On a two-truths phenomenon in spectral graph clustering

Carey E. Priebe[a,b,c,1], Youngser Park[b], Joshua T. Vogelstein[b,d], John M. Conroy[e], Vince Lyzinski[c,f], Minh Tang[a], Avanti Athreya[a], Joshua Cape[a], and Eric Bridgeford[b,g]

[a]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218; [b]Center for Imaging Science, Johns Hopkins University, Baltimore, MD 21218; [c]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD 21218; [d]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218; [e]Institute for Defense Analyses, Center for Computing Sciences, Bowie, MD 20715; [f]Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003; and [g]Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21218

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved February 8, 2019 (received for review August 21, 2018)

Clustering is concerned with coherently grouping observations without any explicit concept of true groupings. Spectral graph clustering—clustering the vertices of a graph based on their spectral embedding—is commonly approached via *K*-means (or, more generally, Gaussian mixture model) clustering composed with either Laplacian spectral embedding (LSE) or adjacency spectral embedding (ASE). Recent theoretical results provide deeper understanding of the problem and solutions and lead us to a "two-truths" LSE vs. ASE spectral graph clustering phenomenon convincingly illustrated here via a diffusion MRI connectome dataset: The different embedding methods yield different clustering results, with LSE capturing left hemisphere/right hemisphere affinity structure and ASE capturing gray matter/white matter core–periphery structure.

spectral embedding | spectral clustering | graph | network | connectome

The purpose of this paper is to cogently present a "two-truths" phenomenon in spectral graph clustering, to understand this phenomenon from a theoretical and methodological perspective, and to demonstrate the phenomenon in a real-data case consisting of multiple graphs each with multiple categorical vertex class labels.

A graph or network consists of a collection of vertices or nodes V representing n entities together with edges or links E representing the observed subset of the $\binom{n}{2}$ possible pairwise relationships between these entities. Graph clustering, often associated with the concept of "community detection," is concerned with partitioning the vertices into coherent groups or clusters. By its very nature, such a partitioning must be based on connectivity patterns.

It is often the case that practitioners cluster the vertices of a graph—say, via K-means clustering composed with Laplacian spectral embedding—and pronounce the method as having performed either well or poorly based on whether the resulting clusters correspond well or poorly with some known or preconceived notion of "correct" clustering. Indeed, such a procedure may be used to compare two clustering methods and to pronounce that one works better (on the particular data under consideration). However, clustering is inherently ill-defined, as there may be multiple meaningful groupings, and two clustering methods that perform differently with respect to one notion of truth may in fact be identifying inherently different, but perhaps complementary, underlying structure. With respect to graph clustering, ref. 1 shows that there can be no algorithm that is optimal for all possible community detection tasks (Fig. 1).

We compare and contrast Laplacian and adjacency spectral embedding as the first step in spectral graph clustering and demonstrate that the two methods, and the two resulting clusterings, identify different—but both meaningful—graph structure. We trust that this simple, clear explication will contribute to an awareness that connectivity-based structure discovery via spectral graph clustering should consider both Laplacian and adjacency spectral embedding and the development of new methodologies based on this awareness.

Spectral Graph Clustering

Given a simple graph $G = (V, E)$ on n vertices, consider the associated $n \times n$ adjacency matrix A in which $A_{ij} = 0$ or 1 encoding whether vertices i and j in V share an edge (i, j) in E. For our simple undirected, unweighted, loopless case, A is binary with $A_{ij} \in \{0, 1\}$, symmetric with $A = A^\top$, and hollow with $diag(A) = \vec{0}$.

The first step of spectral graph clustering (2, 3) involves embedding the graph into Euclidean space via an eigendecomposition. We consider two options: Laplacian spectral embedding (LSE), wherein we decompose the normalized Laplacian of the adjacency matrix, and adjacency spectral embedding (ASE) given by a decomposition of the adjacency matrix itself. With target dimension d, either spectral embedding method produces n points in $\Re^d$, denoted by the $n \times d$ matrix X. ASE employs the eigendecomposition to represent the adjacency matrix via $A = USU^\top$ and chooses the top d eigenvalues by magnitude and their associated vectors to embed the graph via the scaled eigenvectors $U_d|S_d|^{1/2}$. Similarly, LSE embeds the graph via the top scaled eigenvectors of the normalized Laplacian $\mathcal{L}(A) = D^{-1/2}AD^{-1/2}$, where D is the diagonal matrix of vertex degrees.

