

# Joint Embedding of Graphs

Shangsi Wang, Jesús Arroyo, Joshua T. Vogelstein, Carey E. Priebe

**Abstract**—Feature extraction and dimension reduction for networks is critical in a wide variety of domains. Efficiently and accurately learning features for multiple graphs has important applications in statistical inference on graphs. We propose a method to jointly embed multiple undirected graphs. Given a set of graphs, the joint embedding method identifies a linear subspace spanned by rank one symmetric matrices and projects adjacency matrices of graphs into this subspace. The projection coefficients can be treated as features of the graphs, while the embedding components can represent vertex features. We also propose a random graph model for multiple graphs that generalizes other classical models for graphs. We show through theory and numerical experiments that under the model, the joint embedding method produces estimates of parameters with small errors. Via simulation experiments, we demonstrate that the joint embedding method produces features which lead to state of the art performance in classifying graphs. Applying the joint embedding method to human brain graphs, we find it extracts interpretable features with good prediction accuracy in different tasks.

**Index Terms**—graphs, embedding, feature extraction, statistical inference



## 1 INTRODUCTION

IN many problems arising in science and engineering, graphs arise naturally as data structure to capture complex relationships between a set of objects. Graphs have been used in various application domains as diverse as social networks [1], internet mapping [2], brain connectomics [3], political voting networks [4], and many others. The graphs are naturally high dimensional objects with complicated topological structure, which makes graph clustering and classification a challenge to traditional machine learning algorithms. Therefore, feature extraction and dimension reduction techniques are helpful in the applications of learning graph data. In this paper, we propose an algorithm to jointly embed multiple graphs into low dimensional space. We demonstrate through theory and experiments that the joint embedding algorithm produces features which lead to state of the art performance for subsequent inference tasks on graphs.

There exist a few unsupervised approaches to extract features from graphs. First, classical Principal Component Analysis can be applied by treating each edge of a graph as a raw feature [5]. This approach produces features which are linear combinations of edges, but it ignores the topological structure of graphs and the features extracted are not easily interpretable. Second, features can be extracted by computing summary topological and label statistics from graphs [6], [7]. These statistics commonly include number

of edges, number of triangles, average clustering coefficient, maximum effective eccentricity, etc. In general, it is hard to know what intrinsic statistics to compute *a priori* and computing some statistics can be computationally expensive. Third, many frequent subgraph mining algorithms are developed [8]. For example, the fast frequent subgraph mining algorithm can identify all connected subgraphs that occur in a large fraction of graphs in a graph data set [9]. Finally, spectral feature selection can also be applied to graphs. It treats each graph as a node and constructs an object graph based on a similarity measure. Features are computed through the spectral decomposition of this object graph [10].

Adjacency Spectral Embedding (ASE) and Laplacian Eigenmap (LE) are proposed to embed a single graph observation [11], [12]. The inference task considered in these papers is learning of the block structure of the graph or clustering vertices. Given a set of graphs  $\{G_i = (V_i, E_i)\}_{i=1}^m$ , ASE and LE need to embed an adjacency matrix or Laplacian matrix of  $G_i$  individually, and there is no easy way to combine multiple embeddings. The joint embedding method considers the set of graphs together. It takes a matrix factorization approach to extract features for multiple graphs. The algorithm manages to simultaneously identify a set of rank one matrices and project adjacency matrices into the linear subspace spanned by this set of matrices. The joint embedding can be understood as a generalization of ASE for multiple graphs. We demonstrate through simulation experiments that the joint embedding algorithm extracts features which lead to good performance for a variety of inference tasks. In the next section, we review some random graph models and present a model for generating multiple random graphs. In Section 3, we define the joint embedding of graphs and present an algorithm to compute it. In Section 4, we perform some theoretical analyses of our joint embedding. The theoretical results and real data experiments are explored in Section 5. We conclude the paper with a brief discussion of implications and possible future work.

- Shangsi Wang and Carey Priebe are with the Department of Applied Mathematics and Statistics, Johns Hopkins University. E-mail: swang127@jhu.edu, cep@jhu.edu
- Jesús Arroyo is with the Center for Imaging Science, Johns Hopkins University. E-mail: jesus.arroyo@jhu.edu
- Joshua Vogelstein is with the Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University E-mail: jvov@jhu.edu
- Shangsi Wang and Jesús Arroyo equally contributed to the project.
- The authors gratefully acknowledge support from the GRAPHS program of the Defense Advanced Research Projects Agency (DARPA) contract N66001-14-1-4028 and the DARPA SIMPLEX program contract N66001-15-C-4041.

## 2 SETTING

We focus on embedding unweighted and undirected graphs for simplicity, although the joint embedding algorithm works on weighted graphs, and directed graphs with some modifications. Let  $\{G_i = (V_i, E_i)\}_{i=1}^m$  be  $m$  graphs, each with  $n$  vertices, and  $\mathbf{A}_i$  be the adjacency matrix of graph  $G_i$ . The vertices in these graphs should be matched, which means that all the graphs have a common vertex set  $V$ . The joint embedding algorithm embeds all  $G_i$ s simultaneously into  $\mathbb{R}^d$  and represents  $G_i$  by a vector  $\lambda_i \in \mathbb{R}^d$ . Before discussing the joint embedding algorithm, we need a random graph model on multiple graphs, on which the theoretical analysis is based. Let us first recall a model on a single graph: Random Dot Product Graph [13].

**Definition** Random Dot Product Graph (RDPG). Let  $F$  be a distribution on a set  $\mathcal{X} \in \mathbb{R}^d$  satisfying  $x^T y \in [0, 1]$  for all  $x, y \in \mathcal{X}$ . Let  $\mathbf{X} = [x_1^T, x_2^T, \dots, x_n^T] \in \mathcal{X}^n$ . The notation is  $(\mathbf{X}, \mathbf{A}) \sim \text{RDPG}(F)$ , if the  $x_i$  are independent and identically distributed according to  $F$ , and conditioned on  $\mathbf{X}$ , the  $\mathbf{A}_{st}$  are independent Bernoulli random variables,

$$\mathbf{A}_{st} \sim \text{Bernoulli}(x_s^T x_t).$$

Alternatively,

$$P(\mathbf{A}|\mathbf{X}) = \prod_{s < t} (x_s^T x_t)^{\mathbf{A}_{st}} (1 - x_s^T x_t)^{1 - \mathbf{A}_{st}}.$$

Also, define  $\mathbf{P} := \mathbf{X}\mathbf{X}^T$  to be edge probability matrix. When the latent positions  $\mathbf{X}$  is regarded as parameter, the notation becomes  $\mathbf{A} \sim \text{RDPG}(\mathbf{X})$ .

The RDPG is a convenient model which is designed to capture more complex structures than Stochastic Block Model (introduced below). The RDPG can be further generalized to Latent Position Graph by replacing the inner product by a kernel [14]. The Adjacency Spectral Embedding of RDPG adjacency matrix is well studied [15]. Next, we propose a new random graph model which generalizes the RDPG to multiple graphs.

**Definition** Multiple Random Eigen Graphs (MREG). Let  $\{h_k\}_{k=1}^d$  be a set of norm-1 vectors in  $\mathbb{R}^n$ , and  $F$  be a distribution on a set  $\mathcal{X} \in \mathbb{R}^d$ , satisfying  $\sum_{k=1}^d \lambda[k] h_k h_k^T \in [0, 1]^{n \times n}$  for all  $\lambda \in \mathcal{X}$ , where  $\lambda[k]$  is the  $k$ th entry of vector  $\lambda$ . The  $m$  pairs  $\{(\lambda_i, \mathbf{A}_i)\}_{i=1}^m$  follow a  $d$ -dimensional multiple random eigen graphs model, and the notation is

$$\{(\lambda_i, \mathbf{A}_i)\}_{i=1}^m \sim \text{MREG}(F, h_1, \dots, h_d),$$

if  $\{\lambda_i\}_{i=1}^m$  is independent and identically distributed according to distribution  $F$ , and conditioned on  $\lambda_i$ , the entries of  $\mathbf{A}_i$  are independent Bernoulli random variables,

$$\mathbf{A}_i[s, t] \sim \text{Bernoulli}\left(\sum_{k=1}^d \lambda_i[k] h_k[s] h_k[t]\right).$$

$\mathbf{P}_i := \sum_{k=1}^d \lambda_i[k] h_k h_k^T$  is defined to be the edge probability matrix for graph  $i$ . In cases that  $\{\lambda_i\}_{i=1}^m$  are of primary interest, they are treated as parameters, and it is said  $\{\mathbf{A}_i\}_{i=1}^m$  follows a  $m$ -graph  $d$ -dimensional multiple random eigen graphs model with the notation:

$$\{\mathbf{A}_i\}_{i=1}^m \sim \text{MREG}(\lambda_1, \dots, \lambda_m, h_1, \dots, h_d).$$

Compared to the RDPG model, MREG is designed to model multiple graphs. The vectors  $\{h_k\}_{k=1}^d$  are shared across graphs; a  $\lambda_i$  is sampled for each graph. On a single graph, they are equivalent if the edge probability matrix is positive semidefinite. In MREG, we allow self loops to happen. This is mainly for theoretical convenience. Next, we introduce another random graph model: Stochastic Block Model [16], which generalizes the Erdos-Renyi model [17]. SBM is a widely used model to study the community structure of a graph [18], [19].

**Definition** Stochastic Block Model (SBM). Let  $\pi$  be a prior probability vector for block membership which lies in the unit  $K - 1$ -simplex. Denote by  $\tau = (\tau_1, \tau_2, \dots, \tau_n) \in [K]^n$  the block membership vector, where  $\tau$  is a multinomial sequence with probability vector  $\pi$ . Denote by  $\mathbf{B} \in [0, 1]^{K \times K}$  the block connectivity probability matrix. Suppose  $\mathbf{A}$  is a random adjacency matrix given by,

$$P(\mathbf{A}|\tau, \mathbf{B}) = \prod_{i < j} \mathbf{B}_{\tau_i, \tau_j}^{\mathbf{A}_{ij}} (1 - \mathbf{B}_{\tau_i, \tau_j})^{(1 - \mathbf{A}_{ij})}$$

Then,  $\mathbf{A}$  is an adjacency matrix of a  $K$ -block stochastic block model graph, and the notation is  $\mathbf{A} \sim \text{SBM}(\pi, \mathbf{B})$ . Sometimes,  $\tau$  may also be treated as the parameter of interest, in this case the notation is  $\mathbf{A} \sim \text{SBM}(\tau, \mathbf{B})$ .

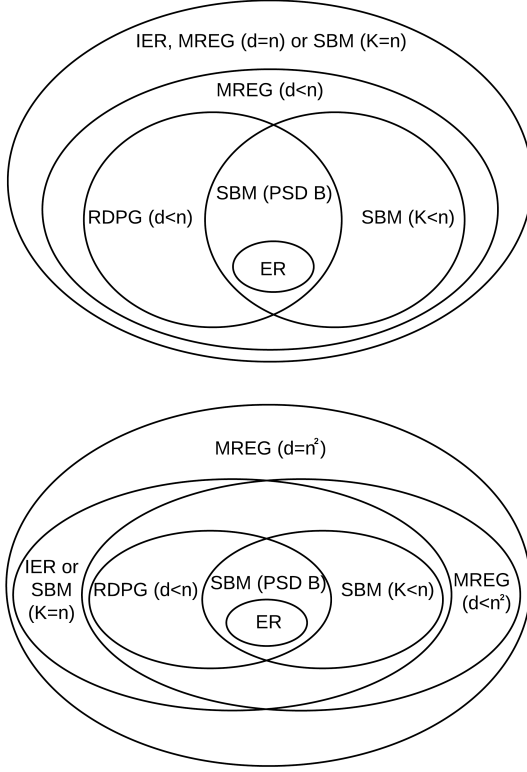
The left panel of Figure 1 shows the relationships between three random graph models defined above and the Erdos-Renyi (ER) model on 1 graph. The models considered are those conditioned on latent positions, that is  $\tau$ ,  $\mathbf{X}$  and  $\lambda$  in SBM, RDPG and MREG respectively are treated as parameters; furthermore, loops are ignored in MREG. If an adjacency matrix  $\mathbf{A} \sim \text{SBM}(\tau, \mathbf{B})$  and the block connectivity matrix  $\mathbf{B}$  is positive semidefinite,  $\mathbf{A}$  can also be written as an  $\text{RDPG}(\mathbf{X})$  with  $\mathbf{X}$  having at most  $K$  distinct rows. If an adjacency matrix  $\mathbf{A} \sim \text{RDPG}(\mathbf{X})$ , then it is also a 1-graph  $\text{MREG}(\lambda_1, h_1, \dots, h_d)$  with  $h_k$  being the normalized  $k$ th column of  $\mathbf{X}$  and  $\lambda_1$  being the vector containing the squared norms of columns of  $\mathbf{X}$ . However, a 1-graph  $\text{MREG}(\lambda_1, h_1, \dots, h_d)$  is not necessarily an RDPG graph since  $\lambda_1$  could contain negative entries which may result in an indefinite edge probability matrix.

The right panel of Figure 1 shows the relationships between the models on multiple graphs. For ER, SBM and RDPG, the graphs are sampled i.i.d. with the same parameters. MREG has the flexibility to have  $\lambda$  differ across graphs, which leads to a more generalized model for multiple graphs. Actually, it turns out that if  $d$  is allowed to be as large as  $\frac{n(n-1)}{2}$ , MREG can represent any distribution on binary graphs, which includes distributions in which edges are not independent.

**Theorem 2.1.** Given any distribution  $\mathcal{F}$  on graphs and a random adjacency matrix  $\mathbf{A} \sim \mathcal{F}$ , there exists a dimension  $d$ , a distribution  $F$  on  $\mathbb{R}^d$ , and a set of vectors  $\{h_k\}_{k=1}^d$ , such that  $\mathbf{A} \sim \text{MREG}(F, h_1, \dots, h_d)$ .

