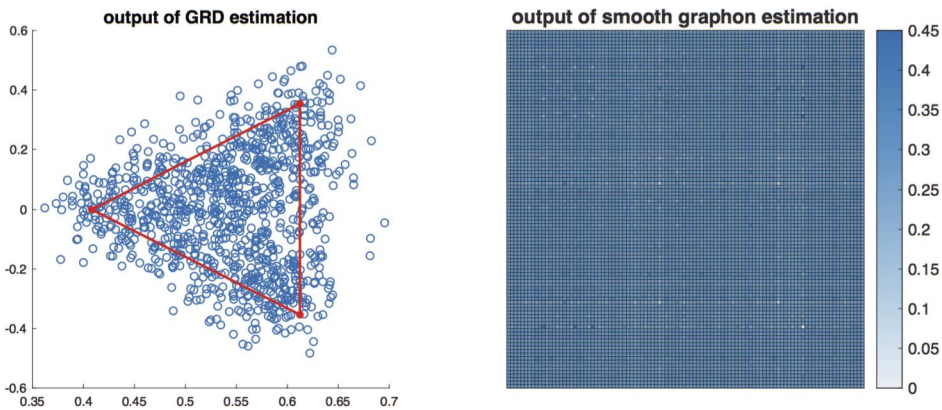

FIG. 3. *Simulation* 1, *MMBM*. *Left*: *truncated and weighted spectral embedding output by the GRD estimation algorithm. The red triangle is the convex polytope theoretically predicted in Section* 2.2. *Right*: *heatmap of estimated probability matrix output by the smooth graphon estimation algorithm with original output node ordering. The heatmap is shown at a lower resolution* (1 : 7) *for better visibility.*

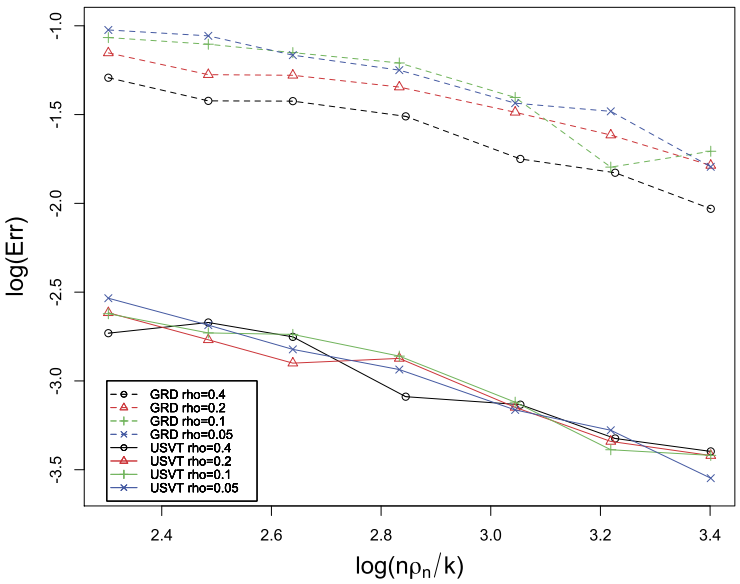

FIG. 4. *Simulation* 2: *SBM. Logarithm of estimation error as function of logarithm of signal strength in stochastic block model with* $k = 4$.

Stochastic block model. In the stochastic block model setting, we consider stochastic block models with $k = 4$ equal sized communities, and the B matrices have randomly generated entries from the uniform distribution on $[0, 1]$ subject to symmetry. We consider four different values of ρ: 0.4, 0.2, 0.1, 0.05 and six values of n such that $\log(n\rho/k)$ takes equally spaced values between 2.2 and 3.2. For each combination of (n, ρ), the simulation is repeated 30 times with independently generated B, community membership, and $\mathbf{A}_n$. The singular value thresholding algorithm for probability matrix estimation uses threshold $2.01\sqrt{n\rho}$. This is to make sure we can reproduce the results in [51]. For GRD estimation, we choose p_1 and p_2 by thresholding the absolute eigenvalues at $2.01\sqrt{n\rho(1-\rho)}$, following the suggestion in [12]. The results are summarized in Figure 4. The error metrics reported here are empirical GRD approximation error and the probability matrix estimation error as introduced in Section 4.4. The similar slopes between empirical GRD errors and probability matrix estimation errors seem to suggest that in this low-dimensional case, the two estimation errors roughly differ by a constant factor.