Significance

Spectral graph clustering—clustering the vertices of a graph based on their spectral embedding—is of significant current interest, finding applications throughout the sciences. But as with clustering in general, what a particular methodology identifies as "clusters" is defined (explicitly, or, more often, implicitly) by the clustering algorithm itself. We provide a clear and concise demonstration of a "two-truths" phenomenon for spectral graph clustering in which the first step—spectral embedding—is either Laplacian spectral embedding, wherein one decomposes the normalized Laplacian of the adjacency matrix, or adjacency spectral embedding given by a decomposition of the adjacency matrix itself. The two resulting clustering methods identify fundamentally different (true and meaningful) structure.

Author contributions: C.E.P., J.T.V., J.M.C., and V.L. designed research; C.E.P., V.L., M.T., A.A., and J.C. performed research; C.E.P., Y.P., and E.B. analyzed data; and C.E.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: cep@jhu.edu.

Published online March 8, 2019.

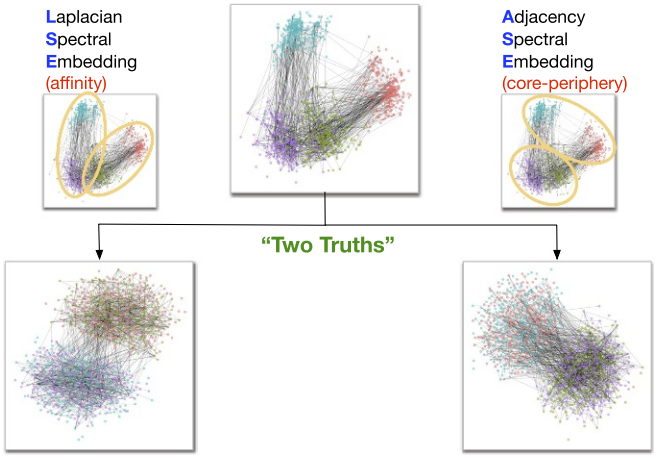

Fig. 1. A two-truths graph (connectome) depicting connectivity structure such that one grouping of the vertices yields affinity structure (e.g., left hemisphere/right hemisphere) and the other grouping yields core–periphery structure (e.g., gray matter/white matter). (*Top Center*) The graph with four vertex colors. (*Top Left* and *Top Right*) LSE groups one way and ASE groups another way. (*Bottom Left*) The LSE truth is two densely connected groups, with sparse interconnectivity between them (affinity structure). (*Bottom Right*) The ASE truth is one densely connected group, with sparse interconnectivity between it and the other group and sparse interconnectivity within the other group (core–periphery structure). This paper demonstrates the two-truths phenomenon illustrated here—that LSE and ASE find fundamentally different but equally meaningful network structure—via theory, simulation, and real data analysis.

In either case, each vertex is mapped to the corresponding row of $X = U_d|S_d|^{1/2}$.

Spectral graph clustering concludes via classical Euclidean clustering of the rows of X. As described below, central limit theorems for spectral embedding of the (sufficiently dense) stochastic block model via either LSE or ASE suggest Gaussian mixture modeling (GMM) for this clustering step. Thus, we consider spectral graph clustering to be GMM composed with LSE or ASE:

$$\mathrm{GMM} \circ \{\mathrm{LSE}, \mathrm{ASE}\}.$$

Stochastic Block Model

The random graph model we use to illustrate our phenomenon is the stochastic block model (SBM), introduced in ref. 4. This model is parameterized by (*i*) a block membership probability vector $\vec{\pi} = [\pi_1, \ldots, \pi_K]^\top$ in the unit simplex and (*ii*) a symmetric $K \times K$ block connectivity probability matrix B with entries in $[0, 1]$ governing the probability of an edge between vertices given their block memberships. Use of the SBM is ubiquitous in theoretical, methodological, and practical graph investigations, and SBMs have been shown to be universal approximators for exchangeable random graphs (5).

For sufficiently dense graphs, both LSE and ASE have a central limit theorem (6–8) demonstrating that, for large n, embedding via the top d eigenvectors from a rank d K-block SBM ($d \equiv rank(B) \leq K$) yields n points in $\Re^d$ behaving approximately as a random sample from a mixture of K Gaussians. That is, given that the ith vertex belongs to block k, the ith row of $X = U_d|S_d|^{1/2}$ will be approximately distributed as a multivariate normal with parameters specific to block k, $X_i \sim \mathcal{MVN}(\mu_k, \Sigma_k)$. The structure of the covariance matrices suggests that the GMM is called for, as an appropriate generalization of K-means clustering. Therefore, GMM(X) via maximum likelihood will produce mixture parameter estimates and associated asymptotically perfect clustering, using either LSE or ASE.

For finite n, however, LSE and ASE yield different clustering performance, and neither one dominates the other.