Theorem 2.1 implies that MREG is really a semi-parametric model, which can capture any distribution on graphs. One can model any set of graphs by MREG with the guarantee that the true distribution is in the model with  $d$  being

large enough. However, in practice, a smaller  $d$  may lead to better inference performance due to reduction in the dimensionality. In the next section, we consider the joint embedding algorithm which can be understood as a parameter estimation procedure for MREG.



**Figure 1:** Relationships between random graph models on 1 graph and multiple graphs. The top panel shows the relationships between the random graph models on 1 graph. The models considered are those conditioned on latent positions, that is  $\tau$ ,  $\mathbf{X}$  and  $\lambda$  in SBM, RDPG and MREG respectively are treated as parameters. ER is a 1-block SBM. If a graph follows SBM with a positive semidefinite edge probability matrix, it also follows the RDPG model. Any SBM and RDPG graph can be represented by a  $d$ -dimensional MREG model with  $d$  being less than or equal to the number of blocks or the dimension of RDPG. On one graph, inhomogeneous ER (IER),  $n$ -dimensional MREG and  $n$ -block SBM are equivalent. The bottom panel shows the relationships between the random graph models on multiple graphs. The models considered are those conditioned on latent positions, and for ER, SBM and RDPG graphs are sampled i.i.d. with the same parameters. In this case, MREG has the flexibility to have  $\lambda$  differ across graphs, which leads to a more generalized model for multiple graphs.

### 3 JOINT EMBEDDING

#### 3.1 Joint Embedding of Graphs

The joint embedding method considers a collection of vertex-aligned graphs, and estimates a common embedding space across all graphs and a loading for each graph.

Specifically, it simultaneously identifies a subspace spanned by a set of rank one symmetric matrices and projects each adjacency matrix  $\mathbf{A}_i$  into the subspace. The coefficients obtained by projecting  $\mathbf{A}_i$  are denoted by  $\hat{\lambda}_i \in \mathbb{R}^d$ , which is called the loading for graph  $i$ . To estimate rank one symmetric matrices and loadings for graphs, the algorithm minimizes the sum of squared Frobenius distances between adjacency matrices and their projections as described below.

**Definition Joint Embedding of Graphs (JE).** Given  $m$  graphs  $\{G_i\}_{i=1}^m$  with  $\mathbf{A}_i$  being the corresponding adjacency matrix, the  $d$ -dimensional joint embedding of graphs  $\{G_i\}_{i=1}^m$  is given by

$$(\hat{\lambda}_1, \dots, \hat{\lambda}_m, \hat{h}_1, \dots, \hat{h}_d) = \underset{\lambda_i, \|h_k\|=1}{\operatorname{argmin}} \sum_{i=1}^m \left\| \mathbf{A}_i - \sum_{k=1}^d \lambda_i[k] h_k h_k^T \right\|^2. \quad (1)$$

Here,  $\|\cdot\|$  denotes the Frobenius norm and  $\lambda_i[k]$  is the  $k$ th entry of vector  $\lambda_i$ .

To make sure that the model is identifiable and avoid the scaling factor,  $h_k$  is required to have norm 1. In addition,  $\{h_k h_k^T\}_{k=1}^d$  must be linearly independent to avoid identifiability issues in estimating  $\lambda_i$ ; however,  $\{h_k\}_{k=1}^d$  needs not to be linearly independent or orthogonal. To ease the notations, let us introduce two matrices  $\mathbf{\Lambda} \in \mathbb{R}^{m \times d}$  and  $\mathbf{H} \in \mathbb{R}^{n \times d}$ , where  $\lambda_i$  is the  $i$ th row of  $\mathbf{\Lambda}$  and  $h_k$  is the  $k$ th row of  $\mathbf{H}$ ; that is,  $\mathbf{\Lambda} = [\lambda_1^T, \dots, \lambda_m^T]$  and  $\mathbf{H} = [h_1, \dots, h_d]$ . The equation (1) can be rewritten using  $\mathbf{\Lambda}$  and  $\mathbf{H}$  as

$$(\hat{\mathbf{\Lambda}}, \hat{\mathbf{H}}) = \underset{\mathbf{\Lambda}, \|h_k\|=1}{\operatorname{argmin}} \sum_{i=1}^m \left\| \mathbf{A}_i - \sum_{k=1}^d \mathbf{\Lambda}_{ik} h_k h_k^T \right\|^2.$$

Denote the function on the left hand side of the equation by  $f(\mathbf{\Lambda}, \mathbf{H})$  which is explicitly a function of  $\lambda_i$ s and  $h_k$ s. There are several alternative ways to formulate the problem. If vector  $\lambda_i$  is converted into a diagonal matrix  $\mathbf{D}_i \in \mathbb{R}^{d \times d}$  by putting entries of  $\lambda_i$  on the diagonal of  $\mathbf{D}_i$ , then solving equation (1) is equivalent to solving

$$\underset{\mathbf{D}_i, \|h_k\|=1}{\operatorname{argmin}} \sum_{i=1}^m \left\| \mathbf{A}_i - \mathbf{H} \mathbf{D}_i \mathbf{H}^T \right\|^2$$

subject to  $\mathbf{D}_i$  being diagonal.

Equation (1) can be also viewed as a tensor factorization problem. If  $\{\mathbf{A}_i\}_{i=1}^m$  are stacked in a 3-D array  $\mathbb{A} \in \mathbb{R}^{m \times n \times n}$ , then solving equation (1) is also equivalent to

$$\underset{\mathbf{\Lambda}, \|h_k\|=1}{\operatorname{argmin}} \left\| \mathbb{A} - \sum_{k=1}^d \mathbf{\Lambda}_{*k} \otimes h_k \otimes h_k \right\|^2,$$

where  $\otimes$  denotes the tensor product and  $\mathbf{\Lambda}_{*k}$  is the  $k$ th column of  $\mathbf{\Lambda}$ . It is well known in the tensor factorization community that the solution to Equation 1 may not necessarily exist for  $d \geq 2$ . This phenomenon was first found by Bini *et al.* [20], and Silva and Lim gives a characterization all such tensors in the order-3 rank-2 case [21]. Although there may not exist a global minimum, finding the local solution in a compact region still provide significant insights to the data. We design an algorithm which is guaranteed to converge, and provide analysis under the  $d = 1$  case.

The joint embedding algorithm assumes the graphs are vertex-aligned and undirected. The vertex-aligned graphs are common in many applications of interest such as neuroimaging [22], multilayer networks [23] or time-varying graphs [7]. In case that the graphs are not aligned, graph matching should be performed before the joint embedding [24], [25]. The mis-alignments of some vertices will have adverse effects in estimating corresponding latent positions in  $\mathbf{H}$ ; however, a small number of mis-aligned vertices should not have a big impact in estimating  $\mathbf{\Lambda}$ . If the graphs have weighted edges, the joint embedding can still be applied. Also, the MREG model can be easily extended to weighted graphs by replacing the Bernoulli distribution with other proper distributions. In fact, in the experiment of section 5.3, the graphs are weighted, where the edge weights are the log of fiber counts across regions of brains. In case of directed graphs, to apply the joint embedding, one can symmetrize the graph by removing the direction of edges. Alternatively,  $h_k h_k^T$  in equation (1) can be replaced by  $h_k g_k^T$ , with  $h_k$  and  $g_k$  representing the in and out latent positions respectively. With this modification, equation (1) becomes the tensor factorization problem [26].

The optimization problem in equation (1) is similar to Principal Component Analysis (PCA) in the sense of minimizing squared reconstruction error to recover loadings and components [5]. However, there are extra symmetries and rank constraints on the components. Specifically, if  $h_k h_k^T$  is replaced by a matrix  $\mathbf{S}_k$  in equation (1)

$$(\hat{\lambda}_1, \dots, \hat{\lambda}_m, \hat{\mathbf{S}}_1, \dots, \hat{\mathbf{S}}_d) = \underset{\lambda_i, \mathbf{S}}{\operatorname{argmin}} \sum_{i=1}^m \left\| \mathbf{A}_i - \sum_{k=1}^d \lambda_i[k] \mathbf{S}_k \right\|^2, \quad (2)$$

the problem can be solved by applying PCA on vectorized adjacency matrices. In this case, there is a  $\mathbf{S}_k$  to estimate for each latent dimension which has  $\frac{n(n+1)}{2}$  parameters. Compared to PCA, the joint embedding estimates a rank one matrix  $h_k h_k^T$  for each latent dimension which has  $n$  parameters, and  $h_k$  can be treated as latent positions for vertices, but the joint embedding yields a larger approximation error due to the extra constraints. Similar optimization problems have also been considered in the simultaneous diagonalization literature [27], [28]. The difference is that the joint embedding is estimating an  $n$ -by- $d$  matrix  $\mathbf{H}$  by minimizing reconstruction error instead of finding a  $n$ -by- $n$  non-singular matrix by trying to simultaneously diagonalize all matrices. The problem in equation (1) has considerably fewer parameters to optimize, which makes it more stable and applicable with  $n$  being moderately large. In case of embedding only one graph, the joint embedding is equivalent to the Adjacency Spectral Embedding solved by singular value decomposition [11]. Next, we describe an algorithm to optimize the objective function  $f(\mathbf{\Lambda}, \mathbf{H})$ .

### 3.2 Alternating Descent Algorithm

The joint embedding of  $\{G_i\}_{i=1}^m$  is estimated by solving the optimization problem in equation (1). There are a few methods proposed to solve similar problems. Alternating least squares (ALS) is a popular method to solve similar problems [26], [29], but often ignore symmetric constraints. Gradient approaches have also been considered for similar

problems [30], [31]. We develop an alternating descent algorithm to minimize  $f(\mathbf{\Lambda}, \mathbf{H})$  that combines ideas from both approaches [32]. The algorithm can also be understood as a block coordinate descent method with  $\mathbf{\Lambda}$  and  $\mathbf{H}$  being the two blocks [33], [34]. The algorithm iteratively updates one of  $\mathbf{\Lambda}$  and  $\mathbf{H}$  while treating the other parameter as fixed. Optimizing  $\mathbf{\Lambda}$  when fixing  $\mathbf{H}$  is straightforward, since it is essentially a least squares problem. However, optimizing  $\mathbf{H}$  when fixing  $\mathbf{\Lambda}$  is hard due to the fact that the problem is non-convex and there is no closed form solution available. In this case, the joint embedding algorithm utilizes gradient information and takes an Armijo backtracking line search strategy to update  $\mathbf{H}$  [35].

Instead of optimizing all columns  $\mathbf{\Lambda}$  and  $\mathbf{H}$  simultaneously, we consider a greedy algorithm which solves the optimization problem by only considering one column of  $\mathbf{H}$  at a time. Specifically, the algorithm fixes all estimates for the first  $k_0 - 1$  columns of  $\mathbf{\Lambda}$  and  $\mathbf{H}$  at iteration  $k_0$ , and then the objective function is minimized by searching through only the  $k_0$ th column of  $\mathbf{\Lambda}$  and  $\mathbf{H}$ . That is,

$$(\hat{\mathbf{\Lambda}}_{*k_0}, \hat{h}_{k_0}) = \underset{\mathbf{\Lambda}_{*k_0}, \|h_{k_0}\|=1}{\operatorname{argmin}} \sum_{i=1}^m \left\| \mathbf{A}_i - \sum_{k=1}^{k_0-1} \hat{\mathbf{\Lambda}}_{ik} \hat{h}_k \hat{h}_k^T - \mathbf{\Lambda}_{ik_0} h_{k_0} h_{k_0}^T \right\|^2. \quad (3)$$

When  $m = 1$ , by the Eckart-Young theorem, the solution for  $(\hat{\mathbf{\Lambda}}_{1,k_0}, \hat{h}_{k_0})$  is given by the leading eigenvalue and eigenvector of  $\mathbf{A}_1 - \sum_{k=1}^{k_0-1} \hat{\mathbf{\Lambda}}_{1k} \hat{h}_k \hat{h}_k^T$ , and hence the solution obtained by our greedy approach coincides with the eigendecomposition of  $\mathbf{A}_1$ . If  $m > 1$  there is no closed-form solution in general, so we develop an optimization method next.

Let  $f(\mathbf{\Lambda}_{*k_0}, h_{k_0})$  denote the sum on the left hand side of the equation. To compute a  $d$ -dimensional joint embedding  $(\hat{\mathbf{\Lambda}}, \hat{\mathbf{H}})$ , the algorithm iteratively finds one component at a time by solving the optimization for the one dimensional embedding problem (3). Once a new component  $k_0$  is found, the matrix of loadings  $\hat{\mathbf{\Lambda}}$  is updated by minimizing the optimization problem (1) with the first  $k_0$  components fixed at  $\hat{h}_1, \dots, \hat{h}_{k_0}$ , resulting in the least squares problem

$$(\hat{\mathbf{\Lambda}}_{*1}, \dots, \hat{\mathbf{\Lambda}}_{*k_0}) = \underset{\mathbf{\Lambda} \in \mathbb{R}^{m \times k_0}}{\operatorname{argmin}} \sum_{i=1}^m \left\| \mathbf{A}_i - \sum_{k=1}^{k_0} \mathbf{\Lambda}_{ik} \hat{h}_k \hat{h}_k^T \right\|^2. \quad (4)$$

This step finds the projection of the graphs into the space spanned by the first  $k_0$  components  $\{\hat{h}_k \hat{h}_k^T\}_{k=1}^{k_0}$ . The algorithm solves the  $d$ -dimensional embedding problem by iteratively updating  $\hat{\mathbf{H}}$  and  $\hat{\mathbf{\Lambda}}$  using equations (3) and (4).

There are a few advantages in iteratively solving one dimensional embedding problems. First, there are fewer parameters to fit at each iteration, since the algorithm is only allowed to vary  $h_{k_0}$  at iteration  $k_0$ . This makes initialization and optimization steps much easier compared to optimizing all columns of  $\mathbf{H}$  simultaneously. Second, it implicitly enforces an ordering on the columns of  $\mathbf{H}$ . This ordering allows us to select the top few columns of  $\mathbf{\Lambda}$  and  $\mathbf{H}$  in cases where model selection is needed after the joint embedding. Third, it allows incremental computation. If  $d$

and  $d'$  dimensional joint embeddings are both computed, the first  $\min(d, d')$  columns of  $\hat{\mathbf{H}}$  will be the same. Fourth, the solution is guaranteed to exist when solving iteratively [21]. Finally, based on numerical experiments, the difference between optimizing iteratively and optimizing all the parameters when  $d$  is small is negligible; however, the iterative algorithm yields a slightly smaller objective function when  $d$  is large. The disadvantage of optimizing each column separately is that the algorithm is more likely to end up at a local minimum when the objective function is structured not in favor of embedding iteratively. In practice, this problem can be mitigated by running the joint embedding algorithm several times with random initializations.