Smooth graphon. In the smooth graphon setting, we use $W(x, y) = \min(x, y)$, whose jth eigenvalue is $\frac{4}{\pi^2(2j-1)^2}$. We consider the same values of ρ: 0.4, 0.2, 0.1, 0.05. Given the small eigenvalues, we consider larger values of $\log(n\rho_n)$, which are equally spaced between 4 and 8.5. Due to the computer memory limit, we carry out the experiment when $n < 1.5 \times 10^4$. That is, for $\rho = 0.4$ the experiment covers n such that $\log(n\rho) \in [4, 8.5]$; for $\rho = 0.2$, it covers $\log(n\rho) \in [4, 8]$; for $\rho = 0.1$, it covers $\log(n\rho) \in [4, 7]$; for $\rho = 0.05$, it covers $\log(n\rho) \in [4, 6.5]$. The results are summarized in the left plot of Figure 5. The plot seems to confirm a slower rate of convergence for GRD estimation error.

The estimation errors of both the empirical GRD and probability matrix exhibits a sharp drop when $\log(n\rho) \approx 7$. To better understand this, we decompose the total estimation error into two parts:

1. The finite-dimensional estimation error: This is the error in approximating the low-rank component of the probability matrix $\mathbf{G}_n = \rho_n^{-1}\mathbb{E}\mathbf{A}_n$. For empirical GRD estimation, this corresponds to $n^{-1}\sum_{i=1}^n \|\rho^{-1/2}\hat{Z}_{i,A}^{(p_1,p_2)} - Z_i^{(p_1,p_2)}\|^2$, where $\hat{Z}_{i,A}^{(p_1,p_2)}$ is the estimated

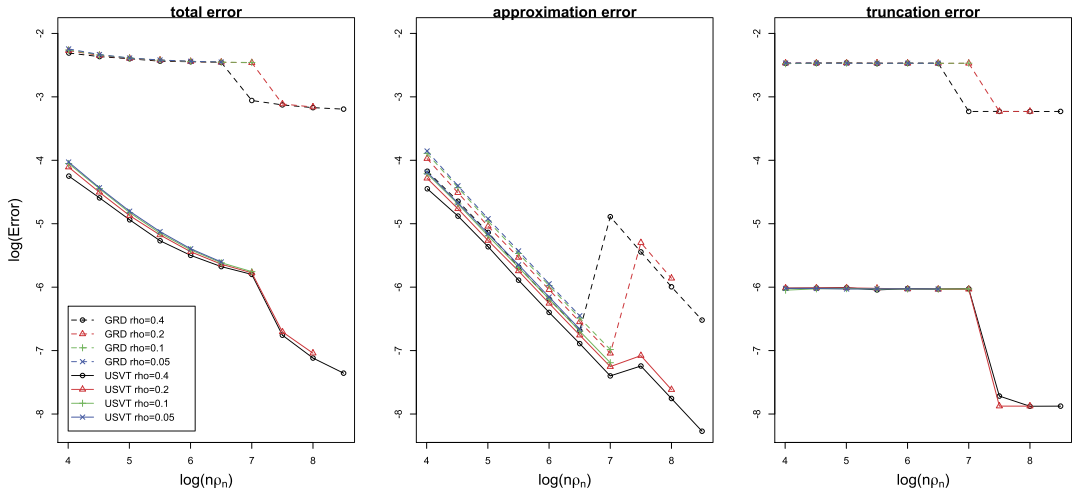

FIG. 5. *Simulation* 2, *smooth graphon*. *Logarithm of estimation error as a function of logarithm of signal strength in smooth graphon*. *Left*: *total estimation error*; *middle*: *finite-dimensional estimation error*; *right*: *truncation error.*