We make significant conceptual use of the positive definite two-block SBM ($K = 2$), with

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

which henceforth we abbreviate as $B = [a, b; b, c]$. In this simple setting, two general/generic cases present themselves: affinity and core–periphery.

Affinity: *a, c* $\gg$ *b*. An SBM with $B = [a, b; b, c]$ is said to exhibit affinity structure if each of the two blocks has a relatively high within-block connectivity probability compared with the between-block connectivity probability.

Core-periphery: *a* $\gg$ *b, c*. An SBM with $B = [a, b; b, c]$ is said to exhibit core–periphery structure if one of the two blocks has a relatively high within-block connectivity probability compared with both the other block's within-block connectivity probability and the between-block connectivity probability.

The relative performance of LSE and ASE for these two cases provides the foundation for our analyses. Informally, LSE outperforms ASE for affinity, and ASE is the better choice for core–periphery. We make this clustering performance assessment analytically precise via Chernoff information, and we demonstrate this in practice via the adjusted Rand index.

Clustering Performance Assessment

We consider two approaches to assessing the performance of a given clustering, defined to be a partition of $[n] \equiv \{1, \ldots, n\}$ into a disjoint union of K partition cells or clusters. For our purposes—demonstrating a two-truths phenomenon in LSE vs. ASE spectral graph clustering—we consider the case in which there is a "true" or meaningful clustering of the vertices against which we can assess performance, but we emphasize that in practice such a truth is neither known nor necessarily unique.

Chernoff Information. Comparing and contrasting the relative performance of LSE vs. ASE via the concept of Chernoff information (9, 10), in the context of their respective central limit theorems (CLTs), provides a limit theorem notion of superiority. Thus, in the SBM, we allude to the GMM provided by the CLT for either LSE or ASE.

The Chernoff information between two distributions is the exponential rate at which the decision-theoretic Bayes error decreases as a function of sample size. In the two-block SBM, with the true clustering of the vertices given by the block memberships, we are interested in the large-sample optimal error rate for recovering the underlying block memberships after the spectral embedding step has been carried out. Thus, we require the Chernoff information $C(F_1, F_2)$ when $F_1 = \mathcal{MVN}(\mu_1, \Sigma_1)$ and $F_2 = \mathcal{MVN}(\mu_2, \Sigma_2)$ are multivariate normals. Letting $\Sigma_t = t\Sigma_1 + (1-t)\Sigma_2$ and

$$h(t; F_1, F_2) = \frac{t(1-t)}{2}(\mu_1 - \mu_2)^\top \Sigma_t^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\log\frac{|\Sigma_t|}{|\Sigma_1|^t|\Sigma_2|^{1-t}},$$

we have

$$\rho_{F_1,F_2} = \sup_{t\in(0,1)} h(t; F_1, F_2).$$

This provides both ρ_L and ρ_A when using the large-sample GMM parameters for F_1, F_2 obtained from the LSE and ASE embeddings, respectively, for a particular two-block SBM distribution (defined by its block membership probability vector $\vec{\pi}$ and block

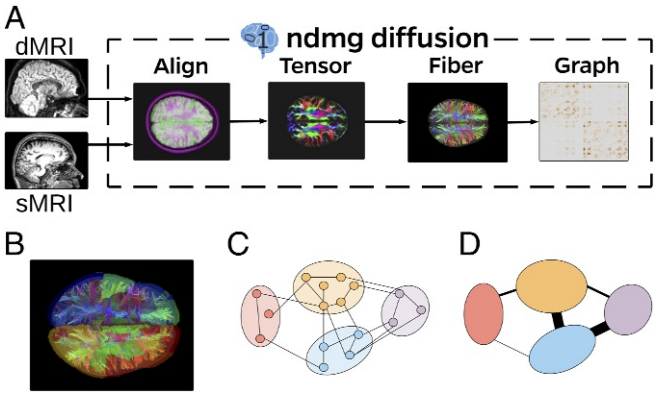

Fig. 2. Connectome data generation. (*A*) The pipeline. (*B*) Voxels and regions in tractography map. (*C*) Voxels and edges. (*D*) Contraction yields vertices and edges. The output is diffusion MRI graphs on ≈1 million vertices. Spatial vertex contraction yields graphs on ≈70,000 vertices from which we extract largest connected components of ≈40,000 vertices with {Left,Right} and {Gray,White} labels for each vertex. Fig. 1 depicts (a subsample from) one such graph.

connectivity probability matrix B). We make use of the Chernoff ratio $\rho = \rho_A/\rho_L$; $\rho > 1$ implies ASE is preferred while $\rho < 1$ implies LSE is preferred. (Recall that as the Chernoff information increases, the large-sample optimal error rate decreases.) Chernoff analysis in the two-block SBM demonstrates that, in general, LSE is preferred for affinity while ASE is preferred for core–periphery (7, 11).