To find  $\Lambda_{*k_0}$  and  $h_{k_0}$  in equation (3), the algorithm needs to evaluate two derivatives:  $\frac{\partial f}{\partial h_{k_0}}$  and  $\frac{\partial f}{\partial \Lambda_{ik_0}}$ . Denote by  $\mathbf{R}_{ik_0}$  the residual matrix after iteration  $k_0 - 1$  which is  $\mathbf{A}_i - \sum_{k=1}^{k_0-1} \hat{\Lambda}_{ik} \hat{h}_k \hat{h}_k^T$ . The gradient of the objective function with respect to  $h_{k_0}$  is given by

$$\frac{\partial f}{\partial h_{k_0}} = -4 \sum_{i=1}^m \Lambda_{ik_0} (\mathbf{R}_{ik_0} - \Lambda_{ik_0} h_{k_0} h_{k_0}^T) h_{k_0}. \quad (5)$$

The derivative of the objective function with respect to  $\Lambda_{ik_0}$  is given by

$$\frac{\partial f}{\partial \Lambda_{ik_0}} = -2 \langle \mathbf{R}_{ik_0} - \Lambda_{ik_0} h_{k_0} h_{k_0}^T, h_{k_0} h_{k_0}^T \rangle.$$

Setting the derivative to 0 yields

$$\hat{\Lambda}_{ik_0} = \langle \mathbf{R}_{ik_0}, h_{k_0} h_{k_0}^T \rangle, \quad (6)$$

where  $\langle \cdot, \cdot \rangle$  denotes the inner product.

Once a new component  $\hat{h}_{k_0} \hat{h}_{k_0}^T$  is obtained, the algorithm proceeds to update  $\Lambda$  by solving the optimization problem (4). Note that the problem can be split into  $m$  least square subproblems, one for each graph  $\mathbf{A}_i$ , of the form

$$(\hat{\Lambda}_{i1}, \dots, \hat{\Lambda}_{ik_0}) = \underset{\Lambda \in \mathbb{R}^{k_0}}{\operatorname{argmin}} \left\| \mathbf{A}_i - \sum_{k=1}^{k_0} \Lambda_k \hat{h}_k \hat{h}_k^T \right\|^2,$$

for each  $i = 1, \dots, m$ . By obtaining the derivative with respect to  $\Lambda$  and setting it equal to zero, we get the solution of the optimization problem by solving a system of linear equations. Let  $\mathbf{\Gamma}$  be a  $k_0 \times k_0$  matrix and  $\mathbf{\Psi}$  be a  $k_0 \times m$  matrix such that  $\mathbf{\Gamma}_{kj} = (\hat{h}_j^T \hat{h}_k)^2$  and  $\mathbf{\Psi}_{ki} = (\hat{h}_k^T \mathbf{A}_i \hat{h}_k)$ . The algorithm updates  $\Lambda$  by solving

$$\mathbf{\Gamma}(\hat{\Lambda}_{*1}, \dots, \hat{\Lambda}_{*k_0})^T = \mathbf{\Psi}. \quad (7)$$

The joint embedding algorithm alternates between updating  $\hat{\Lambda}_{*k_0}$  and  $\hat{h}_{k_0}$  according to equation (5) and (6), after which the values of all the loadings  $(\hat{\Lambda}_{*1}, \dots, \hat{\Lambda}_{*k_0})$  are updated using equation (7). Algorithm 1 describes the general procedure to compute the  $d$ -dimensional joint embedding of graphs  $\{G_i\}_{i=1}^m$ . The algorithm outputs two matrices:  $\hat{\Lambda}$  and  $\hat{\mathbf{H}}$ . The rows of  $\hat{\Lambda}$  denoted by  $\{\hat{\lambda}_i\}_{i=1}^m$  can be treated as estimates of  $\{\lambda_i\}_{i=1}^m$  in MREG and features for graphs. Columns of  $\hat{\mathbf{H}}$  denoted by  $\{\hat{h}_k\}_{k=1}^d$  are estimates of  $\{h_k\}_{k=1}^d$ . If a new graph  $G$  is observed with adjacency matrix  $\mathbf{A}$ ,  $\mathbf{A}$  can be projected into the linear space spanned by  $\{\hat{h}_k \hat{h}_k^T\}_{k=1}^d$  to obtain features for the graph.

In case of  $\mathbf{A}_i$  being large, the updating equations (5) and (6) are not practical due to  $h_k h_k^T$  and  $\mathbf{R}_{ik_0}$  being large and dense. However, they can be rearranged to avoid explicit computation of  $h_k h_k^T$  and  $\mathbf{R}_{ik_0}$ . Equation (5) becomes

$$\begin{aligned} \frac{\partial f}{\partial h_{k_0}} &= -4 \sum_{i=1}^m \Lambda_{ik_0} \mathbf{A}_i h_{k_0} + 4 \sum_{i=1}^m \Lambda_{ik_0} \sum_{k=1}^{k_0-1} \Lambda_{ik} (h_k^T h_{k_0}) h_k \\ &\quad + 4 \sum_{i=1}^m \Lambda_{ik_0}^2 h_{k_0}. \end{aligned}$$

Similarly, equation (6) can be rewritten as

$$\hat{\Lambda}_{ik_0} = h_{k_0}^T \mathbf{A}_i h_{k_0} - \sum_{k=1}^{k_0-1} \Lambda_{ik} (h_k^T h_{k_0})^2.$$

Based on the rearranged equations, efficiently evaluating matrix vector product  $\mathbf{A}_i h_{k_0}$  is needed to calculate the derivatives. This can be completed for a variety of matrices, in particular, sparse matrices [36].

The Algorithm 1 is guaranteed to converge to a stationary point. Specifically, at the termination of iteration  $k_0$ ,  $\frac{\partial f}{\partial h_{k_0}} \approx 0$  and  $\frac{\partial f}{\partial \Lambda_{ik_0}} \approx 0$ . First,  $\frac{\partial f}{\partial \Lambda_{ik_0}} \approx 0$  is ensured due to exact updating by equation (6). Second notice that updating according to equation (5) and (6) always decreases the objective function. Due to the fact that  $h_{k_0}$  lies on the unit sphere and the objective is twice continuous differentiable,  $\frac{\partial f}{\partial h_{k_0}}$  is Lipschitz continuous. This along with Armijo backtracking line search guarantees a "sufficient" decrease  $c \|\frac{\partial f}{\partial h_{k_0}}\|^2$  in objective function each time when the algorithm updates  $h_{k_0}$  [35], where  $c$  is a constant independent of  $h_{k_0}$ . Since the objective function is bounded below by 0, this implies convergence of gradient, that is  $\frac{\partial f}{\partial h_{k_0}} \rightarrow 0$ .

The joint embedding of graphs model requires for identifiability that the basis of the embedding space  $\{h_k h_k^T\}_{k=1}^m$  is linearly independent. This condition also ensures that equation (4) has a unique solution. Although the optimization problem (3) does not directly enforce this constraint, the next theorem guarantees that the components obtained by our method are linearly independent. The proof is given in the appendix.

**Theorem 3.1.** Suppose that  $\hat{\Lambda} = (\hat{\Lambda}_{*1}, \dots, \hat{\Lambda}_{*k_0})$  and  $\hat{\mathbf{H}} = [\hat{h}_1, \dots, \hat{h}_{k_0}]$  are obtained by iteratively solving (3) and (4) for  $k_0$  components, and the first  $k_0 - 1$  components  $\{\hat{h}_k \hat{h}_k^T\}_{k=1}^{k_0-1}$  are linearly independent. Then  $\{\hat{h}_k \hat{h}_k^T\}_{k=1}^{k_0}$  are also linearly independent, or  $\hat{\Lambda}_{ik_0} = 0$  for all  $i = 1, \dots, m$ .

In general, the objective function may have multiple stationary points due to non-convexity. Therefore, the joint embedding algorithm is sensitive to initializations. Actually, like many of the problems in tensor factorization, finding the global minimum in joint embedding is NP-Hard [37]. When time permits, we recommend running the joint embedding several times including many random initializations. In Section 5.1, we study the effects of different initialization approaches through a numerical simulation experiment. For other simulation and real experiments, we initialize  $\hat{h}_{k_0}$  using the top left singular vector of the average residual matrix  $\sum \mathbf{R}_{ik_0}/m$ . Computing this vector is computationally cheap compared to the computational cost of the joint embedding

when the number of graphs  $m$  is larger than 1. When  $m = 1$ , the SVD initialization and the joint embedding solution coincide, and when  $m > 1$  the SVD initialization gives a good approximation to the solution if all the graphs are identically distributed. The optimization algorithm described above may not be the fastest approach to solving the problem; however, numerical optimization is not the focus of this paper. Based on results from numerical applications, our approach works well in estimating parameters and extracting features for subsequent statistical inference. Next, we discuss some variations of the joint embedding algorithm.

---

**Algorithm 1** Joint Embedding

---

```

1: procedure FIND JOINT EMBEDDING  $\hat{\mathbf{A}}, \hat{\mathbf{H}}$  OF  $\{\mathbf{A}_i\}_{i=1}^m$ 
2:   Set residuals:  $\mathbf{R}_{i1} = \mathbf{A}_i$ 
3:   for  $k = 1 : d$  do
4:     Initialize  $h_k$  and  $\mathbf{\Lambda}_{*k}$ 
5:     while not convergent do
6:       Fixing  $\mathbf{\Lambda}_{*k}$ , update  $h_k$  by gradient descent (5)
7:       Project  $h_k$  back to the unit sphere
8:       Fixing  $h_k$ , update  $\mathbf{\Lambda}_{*k}$  by (6)
9:       Compute objective  $\sum_{i=1}^m \|\mathbf{R}_{ik} - \mathbf{\Lambda}_{ik} h_k h_k^T\|^2$ 
10:    end while
11:    Update  $\mathbf{\Lambda}$  by solving (7)
12:    Update residuals:  $\mathbf{R}_{i(k+1)} = \mathbf{A}_i - \sum_{j=1}^k \mathbf{\Lambda}_{ij} h_j h_j^T$ 
13:  end for
14:  Output  $\hat{\mathbf{A}} = [\mathbf{\Lambda}_{*1}, \dots, \mathbf{\Lambda}_{*d}]$  and  $\hat{\mathbf{H}} = [h_1, \dots, h_d]$ 
15: end procedure

```

---

### 3.3 Variations

The joint embedding algorithm described in the previous section can be modified to accommodate several different settings.

**Variation 1.** When all graphs come from the same distribution, we can force estimated loadings  $\hat{\lambda}_i$  to be equal across all graphs. This is useful when the primary inference task is to extract features for vertices. Since all graphs share the same loadings, with slightly abusing notations, let  $\mathbf{\Lambda}$  be a vector in  $\mathbb{R}^d$  and the optimization problem becomes

$$(\hat{\mathbf{A}}, \hat{\mathbf{H}}) = \underset{\mathbf{\Lambda}, \|h_k\|=1}{\operatorname{argmin}} \sum_{i=1}^m \left\| \mathbf{A}_i - \sum_{k=1}^d \mathbf{\Lambda}_k h_k h_k^T \right\|^2,$$

which is equivalent to

$$(\hat{\mathbf{A}}, \hat{\mathbf{H}}) = \underset{\mathbf{\Lambda}, \|h_k\|=1}{\operatorname{argmin}} \left\| \frac{1}{m} \sum_{i=1}^m \mathbf{A}_i - \sum_{k=1}^d \mathbf{\Lambda}_k h_k h_k^T \right\|^2.$$

Therefore, the optimization problem can be solved exactly by finding the singular value decomposition of the average adjacency matrix  $\frac{1}{m} \sum_{i=1}^m \mathbf{A}_i$ .

**Variation 2.** When there is a discrete label  $y_i \in \mathbb{Y}$  associated with the graph  $G_i$  available, we may require all loadings  $\hat{\lambda}_i$  to be equal within class. Let  $\mathbf{\Lambda} \in \mathbb{R}^{|\mathbb{Y}| \times d}$ , the optimization problem becomes

$$(\hat{\mathbf{A}}, \hat{\mathbf{H}}) = \underset{\mathbf{\Lambda}, \|h_k\|=1}{\operatorname{argmin}} \sum_{i=1}^m \left\| \mathbf{A}_i - \sum_{k=1}^d \mathbf{\Lambda}_{y_i k} h_k h_k^T \right\|^2.$$

In this case, when updating  $\mathbf{\Lambda}$  as in equation (6), the algorithm should average  $\mathbf{\Lambda}_{y_k}$  within the same class, that is

$$\hat{\mathbf{\Lambda}}_{y_k} = \left( \sum_{i=1}^m \mathbb{I}\{y_i = y\} \langle \mathbf{R}_{ik}, h_{k0} h_{k0}^T \rangle \right) / \left( \sum_{i=1}^m \mathbb{I}\{y_i = y\} \right).$$

**Variation 3.** In some applications, we may require all  $\mathbf{\Lambda}_{ik}$  to be greater than 0, as in non-negative matrix factorization. One advantage of this constraint is that graph  $G_i$  may be automatically clustered based on the largest entry of  $\hat{\lambda}_i$ . In this case, the optimization problem is

$$(\hat{\mathbf{A}}, \hat{\mathbf{H}}) = \underset{\mathbf{\Lambda} \geq 0, \|h_k\|=1}{\operatorname{argmin}} \sum_{i=1}^m \left\| \mathbf{A}_i - \sum_{k=1}^d \mathbf{\Lambda}_{ik} h_k h_k^T \right\|^2.$$

To guarantee nonnegativity, the algorithm should use non-negative least squares in updating  $\mathbf{\Lambda}$  [38]. Furthermore, a constraint on the number of non-zero elements in  $i$ th row of  $\mathbf{\Lambda}$  can be added as in K-SVD [39], and a basis pursuit algorithm could be used to update  $\mathbf{\Lambda}$  [40], [41]. Next, we discuss some theoretical properties of the MREG model and joint embedding when treated as a parameter estimation procedure for the model.