latent vector Z_i truncated to retain p_1 and p_2 coordinates in the positive and negative parts, respectively, as given in (10). For probability matrix estimation, this corresponds to $n^{-2}\|\hat{\mathbf{G}}_n - \mathbf{G}_n^{(p)}\|_F^2$, where $\mathbf{G}_n^{(p)}$ is the best rank-p approximation to $\mathbf{G}_n$ in Frobenius norm, and p is the number of singular values used in the USVT method.

2. The truncation error: This is the error incurred by ignoring the eigencomponents of $\mathbf{G}_n$ with smaller absolute eigenvalues. In empirical GRD estimation, this error is $n^{-1}\sum_{i=1}^n \|\mathbf{Z}_i - \mathbf{Z}_i^{(p_1,p_2)}\|^2$. In probability matrix estimation, this error is $n^{-2}\|\mathbf{G}_n - \mathbf{G}_n^{(p)}\|_F^2$.

The finite-dimensional error and truncation error are plotted in the middle and right plots in Figure 5, respectively. Near the point $\log(n\rho) = 7$, the signal becomes strong enough to pick up the second eigenvalue of the underlying probability matrix, therefore, the finite-dimensional approximation error increases for both methods, because there are more eigencomponents to estimate. After this increase, the finite-dimensional approximation errors start dropping again with a similar linear slope. The behavior of the truncation error matches the intuition, as it stays constant until a new eigencomponent is picked up when the signal strength increases. In this example, the truncation error is larger and decays more slowly for the empirical GRD estimation than for the probability matrix estimation.

5.3. *The political blogs data*. The political blogs data [1] is one of the most widely studied network data sets with a well-believed degree-corrected community structure [13, 25, 27, 34, 54]. The data set records undirected hyperlinks among 1222 political blogs during the 2004 presidential election, and the nodes have been manually classified as "liberal" and "conservative."

Among many statistical methods applied to this data set, spectral methods are quite popular and have used the top two singular vectors of the adjacency matrix. Here, we apply the truncated and weighted spectral embedding to this data set. The singular value thresholding rule suggests two significant eigencomponents, both of which correspond to the positive component. The embedded nodes in the two-dimensional Kreĭn space reflects a mixture of two components each on a one-dimensional subspace, with each mixture component corresponding to a labeled class. For this data set, we only have one realization of the adjacency matrix so the SBA algorithm is no longer applicable. For comparison, we apply the sorting-and-smoothing (SAS) estimator developed by [11], which adapts the SBA method by sorting the