Adjusted Rand Index. In practice, we wish to empirically assess the performance of a particular clustering algorithm on a given graph. There are numerous cluster assessment criteria available in the literature: the Rand index (RI) (12), normalized mutual information (NMI) (13), variation of information (VI) (14), Jaccard (15), etc. These are typically used to compare either an empirical clustering against a "truth" or two separate empirical clusterings. For concreteness, we consider the well-known adjusted Rand index (ARI), popular in machine learning, which normalizes the RI so that expected chance performance is zero: The ARI is the adjusted-for-chance probability that two partitions of a collection of data points will agree for a randomly chosen pair of data points, putting the pair into the same partition cell in both clusterings or splitting the pair into different cells in both clusterings. (Our empirical connectome results are essentially unchanged when using other cluster assessment criteria.)

In the context of spectral clustering via GMM $\circ$ {LSE, ASE}, we consider $\mathcal{C}_{LSE}$ and $\mathcal{C}_{ASE}$ to be the two clusterings of the vertices of a given graph. Then $\text{ARI}(\mathcal{C}_{LSE}, \mathcal{C}_{ASE})$ assesses their agreement: $\text{ARI}(\mathcal{C}_{LSE}, \mathcal{C}_{ASE}) = 1$ implies that the two clusterings are identical; $\text{ARI}(\mathcal{C}_{LSE}, \mathcal{C}_{ASE}) \approx 0$ implies that the two spectral embedding methods are "operationally orthogonal." (Significance is assessed via permutation testing.)

In the context of two truths, we consider $\mathcal{C}_1$ and $\mathcal{C}_2$ to be two known true or meaningful clusterings of the vertices. Then, with $\mathcal{C}_{SE}$ being either $\mathcal{C}_{LSE}$ or $\mathcal{C}_{ASE}$, $\text{ARI}(\mathcal{C}_{SE}, \mathcal{C}_1) \gg \text{ARI}(\mathcal{C}_{SE}, \mathcal{C}_2)$ implies that the spectral embedding method under consideration is more adept at discovering truth $\mathcal{C}_1$ than truth $\mathcal{C}_2$. Analogous to the theoretical Chernoff analysis, ARI simulation studies in the two-block SBM demonstrate that, in general, LSE is preferred for affinity while ASE is preferred for core–periphery.

Model Selection × 2

To perform the spectral graph clustering GMM $\circ$ {LSE, ASE} in practice, we must address two inherent model selection problems: We must choose the embedding dimension ($\widehat{d}$) and the number of clusters ($\widehat{K}$).

SBM vs. Network Histogram. If the SBM were actually true, then as $n \to \infty$ any reasonable procedure for estimating the singular value decomposition (SVD) rank would yield a consistent estimator $\widehat{d} \to d$ and any reasonable procedure for estimating the number of clusters would yield a consistent estimator $\widehat{K} \to K$. Critically, the universal approximation result of ref. 5 shows that SBMs provide a principled "network histogram" model even without the assumption that the SBM with some fixed (d, K) actually holds. Thus, practical model selection for spectral graph clustering is concerned with choosing $(\widehat{d}, \widehat{K})$ to provide a useful approximation.

The bias–variance tradeoff demonstrates that any quest for a universally optimal methodology for choosing the "best" dimension and number of clusters, in general, for finite n, is a losing proposition. Even for a low-rank model, subsequent inference may be optimized by choosing a dimension smaller than the true signal dimension, and even for a mixture of K Gaussians, inference performance may be optimized by choosing a number of clusters smaller than the true cluster complexity. In the case of semiparametric SBM fitting, wherein low-rank and finite mixtures are used as a practical modeling convenience as opposed to a believed true model, and one presumes that both $\widehat{d}$ and $\widehat{K}$ will tend to infinity as $n \to \infty$, these bias–variance tradeoff considerations are exacerbated.

For $\widehat{d}$ and $\widehat{K}$ below, we make principled methodological choices for simplicity and concreteness, but make no claim that these are best in general or even for the connectome data considered herein. Nevertheless, one must choose an embedding dimension and a mixture complexity, and thus we proceed.