## 4 THEORETICAL RESULTS

In this section, we consider a simple setting where graphs follow a 1-dimensional MREG model, that is  $\{(\lambda_i, \mathbf{A}_i)\}_{i=1}^m \sim MREG(F, h_1)$ . The 1-dimensional joint embedding is well defined in this case, that is  $\hat{\lambda}_i$  and  $\hat{h}_1$  defined in Equation 1 is guaranteed to exist. Under this MREG model, the joint embedding of graphs can be understood as estimators for parameters of the model. Specifically,  $\hat{\lambda}_i$  and  $\hat{h}_1$  are estimates of  $\lambda_i$  and  $h$ . We prove two theorems concerning the asymptotic behavior of estimator  $\hat{h}_1$  produced by joint embedding.

Let  $\hat{h}_1^m$  denote the estimates based on  $m$  graphs and define functions  $\rho$ ,  $D_m$  and  $D$  as below:

$$\rho(\mathbf{A}_i, h) = \|\mathbf{A}_i - \langle \mathbf{A}_i, h h^T \rangle h h^T\|^2,$$

$$D_m(h, h_1) = \frac{1}{m} \sum_{i=1}^m \rho(\mathbf{A}_i, h),$$

$$D(h, h_1) = \mathbb{E}(\rho(\mathbf{A}_i, h)).$$

One can understand  $D_m$  and  $D$  as sample and population approximation errors respectively. By equation (1),

$$\hat{h}_1^m = \underset{\|h\|=1}{\operatorname{argmin}} \underset{\lambda_i}{\operatorname{argmin}} \sum_{i=1}^m \|\mathbf{A}_i - \lambda_i h h^T\|.$$

By equation (6),

$$\langle \mathbf{A}_i, h h^T \rangle = \underset{\lambda_i}{\operatorname{argmin}} \sum_{i=1}^m \|\mathbf{A}_i - \lambda_i h h^T\|.$$

Therefore,

$$\hat{h}_1^m = \underset{\|h\|=1}{\operatorname{argmin}} D_m(h, h_1).$$

The first theorem states that  $\hat{h}_1^m$  converges almost surely to a global minimum of  $D(h, h_1)$ . Alternatively, the theorem

implies the sample minimizer converges to the population minimizer.

**Theorem 4.1.** The estimator  $\hat{h}_1^m$  converges almost surely to the set of global minimizers of  $D(h, h_1)$  as  $m$  goes to infinity. That is,

$$\mathbb{P}\left(\lim_{m \rightarrow \infty} \hat{h}_1^m \in \operatorname{argmin}_h D(h, h_1)\right) = 1.$$

Theorem 4.1 ensures that, in the limit,  $\hat{h}^m$  is arbitrarily close to parameters in the set of global minimizer of  $D(h, h_1)$ . However, the global minimizer is definitely not unique due to the symmetry up to sign flip of  $h$ , that is  $D(h, h_1) = D(-h, h_1)$  for any  $h$ . This problem can be addressed by forcing an orientation of  $\hat{h}_1^m$  or stating that the convergence is up to a sign flip. In this case, Theorem 4.1 does not apply. We are currently only certain that when all graphs are from the Erdos-Renyi random graph model, the global minimizer is unique up to a sign flip. The next theorem concerns the asymptotic bias of  $h'$ . It gives a bound on the difference between the population minimizer  $h'$  and the truth  $h_1$ .

**Theorem 4.2.** If  $h'$  is a minimizer of  $D(h, h_1)$ , then

$$\|h' - h_1\| \leq \frac{2\mathbb{E}(\lambda_i)}{\mathbb{E}(\lambda_i^2)(h_1^T h')^2}.$$

To see an application of Theorem 4.2, let us consider the case in which all graphs are Erdos-Renyi graphs with 100 vertices and edge probability of 0.5. Under this setting, Theorem 4.2 implies  $\|h' - h_1\| \in [0, 0.04] \cup [1.28, 1.52]$ . The second interval is disturbing. It is due to the fact that when  $h_1^T h'$  is small, the bound is useless. We provide some insights as to why the second interval is there and how we can get rid of it with additional assumptions. In the proof of Theorem 4.2, we show that the global optimizer  $h'$  satisfies

$$h' = \operatorname{argmax}_{\|h\|=1} \mathbb{E}(\langle \mathbf{A}_i, h h^T \rangle^2).$$

Taking a closer look at  $E(\langle \mathbf{A}_i, h h^T \rangle^2)$ ,

$$\begin{aligned} \mathbb{E}(\langle \mathbf{A}_i, h h^T \rangle^2) &= \mathbb{E}(\langle \mathbf{P}_i, h h^T \rangle^2) + \mathbb{E}(\langle \mathbf{A}_i - \mathbf{P}_i, h h^T \rangle^2) \\ &= \mathbb{E}(\lambda_i^2)(h_1^T h)^4 + \mathbb{E}((h^T(\mathbf{A}_i - \mathbf{P}_i)h)^2). \end{aligned}$$

Therefore,

$$h' = \operatorname{argmax}_{\|h\|=1} \mathbb{E}(\lambda_i^2)(h_1^T h)^4 + \mathbb{E}((h^T(\mathbf{A}_i - \mathbf{P}_i)h)^2).$$

We can see that  $\mathbb{E}(\lambda_i^2)(h_1^T h)^4$  is maximized when  $h = h_1$ ; however, the noise term  $\mathbb{E}((h^T(\mathbf{A}_i - \mathbf{P}_i)h)^2)$  is generally not maximized at  $h = h_1$ . If  $n$  is large, we can apply a concentration inequality to  $(h^T(\mathbf{A}_i - \mathbf{P}_i)h)^2$  and have an upper bound on  $\mathbb{E}((h^T(\mathbf{A}_i - \mathbf{P}_i)h)^2)$ . If we further assume  $A_i$  is not too sparse, that is  $\mathbb{E}(\lambda_i^2)$  grows with  $n$  fast enough, then the sum of these two terms is dominated by the first term. This provides a way to have a lower bound on  $h_1^T h'$ . We may then replace the denominator of the bound in Theorem 4.2 by the lower bound. In general, if  $n$  is small, the noise term may cause  $h'$  to differ from  $h_1$  by a significant amount. In this chapter, we focus on the case that  $n$  is fixed. The case that  $n$  goes to infinity for Random Dot Product Graph is considered in [42].

The two theorems above concern only the estimation of  $h_1$ , but not  $\lambda_i$ . Based on equation (6), the joint embedding estimates  $\lambda_i$  by

$$\hat{\lambda}_i^m = \langle \mathbf{A}_i, \hat{h}_1^m \hat{h}_1^{mT} \rangle.$$

When  $m$  goes to infinity, we can apply Theorem 4.1,

$$\hat{\lambda}_i^m = \langle \mathbf{A}_i, \hat{h}_1^m \hat{h}_1^{mT} \rangle \xrightarrow{a.s.} \langle \mathbf{A}_i, h' h'^T \rangle = h'^T \mathbf{A}_i h'.$$

Then, applying the bound on  $\|h' - h_1\|$  derived in Theorem 4.2 and utilizing the fact that  $h^T \mathbf{A}_i h$  is continuous in  $h$ , we can obtain an upper bound on  $|\hat{\lambda}_i^m - h_1^T \mathbf{A}_i h_1|$ . When  $\mathbf{A}_i$  is large,  $h_1^T \mathbf{A}_i h_1$  is concentrated around  $\lambda_i$  with high probability. As a consequence, with high probability  $|\hat{\lambda}_i^m - \lambda_i|$  is small. In the next section, we demonstrate properties and utilities of the joint embedding algorithm through experiments.

## 5 EXPERIMENTS

Before going into details of our experiments, we want to discuss how to select the dimensionality  $d$  of the joint embedding. Estimating  $d$  is an important model selection question that has been studied for years under various settings [43]. Model selection is not the focus of this paper, but we still face this problem in numerical experiments. In the simulation experiments of this section, we assume  $d$  is known to us and simply set the dimensionality estimate  $\hat{d}$  equal to  $d$ . In the real data experiment, we recommend two approaches to determine  $\hat{d}$ . Both approaches require first running the  $d'$ -dimensional joint embedding algorithm, where  $d'$  is sufficiently large. We then plot the objective function versus dimension, and determine  $\hat{d}$  to be where the objective starts to flatten out. Alternatively, we can plot  $\{\hat{\mathbf{A}}_{ik}\}_{i=1}^m$  for  $k = 1, \dots, d'$ , and select  $\hat{d}$  when the loadings start to look like noise with 0 mean. These two approaches should yield a similar dimensionality estimate of  $\hat{d}$ .

### 5.1 Simulation Experiment 1: Joint Embedding Under a Simple Model

In the first experiment, we present a simple numerical example to demonstrate some properties of the joint embedding procedure as the number of graphs grows. We repeatedly generate graphs with 20 vertices from 3-dimensional MREG, where  $\lambda_i[1] \sim \text{Uniform}(8, 16)$ ,  $\lambda_i[2] \sim \text{Uniform}(0, 2)$  and  $\lambda_i[3] \sim \text{Uniform}(0, 1)$ , with

$$h_1 = [1, 1, 1, \dots, 1]/\sqrt{20}$$

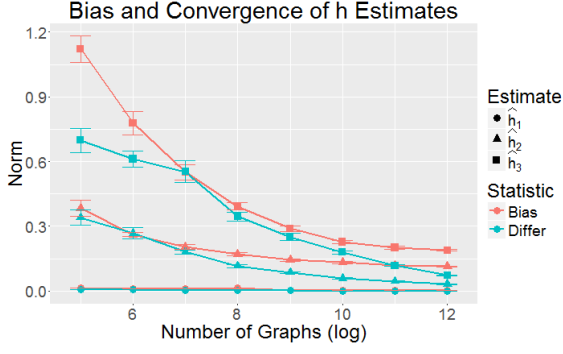
$$h_2 = [1, -1, 1, -1, 1, -1, \dots, -1]/\sqrt{20}$$

$$h_3 = [1, 1, -1, -1, 1, 1, -1, -1, \dots, -1]/\sqrt{20}.$$

We keep doubling the number of graphs  $m$  from  $2^4$  to  $2^{12}$ . At each value of  $m$ , we compute the 3-dimensional joint embedding of graphs. Let the estimated parameters based on  $m$  graphs be denoted by  $\hat{\lambda}_i^m$  and  $\hat{h}_k^m$ . Two quantities based on  $\hat{h}_k^m$  are calculated. The first is the norm difference between the current  $h_k$  estimates and the previous estimates, namely  $\|\hat{h}_k^m - \hat{h}_k^{m/2}\|$ . This provides numerical evidence for the convergence of our principled estimation procedure. The second quantity is  $\|\hat{h}_k^m - h_k\|$ . This investigates whether



$\hat{h}_k$  is an unbiased estimator for  $h_k$ . The procedure described above is repeated 20 times. Figure 2 presents the result.



**Figure 2:** Mean bias ( $\|\hat{h}_k^m - h_k\|$ ) and mean difference between estimates ( $\|\hat{h}_k^m - \hat{h}_k^{m/2}\|$ ) across 20 simulations are shown. The standard errors are also given by error bars. The graphs are generated from a 3-dimensional MREG model as described in section 5.1.  $\hat{h}_k^m$  has small asymptotic bias; however, it seems to converge as  $m$  increases.

Based on the plot, the norm of differences  $\|\hat{h}_k^m - \hat{h}_k^{m/2}\|$  seem to converge to 0 as  $m$  increases. This suggests the convergence of  $\hat{h}_1^m$ . Second, we notice that the bias  $\|\hat{h}_2^m - h_2\|$  and  $\|\hat{h}_3^m - h_3\|$  do not converge to 0; instead, it stops decreasing at around 0.1 and 0.2 respectively. This suggests that  $\hat{h}_k$  is an asymptotically biased estimator for  $h_k$ . Actually, this is as to be expected: when there are infinitely many nuisance parameters present, Neyman and Scott demonstrate that maximum likelihood estimator is inconsistent [44]. In our case, there are infinitely many  $\lambda_i$  as  $m$  grows; therefore, we do not expect the joint embedding to provide an asymptotic consistent estimate of  $h_k$ .

In applications such as clustering or classifying multiple graphs, we may be not interested in  $\hat{h}_k$ .  $\hat{\lambda}_i$  is of primary interest, which provides information specifically about the graphs  $G_i$ . Here, we consider two approaches to estimate  $\lambda_i[1]$ . The first approach is estimating  $\lambda_i[1]$  through joint embedding, that is

$$\hat{\lambda}_i[1] = \langle \mathbf{A}_i, \hat{h}_1^m \hat{h}_1^{mT} \rangle.$$

The second approach estimates  $\lambda_i$  by assuming  $h_1$  is known. In this case, equation (6) gives

$$\hat{\lambda}_i[1] = \langle \mathbf{A}_i, h_1 h_1^T \rangle.$$

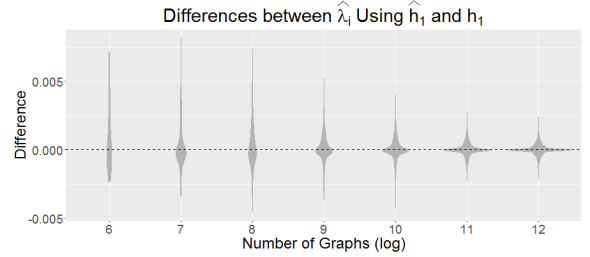
$\hat{\lambda}_i[1]$  calculated this way can be thought as the ‘oracle’ estimate. Figure 3 shows the differences in estimates provided by two approaches. Not surprisingly, the differences are small due to the fact that  $\hat{h}_1^m$  and  $h_1$  are close.