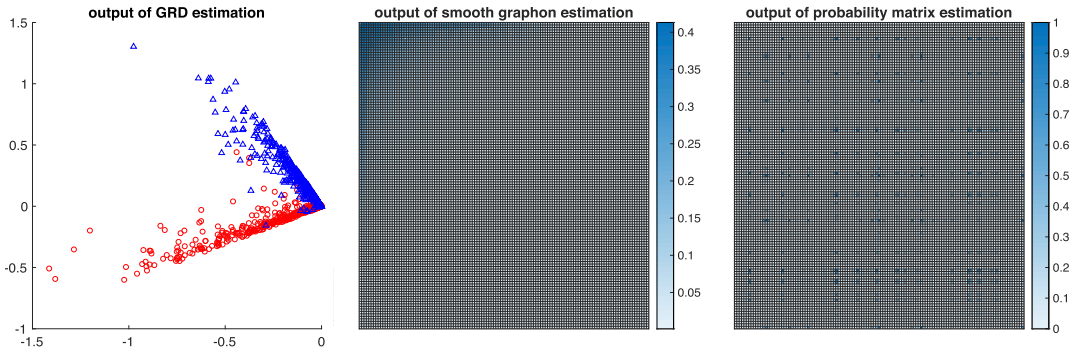

FIG. 6. *Political blogs data. Left*: *node embedding output by the GRD estimation method, colored by the ground truth manual labeling. Middle*: *estimated probability matrix with nodes sorted by the sorting-and-smoothing algorithm. Right*: *estimated probability matrix using USVT with random node ordering. The heatmaps are shown at a lower resolution* $(1:10)$ *for better visibility.*

nodes according to the degrees. We also apply the USVT method to estimate the probability matrix, with the singular value threshold $1.005\sqrt{n}$.

The results are visualized in Figure 6, in a similar fashion as in the simulated examples in Section 5.1. The GRD node embedding scatter plot is colored according to the ground truth of manual labeling of the blogs. It clearly shows that each group is represented by a one-dimensional subspace on which the GRD is supported. The SAS estimator sorts the nodes according to the degrees, and misses the subspace clustering hidden in the data. The USVT probability matrix estimation output is similar to that of SAS, with a different but random sorting of the nodes.

5.4. *The political books data.* The political books data records undirected links among 105 political books with links defined by the co-purchase records on Amazon.com. This data set, available on Mark Newman's website,[1] was collected by Krebs [31] during the 2004 presidential election. The nodes have been manually labeled as one of the three categories: "neutral," "liberal" and "conservative."

Given the three labeled classes, it seems natural to assume three significant eigencomponents. However, the singular value thresholding rule indicates only two significant components, both with positive eigenvalues. Again, we also apply the SAS algorithm and the USVT probability matrix estimator with the threshold $1.005\sqrt{n}$ to this data set.

As shown in Figure 7, the truncated weighted spectral embedding of the first two components (left plot of Figure 7) shows a two-component mixture with each component supported on a one-dimensional subspace, which strongly indicates a two-block DCBM. The "neutral" class, plotted as green square points, appears near the intersection of the other two classes. The SAS estimator and the USVT probability matrix estimator do not explicitly indicate such subspace clustering structure.

6. Discussion.

Kernel based learning. A side result of our theory is the relationship between the generating distribution of a random sample and the distribution of the corresponding kernel/gram matrix. Let F be a probability measure on a separable Hilbert space $\mathcal{X}$. Let $(X_i : i \geq 1)$ be a sequence of independent samples from F, and $\mathbf{G} = (\langle X_i, X_j \rangle, i, j \geq 1)$ be the (infinite size) gram matrix. The perspective of viewing $\mathbf{G}$ as an exchangeable random array allows us to establish

[1] http://www-personal.umich.edu/~mejn/netdata/

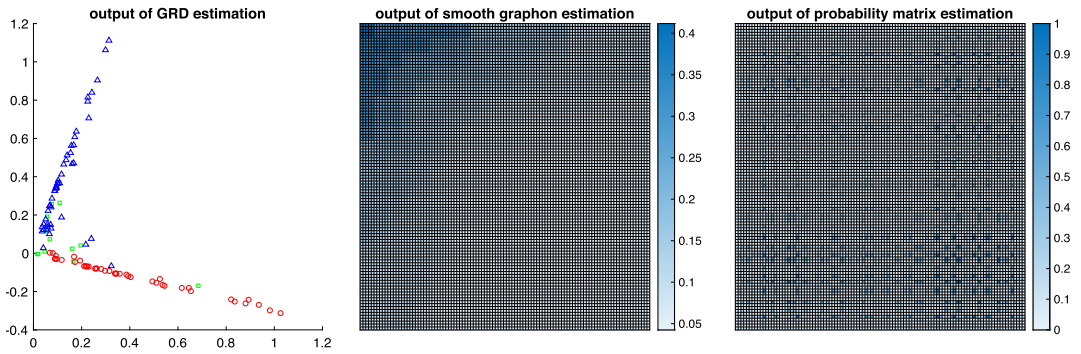

FIG. 7. *Political books data. Left: node embeddings output by the GRD estimation method, colored by the ground truth manual labeling. Middle: estimated probability matrix with nodes sorted by the sorting-and-smoothing algorithm. Right: estimated probability matrix using USVT with random node ordering.*