Choosing the Embedding Dimension $\widehat{d}$. A ubiquitous and principled general methodology for choosing the number of

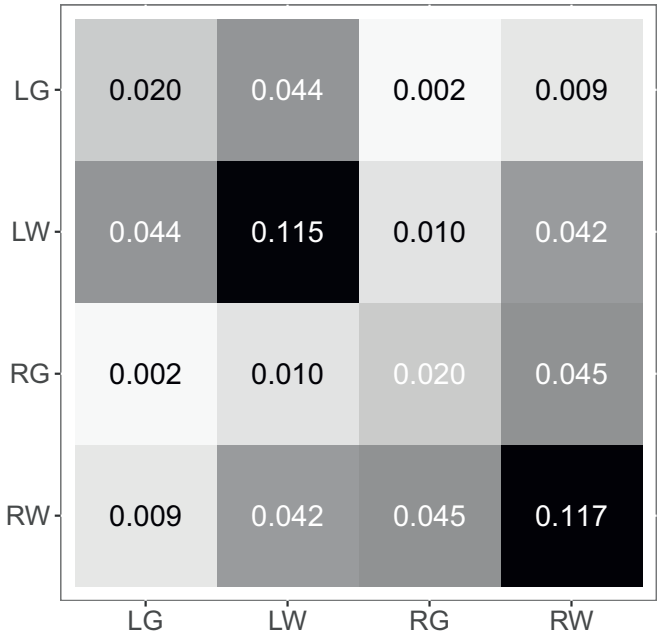

Fig. 3. Block connectivity probability matrix for the {LG,LW,RG,RW} a priori projection of the composite connectome onto the four-block SBM. The two two-block projections ({Left, Right} and {Gray, White}) are shown in Fig. 4. This synthetic SBM exhibits the two-truths phenomenon both theoretically (via Chernoff analysis) and in simulation (via Monte Carlo).

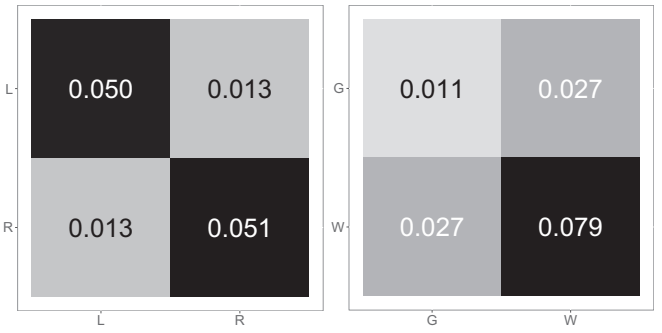

Fig. 4. Block connectivity probability matrices for the a priori projection of the composite connectome onto the two-block SBM for (*Left*) {Left, Right} and (*Right*) {Gray, White}. {Left, Right} exhibits affinity structure, with Chernoff ratio <1; {Gray, White} exhibits core–periphery structure, with Chernoff ratio >1.

dimensions in eigendecompositions and SVDs (e.g., principal components analysis, factor analysis, spectral embedding, etc.) is to examine the so-called scree plot and look for "elbows" defining the cutoff between the top (signal) dimensions and the noise dimensions. There are a plethora of variations for automating this singular value thresholding (SVT); section 2.8 of ref. 16 provides a comprehensive discussion in the context of principal components, and ref. 17 provides a theoretically justified (but perhaps practically suspect, for small n) universal SVT. We consider the profile-likelihood SVT method of ref. 18. Given $A = USU^\top$ (for either LSE or ASE) the singular values S are used to choose the embedding dimension $\widehat{d}$ via

$$\widehat{d} = \arg\max_d \mathit{ProfileLikelihood}_S(d),$$

where $\mathit{ProfileLikelihood}_S(d)$ provides a definition for the magnitude of the "gap" after the first d singular values.

Choosing the Number of Clusters $\widehat{K}$. Choosing the number of clusters in Gaussian mixture models is most often addressed by maximizing a fitness criterion penalized by model complexity. Common approaches include the Akaike information criterion (AIC) (19), the Bayesian information criterion (BIC) (20), minimum description length (MDL) (21), etc. We consider penalized likelihood via the BIC (22). Given n points in $\Re^d$ represented by $X = U_d|S_d|^{1/2}$ (obtained via either LSE or ASE) and letting θ_K represent the GMM parameter vector whose dimension $dim(\theta_K)$ is a function of the data dimension d, the mixture complexity $\widehat{K}$ is chosen via

$$\widehat{K} = \arg\max_K \mathit{PenalizedLikelihood}_X(\widehat{\theta}_K),$$

where $\mathit{PenalizedLikelihood}_X(\widehat{\theta}_K)$ is twice the log-likelihood of the data X evaluated at the GMM with mixture parameter estimate $\widehat{\theta}_K$ penalized by $dim(\theta_K) \cdot \ln\, n$. For spectral clustering, we use the BIC for $\widehat{K}$ after spectral embedding, so $X \in \Re^{\widehat{d}}$ with $\widehat{d}$ chosen as above.