Next, we investigate the effects of four different initialization approaches. The first approach utilizes SVD on average residual matrix to initialize  $h_k$  at each iteration. The second approach randomly samples independent Gaussian variable for each entry of  $h_k$ . The third approach takes the best initialization among 10 random initializations. The

Initialization	Objective	Running time (sec.)
SVD	375.22(1.21)	8.3(1.0)
1 Random	383.29(1.60)	9.2(1.4)
Best of 10 Random	379.63(1.39)	96.5(5.3)
Truth	374.69(1.22)	7.8(1.0)

**TABLE 1:** Objective function and running time of four initialization approaches. The standard error is shown in parenthesis. SVD and truth initializations are significantly better than random initializations on this scenario.

fourth approach initializes  $h_k$  using the truth. To compare these approaches, we generate 16 graphs from the MREG model and jointly embed them with four different initializations. Then, another 16 graphs are generated and the objective function on these graphs are evaluated using  $\hat{\mathbf{H}}$  estimated by joint embedding. This procedure is repeated 100 times. Mean objective function and total running time with standard error of these four approaches are shown in Table 1. Based on Wilcoxon signed-rank test, SVD and truth initializations are significantly better than random initializations on this scenario. For the rest experiments, the initialization is completed by SVD.



**Figure 3:** Distribution of differences between  $\hat{\lambda}_i[1]$  estimated using  $\hat{h}_1^m$  and  $h_1$ . The graphs are generated from the 3-dimensional MREG model as described in section 5.1. The differences are small due to the fact that  $\hat{h}_1^m$  and  $h_1$  are close.

## 5.2 Simulation Experiment 2: Joint Embedding to Classify Graphs

In this experiment, we consider the inference task of classifying graphs. We have  $m$  pairs  $\{(\mathbf{A}_i, y_i)\}_{i=1}^m$  of observations. Each pair consists of an adjacency matrix  $\mathbf{A}_i \in \{0, 1\}^{n \times n}$  and a label  $y_i \in [K]$ . Furthermore, all pairs are assumed to be independent and identically distributed according to an unknown distribution  $\mathbb{F}_{\mathbf{A}, y}$ , that is

$$(\mathbf{A}_1, y_1), (\mathbf{A}_2, y_2), \dots, (\mathbf{A}_m, y_m) \stackrel{i.i.d.}{\sim} \mathbb{F}_{\mathbf{A}, y}.$$

The goal is to find a classifier  $g$  which is a function  $g : \{0, 1\}^{n \times n} \rightarrow [K]$  that has a small classification error  $L_g = P(g(\mathbf{A}) \neq y)$ .

We consider a binary classification problem where  $y$  takes value 1 or 2. 200 graphs with 100 vertices are independently generated. The graphs are sampled from a 2-dimensional MREG model. Let  $h_1$  and  $h_2$  be two vectors in  $\mathbb{R}^{100}$ , and

$$h_1 = [0.1, \dots, 0.1]^T, \text{ and } h_2 = [-0.1, \dots, -0.1, 0.1, \dots, 0.1]^T.$$



Here,  $h_2$  has  $-0.1$  as its first 50 entries and  $0.1$  as its last 50 entries. Graphs are generated according to the MREG model,

$$\{(\lambda_i, \mathbf{A}_i)\}_{i=1}^{200} \sim \text{MREG}(F, h_1, h_2), \quad (8)$$

where  $F$  is a mixture of two point masses with equal probability,

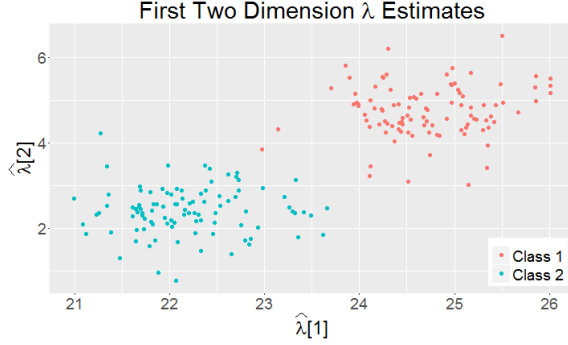
$$F = \frac{1}{2}\mathbb{I}\{\lambda = [25, 5]\} + \frac{1}{2}\mathbb{I}\{\lambda = [22.5, 2.5]\}.$$

We let the class label  $y_i$  indicate which point mass  $\lambda_i$  is sampled from. In terms of SBM, this graph generation scheme is equivalent to

$$A_i|y_i = 1 \sim \text{SBM}((1, \dots, 1, 2, \dots, 2), \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix})$$

$$A_i|y_i = 2 \sim \text{SBM}((1, \dots, 1, 2, \dots, 2), \begin{bmatrix} 0.25 & 0.2 \\ 0.2 & 0.25 \end{bmatrix}).$$

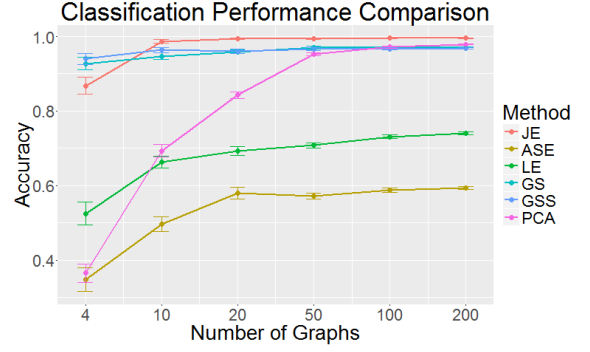
To classify graphs, we first jointly embed 200 graphs. The first two dimensional loadings are shown in Figure 4. We can see two classes are separated after being jointly embedded. Then, a 1-nearest neighbor classifier (1-NN) is



**Figure 4:** Scatter plot of loadings computed by jointly embedding 200 graphs. The graphs are generated from the 2-dimensional MREG model as described in equation (8). The loadings of two classes are separated after being jointly embedded.

constructed based on loadings  $\{\hat{\lambda}_i\}_{i=1}^m$ .

We compare classification performances of using the joint embedding to extract features to five other feature extraction approaches: Adjacency Spectral Embedding, Laplacian Eigenmap, Graph Statistics, Graph Spectral Statistics, and PCA. For Adjacency Spectral Embedding (ASE) and Laplacian Eigenmap (LE), we first embed each adjacency matrix or normalized Laplacian matrix and then compute the Procrustes distance between embeddings. For Graph Statistics (GS), we compute topological statistics of graphs considered by Park *et al.* in [7]. For Graph Spectral Statistics (GSS), we compute the eigenvalues of adjacency matrices and treat them as features [45]. For PCA, we vectorize the adjacency matrices and compute the factors through SVD. After the feature extraction step, we also apply a 1-NN rule to classify graphs. We let the number of graphs  $m$  increase from 4 to 200. For each value of  $m$ , we repeat the simulation 100 times. Figure 5 shows the result. The joint embedding takes advantage of increasing sample size and outperforms other approaches when given more than 10 graphs.



**Figure 5:** Mean classification accuracy of joint embedding, Adjacency Spectral Embedding, Laplacian Eigenmap, Graph Statistics, Graph Spectral Statistics, and PCA with their standard errors are shown. The graphs are generated from a 2-dimensional MREG model as described in the equation (8). The features are first extracted using methods described above; subsequently, we apply a 1-NN to classify graphs. For each value of  $m$ , the simulation is repeated 100 times. ASE, LE, GS and GSS do not take advantage of increasing sample size in the feature extraction step. PCA has poor performance when the sample size is small. Joint embedding takes advantage of increasing sample size and outperforms other approaches when given more than 10 graphs.

### 5.3 Real Data Experiment 1: Subject classification on HNU1 data

In this section, we use the HNU1 data [46] to classify graphs from different subjects. These data consist of diffusion tensor imaging (DTI) records of the brain of 30 healthy different subjects, each of which was scanned 10 times over a period of one month. Each scan was processed with the NeuroData's MRI to Graphs (NDMG) pipeline [47] using the Talairach atlas [48] to register the vertices, and we obtained a sample of 300 graphs (10 graphs per subject), each of which is composed of 1105 vertices.

It has been suggested that the information encoded in patterns of brain connectivity can uniquely identify different subjects [49], [50], and there is some evidence of low-rank structure in those differences [51], [52]. We study how low-dimensional representations can capture inter-individual variability by using the HNU1 data to classify subject scans. The joint embedding of graphs is an ideal method for this task, since it provides a low-rank representation of each of the graphs which is jointly learned from the sample.

The information encoded in the adjacency matrices can be used to accurately discriminate between different individuals. This can be confirmed by a 1-NN classifier using the vectorized adjacency matrices, which gives an almost perfect classification accuracy. However, our goal here is to study the accuracy of low-dimensional representations of the graphs to discriminate between subjects. Thus, we measure how does our joint dimensional embedding method perform as the number of embedding dimensions grows. We compare the performance of our method with PCA, as another dimensionality reduction method. Note that the joint embedding of graphs directly imposes a rank-

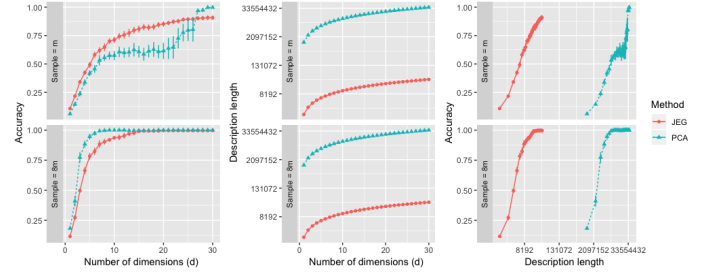
one restriction on the components, enforcing a low-rank representation of the graphs. PCA instead represents the adjacency matrices using components that are usually full rank, and requires many more parameters to represent does components.

Before computing the embedding, all graphs are centered by the mean, that is, we compute  $\hat{\mathbf{A}}_i = \mathbf{A}_i - \frac{1}{300} \sum_{j=1}^{300} \mathbf{A}_j$ . Using a 5-fold cross-validation, the data of each subject is divided into training and test sets. We measure the effect of the sample size using two different scenarios, the first uses 30 graphs on the training set (one scan per subject) and the second uses 240 graphs (eight scans per subject). The rest of the graphs are used in the test set to evaluate the accuracy, and the average over the 5 folds is computed. We follow the same procedure as in experiment 1. Using the training set, we compute an embedding of the graphs into  $d$  dimensions, either by doing JEG or PCA on the vectorized adjacency matrices. After that, all the data is projected into the  $d$ -dimensional embedding, and we use 1-NN to estimate the labels of the test data.

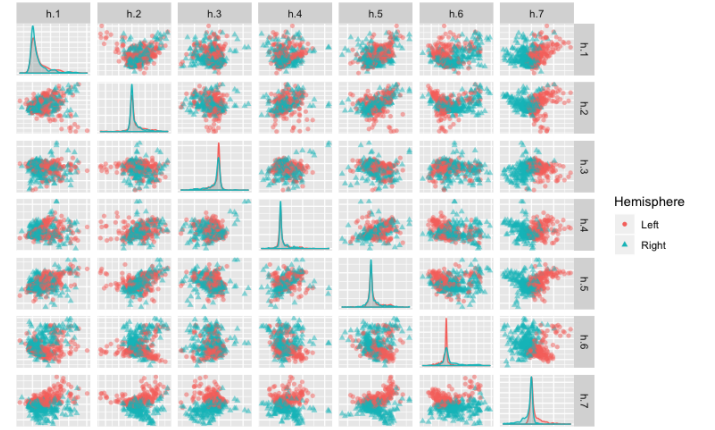
Figure 6 shows a comparison of the average classification accuracy and model complexity for both methods. We measure model complexity as the number of embedding dimensions  $d$ , and as the total number of parameters used in each instance ( $d(n+m)$  for JEG and  $d(n^2+m)$  for PCA). In general, both methods are able to perform very accurate classification when  $d$  is large enough, but JEG uses far fewer parameters in the components for the same number of dimensions (see middle plot). The advantage of JEG is especially remarkable when the sample size is small (first row). In this case, JEG shows a better performance than PCA when  $d$  is not large. For values close to 30, PCA shows almost perfect performance, which is not surprising, since the maximum number of principal components that can be obtained by PCA is  $m = 30$ , and thus PCA is not really performing any dimensionality reduction when  $d$  is close to this value, but rather projecting onto the training data itself. JEG on the other hand is able to provide very accurate classification error in all scenarios using low-rank representations of the graphs with a fewer number of parameters that are interpretable. These can be observed in Figure 7, which shows the latent positions of the vertices obtained by JEG for the first seven dimensions. To construct these plots, only the vertices that are labeled as left (507 of them) or right (525) were considered. As it can be observed, several dimensions of the latent positions show a structure that is clearly related to the hemisphere side. This structure is specially highlighted on the 7th dimension of the embedding.

#### 5.4 Real Data Experiment 2: Predict Composite Creativity Index

In this experiment, we study predicting individual composite creativity index (CCI) through brain connectomes obtained by Multimodal Magnetic Resonance Imaging [53]. Neuroimaging and creativity have been jointly investigated previously. Most studies utilize a statistical testing method and find CCI significantly related or inversely related to the activity of some regions of the brain. For a review, please see Arden *et al.* [54]. We embrace a different approach by directly building a prediction model for CCI. First, we



**Figure 6:** Comparison of average classification accuracy and model complexity (number of embedding dimensions and description length) for JEG and PCA on the HNU1 data. A  $d$  dimensional representation is estimated using a training sample (top row: one graph per subject, bottom row: 8 graphs per subject), after which all data is embedded, and the test data is classified using 1-NN. The representations obtained by JEG are more accurate with fewer model parameters than PCA, especially when the sample size is small.



**Figure 7:** Latent positions of the vertices found by JEG for the first seven dimensions using the HNU1 data. The color indicates the hemisphere side according to the Talairach atlas. The off-diagonal panels contain scatter plots of the vertices and diagonal plots show the density. Several dimensions of the embedding show a relationship between the latent positions and the hemisphere side.

jointly embed graphs of all subjects. Then, we construct a linear regression model by treating the estimated loadings as explanatory variables and CCI as the response variable.