the correspondence between F and the distribution of $\mathbf{G}$. The result essentially says that the gram matrix carries all information about F up to an orthogonal transform. We believe that this result is elementary and highly intuitive, but are not able to find it in the literature.

COROLLARY 6.1. *Let F be a probability measure on a separable Hilbert space $\mathcal{X}$. Denote $\mathcal{G}_F$ the distribution of the corresponding infinite gram matrix $\mathbf{G}$. Then for two probability measures F_1, F_2 on $\mathcal{X}$, $\mathcal{G}_{F_1} = \mathcal{G}_{F_2}$ if and only if $F_1 \overset{o.t.}{=} F_2$, provided that one of the following holds*:

1. *$\mathcal{X}$ is finite dimensional*;
2. $\mathbb{E}_{X\sim F_1}\|X\|^2 < \infty$;
3. $\mathbb{E}_{X\sim F_2}\|X\|^2 < \infty$.

Here, the equivalence relation "$\overset{o.t.}{=}$" is defined as in Definition 4 by treating $\mathcal{H}_+ = \mathcal{X}$ and $\mathcal{H}_- = \varnothing$.

Modeling and inference for relational data. The framework of graph root representation can be extended in several interesting directions. First, one can model the connection probability with a logistic link function so that the two nodes i, j connect with probability $(1 + e^{-\langle Z_i, Z_j\rangle_{\mathcal{K}}})^{-1}$. With such a logistic transform, each distribution on $\mathcal{K}$ can be used to generate an exchangeable random graph and, therefore, can model a wider collection of structures. Moreover, one can also use the same framework to model relational data beyond binary observations. For example, one may observe event counting between a pair of nodes, such as number of email correspondences and frequency of research article citations. In applications such as multivariate time series and multimodal imaging, one may even observe a vector for each pair of nodes.

The graph root embedding also facilitates many subsequent inferences. We have discussed two examples in Section 2.3. There are other potential uses of GRD representations of networks. For example, in addition to clustering the embedded nodes as in SBM and DCBM, one can also test for specific structures of the graph root distribution, or compare the graph root distributions for multiple networks. See [49] for an example of two-sample comparison for random dot-product graphs. Another way to make use of the graph root embedding is to model the node movement in temporal networks. A challenge is to find the orthogonal transforms to match the embeddings at different time points. See [47] for an example using a similar idea with a different latent space model.

Acknowledgements. The author would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding. Jing Lei's research was partially supported by NSF Grants DMS-1553884 and DMS-2015492.

SUPPLEMENTARY MATERIAL

Supplement to "Network representation using graph root distributions" (DOI: 10.1214/20-AOS1976SUPP; .pdf). The supplementary file contains detailed explanations for the examples in Subsection 2.2 and proofs of the theoretical results in Sections 3 and 4.

REFERENCES

[1] ADAMIC, L. A. and GLANCE, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery* 36–43. ACM.

[2] AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

[3] AIROLDI, E. M., COSTA, T. B. and CHAN, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems* 692–700.

[4] ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598. MR0637937 https://doi.org/10.1016/0047-259X(81)90099-3

[5] ATHREYA, A., FISHKIND, D. E., TANG, M., PRIEBE, C. E., PARK, Y., VOGELSTEIN, J. T., LEVIN, K., LYZINSKI, V., QIN, Y. et al. (2017). Statistical inference on random dot product graphs: A survey. *J. Mach. Learn. Res.* **18** Paper No. 226, 92. MR3827114

[6] BHATTACHARYYA, S. and CHATTERJEE, S. (2018). Spectral clustering for multiple sparse networks: I. arXiv preprint arXiv:1805.10594.