Connectome Data

We consider for illustration a diffusion MRI dataset consisting of 114 connectomes (57 subjects, two scans each) with 72,783 vertices each and both left/right/other hemispheric and gray/white/other tissue attributes for each vertex. Graphs were estimated using the NeuroData's MR Graphs pipeline (23), with vertices representing subregions defined via spatial proximity and edges defined by tensor-based fiber streamlines connecting these regions (Fig. 2).

The actual graphs we consider are the largest connected component (LCC) of the induced subgraph on the vertices labeled as both left or right and gray or white. This yields $m = 114$ connected graphs on $n \approx 40,000$ vertices. Additionally, for each graph every vertex has a {Left,Right} label and a {Gray,White} label, which we sometimes find convenient to consider as a single label in {LG,LW,RG,RW}.

Sparsity. The only notions of sparsity relevant here are linear algebraic: whether there are enough edges in the graph to support spectral embedding and whether there are few enough to allow for sparse matrix computations. We have a collection of observed connectomes and we want to cluster the vertices in these graphs, as opposed to in an unobserved sequence with the number of vertices tending to infinity. Our connectomes have, on average, $n \approx 40,000$ vertices and $e \approx 2,000,000$ edges, for an average degree $2e/n \approx 100$ and a graph density $e/\binom{n}{2} \approx 0.0025$.

Synthetic Analysis. We consider a synthetic data analysis via a priori projections onto the SBM—block model estimates based on known or assumed block memberships. Averaging the collection of $m = 114$ connectomes yields the composite (weighted) graph adjacency matrix $\bar{A}$. The {LG,LW,RG,RW} projection of the binarized $\bar{A}$ onto the four-block SBM yields the block connectivity probability matrix B presented in Fig. 3 and the block membership probability vector $\vec{\pi} = [0.28, 0.22, 0.28, 0.22]^\top$. Limit theory demonstrates that spectral graph clustering using $d = K = 4$ will, for large n, correctly identify block memberships

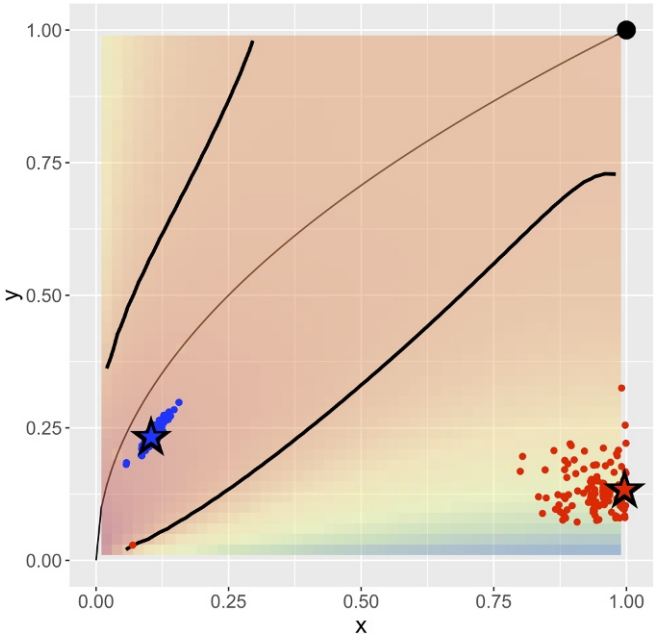

Fig. 5. For each of our 114 connectomes, we plot the a priori two-block SBM projections for {Left, Right} in red and {Gray, White} in blue. The coordinates are given by $x = \min(a, c)/\max(a, c)$ and $y = b/\max(a, c)$, where $B = [a, b; b, c]$ is the observed block connectivity probability matrix. The thin black curve $y = \sqrt{x}$ represents the rank 1 submodel separating positive definite (lower right) from indefinite (upper left). The background color shading is Chernoff ratio ρ, and the thick black curves are $\rho = 1$ separating the region where ASE is preferred (between the curves) from where LSE is preferred. The point (1, 1) represents Erdős–Rényi ($a = b = c$). The large stars are from the a priori composite connectome projections (Fig. 4). We see that the red {Left, Right} projections are in the affinity region where $\rho < 1$ and LSE is preferred while the blue {Gray, White} projections are in the core–periphery region where $\rho > 1$ and ASE is preferred. This analytical finding based on projections onto the SBM carries over to empirical spectral clustering results on the individual connectomes (Fig. 7).