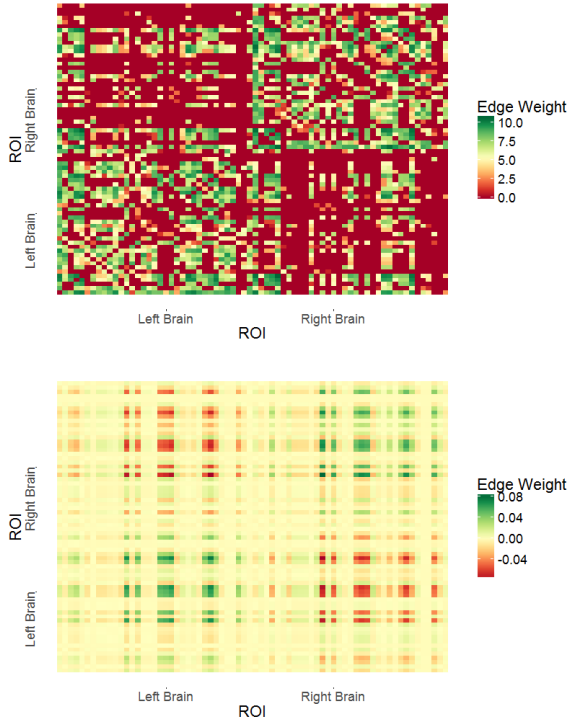
In total, 113 healthy, young adult subjects were scanned using a Siemens TrioTim scanner. 3D-MPRAGE and DTI in 35 directions of the subjects were acquired [55]. The images were then registered by Desikan-Killiany Atlas [56], and a graph of 70 vertices is constructed. The process of transforming MRI to graphs was completed by NDMG pipeline [47]. The graphs derived have weighted edges. One example of a graph is shown in the top panel of Figure 8. For each subject, a divergent thinking measure was scored by independent judges using the Consensual Assessment Technique [57], from which the CCI is derived.

To predict the CCI, we first jointly embed 113 graphs

with  $d = 10$ , and then fit a linear model by regressing CCI on  $\hat{\lambda}_i$ , that is

$$CCI_i \sim \beta_0 + \hat{\lambda}_i^T \beta + \epsilon_i.$$

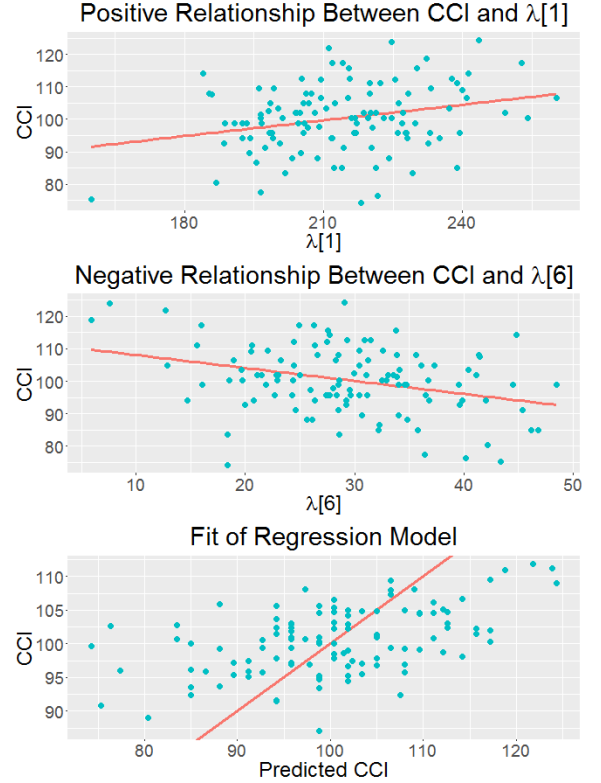
We consider two linear regression models. One using only the loadings obtained from the first component  $\hat{\lambda}_i[1]$  as the explanatory variable, and another one using the vector  $\hat{\lambda}_i$  as explanatory variables. If only the first dimensional loadings are used, the top panel of Figure 9 shows the result. There is a significant positive linear relationship between CCI and the first dimensional loadings. The first dimensional loadings generally capture the overall connectivity of graphs. In this case, the correlation between the first dimensional loadings and the sum of edge weights is around 0.98. This model suggests that the individual tends to be more creative when there is more brain connectivity. The R-square of this model is 0.07248, and the model is statistically significantly better when compared to the null model with a p-value 0.0039, according to the F-test.



**Figure 8:** The top panel shows the graph derived from a typical subject. There is much more neural connectivity within each hemisphere. The bottom panel shows the rank one matrix  $\hat{h}_6^T \hat{h}_6$ , which has positive connectivity within each hemisphere, but negative connectivity across hemispheres.

If CCI is regressed on the 10 dimensional loadings, a summary of the linear model is provided in the appendix and a scatter plot of fitted CCI versus true CCI is provided in the bottom panel of Figure 9. The R-square is 0.2325 and the model is statistically significantly better than the null model with a p-value 0.0018 according to the F-test. It is also significantly better than the model with only  $\hat{\lambda}_i[1]$ . Although there is still a positive relationship between CCI and the first dimensional loadings, it is no longer significant due to the

inclusion of more explanatory variables. In this model, there is a significant negative relationship between CCI and  $\hat{\lambda}_i[6]$  based on the t-test. The scatter plot of CCI against  $\hat{\lambda}_i[6]$  is given in the middle panel of Figure 9. We look into the rank one matrix  $\hat{h}_6^T \hat{h}_6$ , which is shown in the bottom panel of Figure 8. It has positive connectivity within each hemisphere of the brain, but negative connectivity across hemispheres. This suggests that compared to within-hemisphere connectivity, across-hemisphere connectivity tends to have a more positive impact on human creativity.

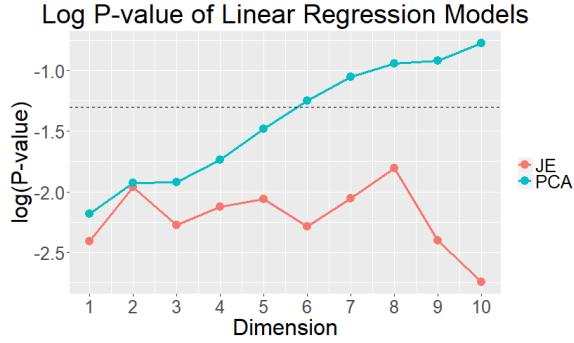


**Figure 9:** The top panel shows the scatter plot of CCI against  $\hat{\lambda}_i[1]$  with the regression line. There is a positive relationship between CCI and first dimensional loadings. The middle panel shows the scatter plot of CCI against  $\hat{\lambda}_i[6]$  with regression line. There is a negative relationship between CCI and sixth dimensional loadings. The bottom panel shows the predicted CCI versus true CCI with the identity line.

We also applied PCA to the vectorized adjacency matrices as described in (2), and then we regress CCI on PCA loadings. The log P-value of regression models are shown in Figure 10, with  $\log(0.05)$  represented by the dash line. When the number of dimensions is small, our joint embedding and PCA yield similar results; however, the performance of PCA degrades quickly as the number of dimensions increase. When regressed on 10 dimensional loadings, the p-values of joint embedding and PCA are 0.0018 and 0.16 respectively. The reason that joint embedding outperforms PCA in this setup is as explained before: the joint embedding only needs to estimate a rank one graph for each dimension, which has 70 parameters in this application; however, the PCA has



$2415 = \frac{70(70-1)}{2}$  parameters to fit for each dimension.



**Figure 10:** The log p-value of linear regression models using joint embedding and PCA are shown. The dash line is drawn at  $\log(0.05)$ . When the number of dimensions is small, the joint embedding and PCA yield similar results; however, the performance of PCA degrades quickly as the number of dimensions increase.

### 5.5 Real Data Experiment 3: Joint Embedding to Cluster Vertices

In the previous experiments, we focus on feature extraction for graphs. Here, we consider a different task, that is spectral clustering through the joint embedding. In general, spectral clustering first computes (generalized) eigenvalues and eigenvectors of adjacency matrix or Laplacian matrix, then clusters the latent positions into groups [11], [12]. The cluster identities of latent positions become the cluster identities of vertices of the original graph. Adjacency Spectral Embedding (ASE) is one of the spectral clustering algorithms used to find the latent positions of the vertices to fit the SBM and RDPG [58], which are special cases of our MREG model. When applied to one graph, the joint embedding is equivalent to Adjacency Spectral Embedding (ASE). When given multiple graphs, the joint embedding can estimate latent positions for graph  $i$  as  $[\hat{\lambda}_i[1]^{\frac{1}{2}} \hat{h}_1, \hat{\lambda}_i[2]^{\frac{1}{2}} \hat{h}_2, \dots, \hat{\lambda}_i[d]^{\frac{1}{2}} \hat{h}_d]$  or equivalently  $\hat{\mathbf{H}} \hat{\mathbf{D}}_i^{\frac{1}{2}}$ . Alternatively, the matrix of components  $\hat{H} = [\hat{h}_1, \dots, \hat{h}_d]$  gives joint latent positions for all the graphs. Then, a clustering algorithm can be applied to the latent positions to obtain communities.

We apply this spectral clustering approaches to two Wikipedia graphs [59]. The vertices of these graphs represent Wikipedia article pages. Two vertices are connected by an edge if either of the associated pages hyperlinks to the other. Two graphs are constructed based on English webpages and French webpages. The full graph has 1382 vertices which represents articles within 2-neighborhood of "Algebraic Geometry". Based on the content of the associated articles, they are grouped by hand into 6 categories: people, places, dates, things, math terms, and categories.

We consider a subset of vertices from 3 categories: people, things, and math terms. After taking the induced subgraph of these vertices and removing isolated vertices,

Method	ARI	Purity
ASE+EN	0.152 (0.02)	<b>0.555</b> (0.018)
JE+EN	<b>0.154</b> (0.019)	0.55 (0.019)
ASE+FR	0.114 (0.019)	0.521 (0.017)
JE+FR	<b>0.154</b> (0.019)	<b>0.549</b> (0.017)
ASE+(EN+FR)	0.155 (0.02)	0.547 (0.016)
JE+(EN, FR)	<b>0.156</b> (0.02)	<b>0.551</b> (0.021)

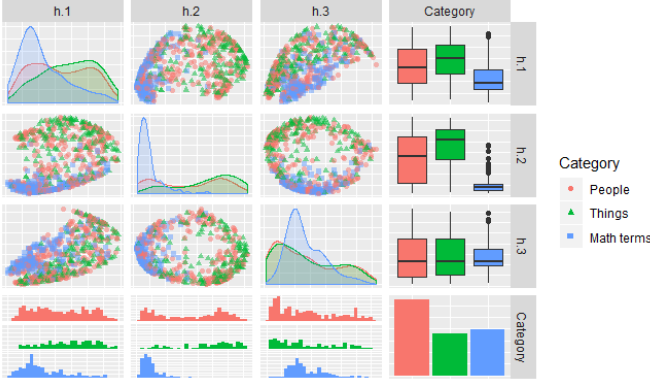
**TABLE 2:** Clustering Performance on Wikipedia Graphs. The adjusted rand index (ARI) and the purity of clustering of different spectral clustering approaches are shown. The first two scenarios consider each graph separately, and the third considers them jointly. The standard error (included in parenthesis) is estimated through repeatedly clustering bootstrapped latent positions. The joint embedding estimates latent positions which combine the information in both graphs with good clustering performance.

there are  $n = 704$  vertices left. Specifically, 326, 181, and 197 vertices are from people, things and math terms respectively. We consider approaches to embed the graphs to obtain latent positions. First, we consider clustering of each graph separately by doing ASE on the English graph  $\mathbf{A}_{en}$  (ASE+EN), or equivalently, JE on  $\mathbf{A}_{en}$ , and ASE on the French Graph  $\mathbf{A}_{fr}$  (ASE+FR), and compare with the individual latent positions obtained by JE  $\hat{\mathbf{H}} \hat{\mathbf{D}}_{en}^{\frac{1}{2}}$  (JE+EN) and  $\hat{\mathbf{H}} \hat{\mathbf{D}}_{fr}^{\frac{1}{2}}$  (JE+FR). We also consider methods to estimate joint latent positions, by doing ASE on the mean of both graphs  $\bar{\mathbf{A}} = (\mathbf{A}_{en} + \mathbf{A}_{fr})/2$  (ASE+(EN+FR)) [51], and the matrix  $\hat{\mathbf{H}}$  obtained by JE on both graphs (JE+(EN,FR)). The dimension  $d$  is set to 3 for all approaches, and the latent positions are scaled to have norm 1 for degree correction. Then, we apply 3-means algorithm to the latent positions.

The joint latent positions of the graphs  $\hat{\mathbf{H}}$  estimated by the joint embedding are provided on Figure 11. The latent positions of math terms are separated from the other two clusters. However, the latent positions of people and things are mixed. Table 2 shows the clustering results measured by adjusted rand index and the purity of clustering [60], [61]. The standard error is estimated through repeatedly clustering bootstrapped latent positions. All methods yield clustering results which are significantly better than random. The English graph demonstrates clearer community structure than the French graph. The joint embedding produces latent positions that combine the information in both graphs, and leads to better clustering performance. The JE and ASE give similar results on the English graph, but JE is able to improve the clustering performance on the French graph significantly. The JE also shows better performance than ASE on  $\bar{\mathbf{A}}$ , which also uses the information of both graphs. These results show that our method provide interpretable representations for the vertices, with an accuracy comparable to state of the art methods for spectral clustering.