[7] BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

[8] BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39** 2280–2301. MR2906868 https://doi.org/10.1214/11-AOS904

[9] BOLLOBÁS, B., JANSON, S. and RIORDAN, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures Algorithms* **31** 3–122. MR2337396 https://doi.org/10.1002/rsa.20168

[10] CARON, F. and FOX, E. B. (2017). Sparse graphs using exchangeable random measures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1295–1366. MR3731666 https://doi.org/10.1111/rssb.12233

[11] CHAN, S. and AIROLDI, E. (2014). A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning* 208–216.

[12] CHATTERJEE, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43** 177–214. MR3285604 https://doi.org/10.1214/14-AOS1272

[13] CHEN, K. and LEI, J. (2018). Network cross-validation for determining the number of communities in network data. *J. Amer. Statist. Assoc.* **113** 241–251. MR3803461 https://doi.org/10.1080/01621459.2016.1246365

[14] CHIN, P., RAO, A. and VU, V. (2015). Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory* 391–423.

[15] COJA-OGHLAN, A. (2010). Graph partitioning via adaptive spectral techniques. *Combin. Probab. Comput.* **19** 227–284. MR2593622 https://doi.org/10.1017/S0963548309990514

[16] CRANE, H. and DEMPSEY, W. (2018). Edge exchangeable models for interaction networks. *J. Amer. Statist. Assoc.* **113** 1311–1326. MR3862359 https://doi.org/10.1080/01621459.2017.1341413

[17] GAO, C., LU, Y. and ZHOU, H. H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652. MR3405606 https://doi.org/10.1214/15-AOS1354

[18] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2018). Community detection in degree-corrected block models. *Ann. Statist.* **46** 2153–2185. MR3845014 https://doi.org/10.1214/17-AOS1615

[19] GOHBERG, I. C. and KREĬN, M. G. (1988). *Introduction to the Theory of Linear Nonselfadjoint Operators*. **18**. Amer. Math. Soc., Providence, R.I. MR0246142

[20] GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.