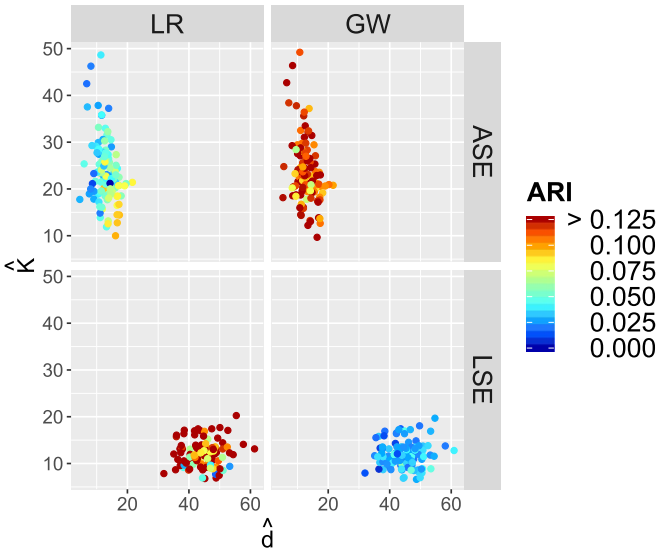

Fig. 6. Results of the $(\widehat{d}, \widehat{K})$ model selection for spectral graph clustering for each of our 114 connectomes. For LSE we see $\widehat{d} \in \{30, \ldots, 60\}$ and $\widehat{K} \in \{2, \ldots, 20\}$; for ASE we see $\widehat{d} \in \{2, \ldots, 20\}$ and $\widehat{K} \in \{10, \ldots, 50\}$. The color coding represents clustering performance in terms of ARI for each of LSE and ASE against each of the two truths {Left, Right} and {Gray, White} and shows that LSE clustering identifies {Left, Right} better than {Gray, White} and ASE identifies {Gray, White} better than {Left, Right}. Our two-truths phenomenon is conclusively demonstrated: LSE finds {Left, Right} (affinity) while ASE finds {Gray, White} (core–periphery).

for this four-block case when using either LSE or ASE. Our interest is to compare and contrast the two spectral embedding methods for clustering into two clusters. We demonstrate that this synthetic case exhibits the two-truths phenomenon both theoretically and in simulation—the {LG,LW,RG,RW} a priori projection of our composite connectome yields a four-block two-truths SBM.

Two-Block Projections. A priori projections onto the two-block SBM for {Left,Right} and {Gray,White} yield the two-block connectivity probability matrices shown in Fig. 4. It is apparent that the {Left,Right} a priori block connectivity probability matrix $B = [a, b; b, c]$ represents an affinity SBM with $a \approx c \gg b$ and the {Gray,White} a priori projection yields a core–periphery SBM with $c \gg a \approx b$. It remains to investigate the extent to which the Chernoff analysis from the two-block setting (LSE is preferred for affinity while ASE is preferred for core–periphery) extends to such a four-block two-truths case; we do so theoretically and in simulation using this synthetic model derived from the {LG,LW,RG,RW} a priori projection of our composite connectome in *Theoretical Results* and *Simulation Results* and then empirically on the original connectomes in *Connectome Results*.

Theoretical Results. Analysis using the large-sample Gaussian mixture model approximations from the LSE and ASE CLTs shows that the 2D embedding of the four-block model, when clustered into two clusters, will yield { {LG,LW}, {RG,RW} } (i.e., {Left, Right}) when embedding via LSE and { {LG,RG}, {LW,RW} } (i.e., {Gray, White}) when using ASE. That is, using numerical integration for the $d = K = 2$ GMM ∘ LSE, the largest Kullback–Leibler divergence (as a surrogate for Chernoff information) among the 10 possible ways of grouping the four Gaussians into two clusters is for the { {LG,LW}, {RG,RW} } grouping, and the largest of these values for the GMM ∘ ASE is for the { {LG,RG}, {LW,RW} } grouping.

Simulation Results. We augment the Chernoff limit theory via Monte Carlo simulation, sampling graphs from the four-block model and running the GMM ∘ {LSE, ASE} algorithm specifying $\widehat{d} = \widehat{K} = 2$. This results in LSE finding {Left, Right} (ARI > 0.95) with probability >0.95 and ASE finding {Gray, White} (ARI > 0.95) with probability >0.95.

Connectome Results. Figs. 5–7 present empirical results for the connectome dataset, $m = 114$ graphs each on $n \approx 40,000$ vertices. We note that these connectomes are most assuredly not four-block two-truths SBMs of the kind presented in Figs. 3 and 4, but they do have two truths ({Left, Right} and {Gray, White}) and, as we shall see, they do exhibit a real-data version of the synthetic results presented above, in the spirit of semiparametric SBM fitting.