## 6 CONCLUSION

In summary, we developed a joint embedding method that can simultaneously embed multiple graphs into low dimensional space. The joint embedding can be utilized to estimate features for inference problems on multiple vertex



**Figure 11:** The joint latent positions of  $\hat{\mathbf{H}}$  estimated by the JE are shown. The first three plots on the diagonal are density estimates of latent positions for each dimension and category, and the last plot shows the number of points from each category. The first three plots of the last row show latent positions histograms for each dimension and category, and the first three plots of the last column are the corresponding box plots. The pairs plots of latent positions are given in the off-diagonal panels. The latent positions of math terms are separated from the other two clusters.

matched graphs. Learning on multiple graphs has significant applications in diverse fields and our results have both theoretical and practical implications for the problem. As the real data experiment illustrates, the joint embedding is a practically viable inference procedure. We also proposed a Multiple Random Eigen Graphs model. It can be understood as a generalization of the Random Dot Product Graph model or the Stochastic Block Model for multiple random graphs. We analyzed the performance of joint embedding on this model under simple settings. We demonstrated that the joint embedding method provides estimates with bounded error. Our approach is intimately related to other matrix and tensor factorization approaches such as singular value decomposition and CP decomposition. Indeed, the joint embedding and these algorithms all try to estimate a low dimensional representation of high dimensional objects through minimizing a reconstruction error. We are currently investigating the utility of joint embedding with more or less regularizations on parameters and under different set ups. We are optimistic that our method provides a viable tool for analyzing multiple graphs and can contribute to a deeper understanding of the joint structure of networks.

## APPENDIX A

**Proof of Theorem 2.1** Denote the probability of observing a particular adjacency matrix  $\mathbf{A}_i$  under distribution  $\mathcal{F}$  by  $p_i$ . It suffices to show that there is a set of parameters for MREG such that observing  $\mathbf{A}_i$  under MREG is also  $p_i$ .

For undirected graphs with loops on  $n$  vertices, there are  $\binom{n+1}{2}$  possible edges. Let  $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_{\binom{n+1}{2}}$  be all the possible adjacency matrices. Since real symmetric matrix of size  $n$  has  $\binom{n+1}{2}$  free entries which lies in a linear space, if there exists  $\binom{n+1}{2}$  linearly independent rank one symmetric

matrices, they form a basis for this space. It turns out that the rank one symmetric matrices generated by vectors  $\{e_i\}_{i=1}^n \cup \{e_i + e_j\}_{i < j}$  are linearly independent, where  $\{e_i\}_{i=1}^n$  is the standard basis for  $n$ -dimensional Euclidean space.

Next, we construct parameters for the MREG. Let  $d$  be  $\binom{n+1}{2}$  and

$$\{h_k\}_{k=1}^d = \{e_i\}_{i=1}^n \cup \left\{ \frac{e_i + e_j}{\sqrt{2}} \right\}_{i < j}.$$

Since  $\{h_k h_k^T\}_{k=1}^d$  forms a basis for real symmetric matrices, for each adjacency matrix  $\mathbf{A}_i$ , there exists a vector  $\lambda_i$ , such that

$$\mathbf{A}_i = \sum_k \lambda_i[k] h_k h_k^T.$$

Let  $F$  be a finite mixture distribution on points  $\{\lambda_i\}_{i=1}^{\binom{n+1}{2}}$ , that is

$$F = \sum p_i \mathbb{I}\{\lambda = \lambda_i\}.$$

Under this MREG model, for any adjacency matrix  $\mathbf{A}_i$

$$P(\mathbf{A} = \mathbf{A}_i) = P(\lambda = \lambda_i) = p_i.$$

This concludes that the distribution  $\mathcal{F}$  and  $MREG(F, h_1, \dots, h_d)$  are equal.

**Proof of Theorem 3.1** Suppose that  $\hat{h}_{k_0} \hat{h}_{k_0}^T = \sum_{k=1}^{k_0-1} \alpha_k \hat{h}_k \hat{h}_k^T$ . Therefore, from equation (3) it can be noted that the solution for  $\hat{\Lambda}_{*k_0}$  is given by

$$\begin{aligned} \hat{\Lambda}_{ik_0} &= \operatorname{argmin}_{\Lambda_{ik_0} \in \mathbb{R}} \left\| \mathbf{A}_i - \sum_{k=1}^{k_0-1} \hat{\Lambda}_{ik} \hat{h}_k \hat{h}_k^T - \Lambda_{ik_0} h_{k_0} h_{k_0}^T \right\|^2. \\ &= \operatorname{argmin}_{\Lambda_{ik_0} \in \mathbb{R}} \left\| \mathbf{A}_i - \sum_{k=1}^{k_0-1} (\hat{\Lambda}_{ik} + \alpha_k \Lambda_{ik_0}) \hat{h}_k \hat{h}_k^T \right\|^2. \end{aligned}$$

Note that equation (4) implies that for any  $i = 1, \dots, m$ ,

$$\left\| \mathbf{A}_i - \sum_{k=1}^{k_0-1} (\hat{\Lambda}_{ik} + \alpha_k \Lambda_{ik_0}) \hat{h}_k \hat{h}_k^T \right\|^2 \geq \left\| \mathbf{A}_i - \sum_{k=1}^{k_0-1} \hat{\Lambda}_{ik} \hat{h}_k \hat{h}_k^T \right\|^2,$$

and hence,  $\hat{\Lambda}_{ik_0} \sum_{k=1}^{k_0-1} \alpha_k \hat{h}_k \hat{h}_k^T = 0$ , so  $\hat{\Lambda}_{ik_0} = 0$  for all  $i = 1, \dots, m$ .

**Proof of Theorem 4.1** First, we show that  $|D_n(h, h_1) - D(h, h_1)|$  converges uniformly to 0. To begin with, notice three facts:

- (1) the set  $\{h : \|h\| = 1\}$  is compact;
- (2) for all  $h$ , the function  $\rho(\cdot, h)$  is continuous
- (3) for all  $h$ , the function  $\rho(\cdot, h)$  is bounded by  $n^2$ .

Therefore, by the uniform law of large numbers [62], we have

$$\sup_h |D_m(h, h_1) - D(h, h_1)| \xrightarrow{a.s.} 0.$$

To prove the claim of the theorem, we use a technique similar to that employed by Bickel and Doksum [63]. Let  $h'$  be a global minimizer of  $D(h, h_1)$ . By definition, we must have

$D_m(\hat{h}_1^m, h_1) \leq D_m(h', h_1)$  and  $D(h', h_1) \leq D(\hat{h}_1^m, h_1)$ . From these two inequalities,

$$\begin{aligned} D_m(h', h_1) - D(h', h_1) &\geq D_m(\hat{h}_1^m, h_1) - D(h', h_1) \\ &\geq D_m(\hat{h}_1^m, h_1) - D(\hat{h}_1^m, h_1). \end{aligned}$$

Therefore,

$$|D_m(\hat{h}_1^m, h_1) - D(h', h_1)| \leq \max(|D_m(h', h_1) - D(h', h_1)|, |D_m(\hat{h}_1^m, h_1) - D(\hat{h}_1^m, h_1)|). \quad |h'^T E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i) h''| = |E(\langle \mathbf{A}_i, h'h'^T \rangle \langle \mathbf{A}_i, h''h''^T \rangle)| > E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i);$$

This implies

$$|D_m(\hat{h}_1^m, h_1) - D(h', h_1)| \leq \sup_h |D_m(h, h_1) - D(h, h_1)|.$$

Hence,  $|D_m(\hat{h}_1^m, h_1) - D(h', h_1)|$  must converge almost surely to 0, that is

$$|D_m(\hat{h}_1^m, h_1) - D(h', h_1)| \xrightarrow{a.s.} 0.$$

If  $\hat{h}_1^m$  does not converge almost surely to a global minimizer of  $D(h, h_1)$ , then for all  $h' \in \arg \min D(h, h_1)$ ,  $\|\hat{h}_1^m - h'\| \geq \epsilon$  for some  $\epsilon$  and infinitely many values of  $m$ . Since  $h'$  is a global minimum,  $|D(\hat{h}_1^m, h_1) - D(h', h_1)| > \epsilon'$  for infinitely many values of  $m$  and some  $\epsilon'$ . This contradicts with the previous equation. Therefore,  $\hat{h}_1^m$  must converge almost surely to  $\arg \min D(h', h_1)$ .

**Proof of Theorem 4.2** The proof of theorem relies on two lemmas. The first lemma shows that  $h'$  is the eigenvector corresponding to the largest eigenvalue of  $E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i)$ . The second lemma shows that  $E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i)$  is close to  $E(\lambda_i^2)(h_1^T h')^2 h_1 h_1^T$  under Frobenius norm. Then, we apply Davis-Kahan theorem [64] to establish the result of theorem.

**Lemma 6.1.** The vector  $h'$  is the eigenvector corresponding to the largest eigenvalue of  $E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i)$ .

We notice that

$$\begin{aligned} \min_{\|h\|=1} D(h, h_1) &= \min_{\|h\|=1} E(\|\mathbf{A}_i - \langle \mathbf{A}_i, hh^T \rangle hh^T\|^2) \\ &= \min_{\|h\|=1} E(\langle \mathbf{A}_i, \mathbf{A}_i \rangle - \langle \mathbf{A}_i, hh^T \rangle^2) \\ &= E(\langle \mathbf{A}_i, \mathbf{A}_i \rangle) - \max_{\|h\|=1} E(\langle \mathbf{A}_i, hh^T \rangle^2). \end{aligned}$$

Therefore,

$$h' = \operatorname{argmin}_{\|h\|=1} D(h, h_1) = \operatorname{argmax}_{\|h\|=1} E(\langle \mathbf{A}_i, hh^T \rangle^2). \quad (9)$$

Taking the derivative of  $E(\langle \mathbf{A}_i, hh^T \rangle^2) + c(h^T h - 1)$  with respect to  $h$ ,

$$\begin{aligned} \frac{\partial E(\langle \mathbf{A}_i, hh^T \rangle^2) + c(h^T h - 1)}{\partial h} &= E\left(\frac{\partial \langle \mathbf{A}_i, hh^T \rangle^2}{\partial h}\right) + 2ch \\ &= 4E(\langle \mathbf{A}_i, hh^T \rangle \mathbf{A}_i)h + 2ch. \end{aligned}$$

Setting this expression to 0 yields,

$$E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i)h' = -\frac{1}{2}ch'.$$

Using the fact that  $\|h'\| = 1$ , we can solve for  $c$ :

$$c = -2h'^T E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i)h' = -2E(\langle \mathbf{A}_i, h'h'^T \rangle^2).$$

Then, substituting for  $c$ ,

$$E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i)h' = E(\langle \mathbf{A}_i, h'h'^T \rangle^2)h'.$$

Therefore, we see that  $h'$  is an eigenvector of  $E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i)$  and the corresponding eigenvalue is  $E(\langle \mathbf{A}_i, h'h'^T \rangle^2)$ . Furthermore,  $E(\langle \mathbf{A}_i, h'h'^T \rangle^2)$  must be the eigenvalue with the largest magnitude. For if not, then there exists an  $h''$  with norm 1 such that

$$E(\langle \mathbf{A}_i, h''h''^T \rangle^2)E(\langle \mathbf{A}_i, h'h'^T \rangle^2) > |E(\langle \mathbf{A}_i, h'h'^T \rangle \langle \mathbf{A}_i, h''h''^T \rangle)|^2,$$

implying  $E(\langle \mathbf{A}_i, h''h''^T \rangle^2) > E(\langle \mathbf{A}_i, h'h'^T \rangle^2)$ , which contradicts equation (9) of this appendix. Thus, we conclude that  $h'$  is the eigenvector corresponding to the largest eigenvalue of  $E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i)$ .

**Lemma 6.2.**

$$\|E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i) - E(\lambda_i^2)(h_1^T h')^2 h_1 h_1^T\| \leq 2E(\lambda_i).$$

We compute  $E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i)$  by conditioning on  $P_i$ .

$$\begin{aligned} E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i | P_i) &= E(\langle \mathbf{A}_i - P_i, h'h'^T \rangle (\mathbf{A}_i - P_i) | P_i) + E(\langle \mathbf{A}_i - P_i, h'h'^T \rangle P_i | P_i) \\ &\quad + E(\langle P_i, h'h'^T \rangle (\mathbf{A}_i - P_i) | P_i) + E(\langle P_i, h'h'^T \rangle P_i | P_i) \\ &= E(\langle \mathbf{A}_i - P_i, h'h'^T \rangle (\mathbf{A}_i - P_i) | P_i) + \lambda_i (h_1^T h')^2 P_i \\ &= 2h'h'^T * P_i * (\mathbf{J} - P_i) - \text{DIAG}(h_1 h_1^T * P_i * (\mathbf{J} - P_i)) \\ &\quad + \lambda_i (h_1^T h')^2 P_i. \end{aligned}$$

Here,  $\text{DIAG}()$  means only keep the diagonal of the matrix;  $*$  means the Hadamard product, and  $\mathbf{J}$  is a matrix of all ones. Using the fact that  $P_i = \lambda_i h_1 h_1^T$ , we have

$$\begin{aligned} E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i) - E(\lambda_i^2)(h_1^T h')^2 h_1 h_1^T &= E(E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i | P_i) - \lambda_i (h_1^T h')^2 P_i) \\ &= E(2h'h'^T * P_i * (\mathbf{J} - P_i) - \text{DIAG}(h'h'^T * P_i * (\mathbf{J} - P_i))). \end{aligned}$$

If we consider the norm difference between  $E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i)$  and  $E(\lambda_i^2)(h_1^T h')^2 h_1 h_1^T$ , we have

$$\begin{aligned} \|E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i) - E(\lambda_i^2)(h_1^T h')^2 h_1 h_1^T\| &= \|E(2h'h'^T * P_i * (\mathbf{J} - P_i) - \text{DIAG}(h'h'^T * P_i * (\mathbf{J} - P_i)))\| \\ &\leq E(\|2h'h'^T * P_i * (\mathbf{J} - P_i) - \text{DIAG}(h'h'^T * P_i * (\mathbf{J} - P_i))\|) \\ &\leq E(\|2h'h'^T * P_i * (\mathbf{J} - P_i)\|) \\ &\leq E(\|2h'h'^T * P_i\|) \\ &\leq 2E(\lambda_i)\|h'h'^T * h_1 h_1^T\| \\ &= 2E(\lambda_i). \end{aligned}$$

This finishes the proof for the lemma.