[21] HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269 https://doi.org/10.1214/009053606000000957
[22] HOFF, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems* 657–664.
[23] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7
[24] HOOVER, D. N. (1982). Row-column exchangeability and a generalized model for probability. In *Exchangeability in Probability and Statistics* (*Rome*, 1981) 281–291. North-Holland, Amsterdam. MR0675982
[25] JIN, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57–89. MR3285600 https://doi.org/10.1214/14-AOS1265
[26] KALLENBERG, O. (1989). On the representation theorem for exchangeable arrays. *J. Multivariate Anal.* **30** 137–154. MR1003713 https://doi.org/10.1016/0047-259X(89)90092-4
[27] KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107, 10. MR2788206 https://doi.org/10.1103/PhysRevE.83.016107
[28] KLOPP, O., TSYBAKOV, A. B. and VERZELEN, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* **45** 316–354. MR3611494 https://doi.org/10.1214/16-AOS1454
[29] KLOPP, O. and VERZELEN, N. (2019). Optimal graphon estimation in cut distance. *Probab. Theory Related Fields* **174** 1033–1090. MR3980311 https://doi.org/10.1007/s00440-018-0878-1
[30] KOLACZYK, E. D. (2009). *Statistical Analysis of Network Data*: *Methods and Models*. *Springer Series in Statistics*. Springer, New York. MR2724362 https://doi.org/10.1007/978-0-387-88146-1
[31] KREBS, V. (2004). Social network analysis software & services for organizations, communities, and their consultants. http://www.orgnet.com/.
[32] LAX, P. D. (2002). *Functional Analysis*. *Pure and Applied Mathematics* (*New York*). Wiley Interscience, New York. MR1892228
[33] LEI, J. (2014). Adaptive global testing for functional linear models. *J. Amer. Statist. Assoc.* **109** 624–634. MR3223738 https://doi.org/10.1080/01621459.2013.856794
[34] LEI, J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist.* **44** 401–424. MR3449773 https://doi.org/10.1214/15-AOS1370
[35] LEI, J. (2020). Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli* **26** 767–798. MR4036051 https://doi.org/10.3150/19-BEJ1151
[36] LEI, J. (2021). Supplement to "Network representation using graph root distributions." https://doi.org/10.1214/20-AOS1976SUPP
[37] LOVÁSZ, L. (2012). *Large Networks and Graph Limits*. *American Mathematical Society Colloquium Publications* **60**. Amer. Math. Soc., Providence, RI. MR3012035 https://doi.org/10.1090/coll/060
[38] MCSHERRY, F. (2001). Spectral partitioning of random graphs. In 42*nd IEEE Symposium on Foundations of Computer Science* (*Las Vegas*, *NV*, 2001) 529–537. IEEE Computer Soc., Los Alamitos, CA. MR1948742
[39] MEISTER, A. (2011). Asymptotic equivalence of functional linear regression and a white noise inverse problem. *Ann. Statist.* **39** 1471–1495. MR2850209 https://doi.org/10.1214/10-AOS872
[40] NEWMAN, M. E. J. (2010). *Networks*: *An Introduction*. Oxford Univ. Press, Oxford. MR2676073 https://doi.org/10.1093/acprof:oso/9780199206650.001.0001
[41] NICKEL, C. L. M. (2007). *Random Dot Product Graphs*: *A Model for Social Networks*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–The Johns Hopkins University. MR2710154
[42] ONG, C. S., MARY, X., CANU, S. and SMOLA, A. J. (2004). Learning with non-positive kernels. In *Proceedings of the Twenty-First International Conference on Machine Learning* 81. ACM.
[43] ORBANZ, P. and ROY, D. M. (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 437–461.
[44] PENROSE, M. (2003). *Random Geometric Graphs*. *Oxford Studies in Probability* **5**. Oxford Univ. Press, Oxford. MR1986198 https://doi.org/10.1093/acprof:oso/9780198506263.001.0001
[45] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887
[46] RUBIN-DELANCHY, P., PRIEBE, C. E. and TANG, M. (2017). The generalised random dot product graph. arXiv preprint arXiv:1709.05506.
[47] SEWELL, D. K. and CHEN, Y. (2015). Latent space models for dynamic networks. *J. Amer. Statist. Assoc.* **110** 1646–1657. MR3449061 https://doi.org/10.1080/01621459.2014.988214
[48] SUSSMAN, D. L., TANG, M., FISHKIND, D. E. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* **107** 1119–1128. MR3010899 https://doi.org/10.1080/01621459.2012.699795

[49] TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V. and PRIEBE, C. E. (2017). A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli* **23** 1599–1630. MR3624872 https://doi.org/10.3150/15-BEJ789

[50] WOLFE, P. J. and OLHEDE, S. C. (2013). Nonparametric graphon estimation. arXiv preprint arXiv:1309.5936.

[51] XU, J. (2018). Rates of convergence of spectral methods for graphon estimation. In *Proceedings of the* 35*th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 5433–5442. PMLR, Stockholmsmässan, Stockholm Sweden.

[52] XU, J., MASSOULIÉ, L. and LELARGE, M. (2014). Edge label inference in generalized stochastic block models: From spectral theory to impossibility results. In *Conference on Learning Theory* 903–920.

[53] ZHANG, Y., LEVINA, E. and ZHU, J. (2017). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika* **104** 771–783. MR3737303 https://doi.org/10.1093/biomet/asx042

[54] ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083 https://doi.org/10.1214/12-AOS1036