Notice that the only non-zero eigenvector of  $E(\lambda_i^2)(h_1^T h')^2 h_1 h_1^T$  is  $h_1$  and the corresponding eigenvalue is  $E(\lambda_i^2)(h_1^T h')^2$ . We apply the Davis-Kahan theorem [64] to the eigenvector corresponding to the largest eigenvalue of matrices  $E(\langle \mathbf{A}_i, h'h'^T \rangle \mathbf{A}_i)$  and  $E(\lambda_i^2)(h_1^T h')^2 h_1 h_1^T$ , yielding

$$\|h' - h_1\| \leq \frac{2E(\lambda_i)}{E(\lambda_i^2)(h_1^T h')^2}.$$

## APPENDIX B

### Linear Regression Model Summary

```
> model <- lm(cci ~ Lambda+1)
> summary(model)
```

Call:

```
lm(formula = cci ~ Lambda + 1)
```

Residuals:

Min	1Q	Median	3Q	Max
-26.3432	-6.144	-0.7578	7.1032	16.9004

Coefficients:	Estimate	Pr(> t )	
(Intercept)	1.275e+02	0.000275	***
Lambda1	2.421e-04	0.997981	
Lambda2	-2.326e-01	0.070110	.
Lambda3	-3.716e-02	0.822592	
Lambda4	8.049e-02	0.687628	
Lambda5	-2.925e-01	0.421858	
Lambda6	-4.285e-01	0.009088	**
Lambda7	-1.745e-01	0.590533	
Lambda8	-3.465e-01	0.240093	
Lambda9	-8.970e-01	0.007999	**
Lambda10	-8.955e-01	0.052839	.

Signif. codes:	0	***	0.001	**
0.01	*	0.05	.	0.1
				1

Residual standard error: 9.437

on 102 degrees of freedom

Multiple R-squared: 0.2325,

Adjusted R-squared: 0.1572

F-statistic: 3.09 on 10 and 102 DF,

p-value: 0.001795

## REFERENCES

- [1] E. Otte and R. Rousseau, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of Information Science*, vol. 28, no. 6, pp. 441–453, 2002.
- [2] R. Govindan and H. Tangmunarunkit, "Heuristics for internet map discovery," in *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3, pp. 1371–1380, IEEE, 2000.
- [3] E. T. Bullmore and D. S. Bassett, "Brain graphs: graphical models of the human brain connectome," *Annual Review of clinical Psychology*, vol. 7, pp. 113–140, 2011.
- [4] M. D. Ward, K. Stovel, and A. Sacks, "Network analysis and political science," *Annual Review of Political Science*, vol. 14, pp. 245–264, 2011.
- [5] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.
- [6] G. Li, M. Semerci, B. Yener, and M. J. Zaki, "Graph classification via topological and label attributes," in *Proceedings of the 9th international workshop on mining and learning with graphs (MLG)*, San Diego, USA, vol. 2, 2011.
- [7] Y. Park, C. E. Priebe, and A. Youssef, "Anomaly detection in time series of graphs using fusion of graph invariants," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 67–75, 2013.
- [8] C. Jiang, F. Coenen, and M. Zito, "A survey of frequent subgraph mining algorithms," *The Knowledge Engineering Review*, vol. 28, no. 01, pp. 75–105, 2013.
- [9] J. Huan, W. Wang, and J. Prins, "Efficient mining of frequent subgraphs in the presence of isomorphism," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pp. 549–552, IEEE, 2003.
- [10] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24th International Conference on Machine Learning*, pp. 1151–1157, ACM, 2007.
- [11] D. L. Sussman, M. Tang, D. E. Fishkind, and C. E. Priebe, "A consistent adjacency spectral embedding for stochastic blockmodel graphs," *Journal of the American Statistical Association*, vol. 107, no. 499, pp. 1119–1128, 2012.
- [12] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.
- [13] S. J. Young and E. R. Scheinerman, "Random dot product graph models for social networks," in *International Workshop on Algorithms and Models for the Web-Graph*, pp. 138–149, Springer, 2007.
- [14] M. Tang, D. L. Sussman, and C. E. Priebe, "Universally consistent vertex classification for latent positions graphs," *The Annals of Statistics*, vol. 41, no. 3, pp. 1406–1430, 2013.
- [15] D. L. Sussman, M. Tang, and C. E. Priebe, "Consistent latent position estimation and vertex classification for random dot product graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 48–57, 2014.
- [16] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.
- [17] P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.
- [18] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, p. 016107, 2011.
- [19] V. Lyzinski, M. Tang, A. Athreya, Y. Park, and C. E. Priebe, "Community detection and classification in hierarchical stochastic blockmodels," *IEEE Transactions on Network Science and Engineering*, vol. 4, no. 1, pp. 13–26, 2017.
- [20] D. Bini, M. Capovani, F. Romani, and G. Lotti, "O (n2.7799) complexity for n x n approximate matrix multiplication," *Information Processing Letters*, vol. 8, no. 5, pp. 234–235, 1979.
- [21] V. De Silva and L.-H. Lim, "Tensor rank and the ill-posedness of the best low-rank approximation problem," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1084–1127, 2008.
- [22] J. Nolte, "The human brain: an introduction to its functional anatomy," 2002.
- [23] M. Kivelä, A. Arenas, M. Barthélemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of Complex Networks*, vol. 2, no. 3, pp. 203–271, 2014.
- [24] J. Yan, Y. Li, W. Liu, H. Zha, X. Yang, and S. M. Chu, "Graduated consistency-regularized optimization for multi-graph matching," in *European Conference on Computer Vision*, pp. 407–422, Springer, 2014.
- [25] H.-M. Park and K.-J. Yoon, "Encouraging second-order consistency for multiple graph matching," *Machine Vision and Applications*, vol. 27, no. 7, pp. 1021–1034, 2016.
- [26] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.
- [27] B. N. Flury and W. Gautschi, "An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form," *SIAM Journal on Scientific and Statistical Computing*, vol. 7, no. 1, pp. 169–184, 1986.
- [28] A. Ziehe, P. Laskov, G. Nolte, and K.-R. Müller, "A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation," *Journal of Machine Learning Research*, vol. 5, no. Jul, pp. 777–800, 2004.
- [29] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.
- [30] W. Tang, Z. Lu, and I. S. Dhillon, "Clustering with multiple graphs," in *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pp. 1016–1021, IEEE, 2009.
- [31] T. G. Kolda, "Numerical optimization for symmetric tensor decomposition," *Mathematical Programming*, vol. 151, no. 1, pp. 225–248, 2015.
- [32] J. C. Bezdek and R. J. Hathaway, "Convergence of alternating optimization," *Neural, Parallel & Scientific Computations*, vol. 11, no. 4, pp. 351–368, 2003.
- [33] S. J. Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.



- [34] A. Beck and L. Tetruashvili, "On the convergence of block coordinate descent type methods," *SIAM journal on Optimization*, vol. 23, no. 4, pp. 2037–2060, 2013.
- [35] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.
- [36] N. Bell and M. Garland, "Implementing sparse matrix-vector multiplication on throughput-oriented processors," in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, p. 18, ACM, 2009.
- [37] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *Journal of the ACM (JACM)*, vol. 60, no. 6, p. 45, 2013.
- [38] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM journal on matrix analysis and applications*, vol. 30, no. 2, pp. 713–730, 2008.
- [39] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [40] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.
- [41] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on information theory*, vol. 53, no. 12, pp. 4655–4666, 2007.
- [42] A. Athreya, C. Priebe, M. Tang, V. Lyzinski, D. Marchette, and D. Sussman, "A limit theorem for scaled eigenvectors of random dot product graphs," *Sankhya A*, pp. 1–18, 2013.
- [43] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14, pp. 1137–1145, 1995.
- [44] J. Neyman and E. L. Scott, "Consistent estimates based on partially consistent observations," *Econometrica: Journal of the Econometric Society*, pp. 1–32, 1948.
- [45] S. N. Dorogovtsev, A. V. Goltsev, J. F. Mendes, and A. N. Samukhin, "Spectra of complex networks," *Physical Review E*, vol. 68, no. 4, p. 046109, 2003.
- [46] X.-N. Zuo, J. S. Anderson, P. Bellec, R. M. Birn, B. B. Biswal, J. Blautzik, J. C. Breitner, R. L. Buckner, V. D. Calhoun, and F. X. Castellanos, "An open science resource for establishing reliability and reproducibility in functional connectomics," *Scientific data*, vol. 1, p. 140049, 2014.
- [47] G. Kiar, W. Gray Roncal, D. Mhembe, E. Bridgeford, R. Burns, and J. Vogelstein, "ndmg: Neurodata's mri graphs pipeline," Aug. 2016. open-source code.
- [48] J. Lancaster, "The Talairach Daemon, a database server for Talairach atlas labels," *Neuroimage*, vol. 5, no. 4, pp. 238–242, 1997.
- [49] Shangsi Wang and Carey Priebe and Cameron Craddock and Xi-Nian Zuo and Michael Milham and Joshua T. Vogelstein, "Optimal design for discovery science: Applications in neuroimaging," in *Organization for Human Brain Mapping (OHBM)*.
- [50] E. S. Finn, X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable, "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity," *Nature Neuroscience*, vol. 18, no. 11, p. 1664, 2015.
- [51] R. Tang, M. Ketcha, A. Badea, E. D. Calabrese, D. S. Margulies, J. T. Vogelstein, C. E. Priebe, and D. L. Sussman, "Connectome smoothing via low-rank approximations," *arXiv preprint arXiv:1609.01672*, 2016.
- [52] C. Sripada, M. Angstadt, S. Rutherford, D. Kessler, Y. Kim, M. Yee, and L. Levina, "Fundamental units of inter-individual variation in resting state connectomes," *bioRxiv*, p. 326082, 2018.
- [53] D. Koutra, J. T. Vogelstein, and C. Faloutsos, "DeltaCon: A principled massive-graph similarity function," in *Proceedings of the SIAM International Conference in Data Mining. Society for Industrial and Applied Mathematics*, pp. 162–170, SIAM, 2013.
- [54] R. Arden, R. S. Chavez, R. Grazioplene, and R. E. Jung, "Neuroimaging creativity: a psychometric view," *Behavioural Brain Research*, vol. 214, no. 2, pp. 143–156, 2010.
- [55] M. Brant-Zawadzki, G. D. Gillan, and W. R. Nitz, "MP RAGE: a three-dimensional, T1-weighted, gradient-echo sequence—initial experience in the brain," *Radiology*, vol. 182, no. 3, pp. 769–775, 1992.
- [56] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, and B. T. Hyman, "An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest," *Neuroimage*, vol. 31, no. 3, pp. 968–980, 2006.
- [57] T. M. Amabile, "The social psychology of creativity: A componential conceptualization," *Journal of Personality and Social Psychology*, vol. 45, no. 2, p. 357, 1983.
- [58] V. Lyzinski, D. L. Sussman, M. Tang, A. Athreya, and C. E. Priebe, "Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding," *Electronic Journal of Statistics*, vol. 8, no. 2, pp. 2905–2922, 2014.
- [59] S. Suwan, D. S. Lee, R. Tang, D. L. Sussman, M. Tang, and C. E. Priebe, "Empirical bayes estimation for the stochastic block-model," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 761–782, 2016.
- [60] D. Steinley, "Properties of the hubert-arable adjusted rand index," *Psychological methods*, vol. 9, no. 3, p. 386, 2004.
- [61] E. Rendón, I. Abundez, A. Arizmendi, and E. Quiroz, "Internal versus external cluster validation indexes," *International Journal of Computers and Communications*, vol. 5, no. 1, pp. 27–34, 2011.
- [62] R. I. Jennrich, "Asymptotic properties of non-linear least squares estimators," *The Annals of Mathematical Statistics*, vol. 40, no. 2, pp. 633–643, 1969.
- [63] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics, volume I*, vol. 117, ch. 6. CRC Press, 2015.
- [64] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation," *SIAM Journal on Numerical Analysis*, vol. 7, no. 1, pp. 1–46, 1970.



**Shangsi Wang** received the BS degree in mathematics and actuarial science from University of Waterloo in 2012, the MA degree from the Department of Applied Mathematics and Statistics at Johns Hopkins University, in 2014. He is currently a graduate student at the Johns Hopkins University, working toward the PhD degree in statistics. His research interests include statistical inference on random networks and pattern recognition in graph datasets.



**Jesús Arroyo** received the BS degrees in applied mathematics and computer engineering from the Instituto Tecnológico Autónomo de México (ITAM) at Mexico City in 2013, and the MA and PhD degrees from the Department of Statistics at the University of Michigan, Ann Arbor, in 2018. He is postdoctoral fellow in the Center for Imaging Science at JHU since 2018. His research interests include statistical analysis of graphs and high dimensional data, and applications to neuroimaging.



**Joshua T. Vogelstein** received the BS degree from the Department of Biomedical Engineering (BME) at Washington University in St. Louis, in 2002, the MS degree from the Department of Applied Mathematics and Statistics at Johns Hopkins University (JHU), in 2009, and the PhD degree from the Department of Neuroscience at JHU in 2009. He was a postdoctoral fellow in AMS at JHU from 2009 until 2011, when he was appointed an assistant research scientist, and became a member of the Institute for Data Intensive Science and Engineering. He spent two years at Information Initiative at Duke, before becoming Assistant Professor in BME at JHU, and core faculty in the Institute for Computational Medicine and the Center for Imaging Science. His research interests include computational statistics, focusing on ultrahigh-dimensional and non-Euclidean neuroscience data, especially connectomics.



**Carey E. Priebe** received the BS degree in mathematics from Purdue University, in 1984, the MS degree in computer science from San Diego State University, in 1988, and the PhD degree in information technology (computational statistics) from George Mason University, in 1993. From 1985 to 1994, he worked as a mathematician and scientist in the US Navy research and development laboratory system. Since 1994, he has been a professor in the Department of Applied Mathematics and Statistics, Johns Hopkins University (JHU). His research interests include computational statistics, kernel and mixture estimates, statistical pattern recognition, statistical image analysis, dimensionality reduction, model selection, and statistical inference for high-dimensional and graph data. He is a Senior Member of the IEEE, a Lifetime Member of the Institute of Mathematical Statistics, an Elected Member of the International Statistical Institute, and a Fellow of the American Statistical Association.