First, in Fig. 5, we consider a priori projections of the individual connectomes, analogous to the Fig. 4 projections of the composite connectome. Letting $B = [a, b; b, c]$ be the observed block connectivity probability matrix for the a priori two-block SBM projection ({Left, Right} or {Gray, White}) of a given individual connectome, the coordinates in Fig. 5 are given by $x = \min(a, c)/\max(a, c)$ and $y = b/\max(a, c)$. Each graph yields two points, one for each of {Left, Right} and {Gray, White}. We see that the {Left, Right} projections are in the affinity region (large x and small y imply $a \approx c \gg b$, where Chernoff ratio $\rho < 1$ and LSE is preferred) while the {Gray, White} projections are in the core–periphery region [small x and small y imply $\max(a, c) \gg b \approx \min(a, c)$, where $\rho > 1$ and ASE is preferred]. This exploratory data analysis finding indicates complex two-truths structure in our connectome dataset. [Of independent interest, we propose Fig. 5 as the representative for an illustrative two-truths exploratory data analysis (EDA) plot

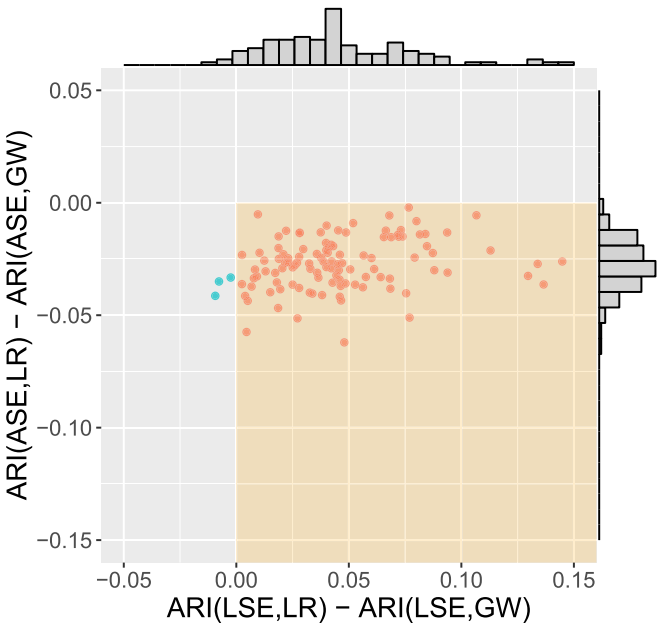

Fig. 7. Spectral graph clustering assessment via ARI. For each of our 114 connectomes, we plot the difference in ARI for the {Left, Right} truth against the difference in ARI for the {Gray, White} truth for the clusterings produced by each of LSE and ASE: x = ARI(LSE,LR) − ARI(LSE,GW) vs. y = ARI(ASE,LR) − ARI(ASE,GW). A point in the (+, −) quadrant indicates that for that connectome the LSE clustering identified {Left, Right} better than {Gray, White} and ASE identified {Gray, White} better than {Left, Right}. Marginal histograms are provided. Our two-truths phenomenon is conclusively demonstrated: LSE identifies {Left, Right} (affinity) while ASE identifies {Gray, White} (core–periphery).

for a dataset of m graphs with multiple categorical vertex labels.]

In Figs. 6 and 7 we present the results of $m = 114$ runs of the spectral clustering algorithm GMM $\circ$ $\{\text{LSE}, \text{ASE}\}$. We consider each of LSE and ASE, choosing $\widehat{d}$ and $\widehat{K}$ as described above. The resulting empirical clusterings are evaluated via the ARI against each of the $\{$Left, Right$\}$ and $\{$Gray, White$\}$ truths. In Fig. 6 we present the results of the $(\widehat{d}, \widehat{K})$ model selection, and we observe that ASE is choosing $\widehat{d} \in \{2, \ldots, 20\}$ and LSE is choosing $\widehat{d} \in \{30, \ldots, 60\}$, while ASE is choosing $\widehat{K} \in \{10, \ldots, 50\}$ and LSE is choosing $\widehat{K} \in \{2, \ldots, 20\}$. In Fig. 7, each graph is represented by a single point, plotting x = ARI(LSE,LR) – ARI(LSE,GW) vs. y = ARI(ASE,LR) – ARI(ASE,GW), where "LSE" (resp. "ASE") represents the empirical clustering $\mathcal{C}_{LSE}$ (resp. $\mathcal{C}_{ASE}$) and "LR" (resp. "GW") represents the true clustering $\mathcal{C}_{\{\text{Left,Right}\}}$ (resp. $\mathcal{C}_{\{\text{Gray,White}\}}$). We see that almost all of the points lie in the $(+,-)$ quadrant, indicating ARI(LSE,LR) $>$ ARI(LSE,GW) and ARI(ASE,LR) $<$ ARI(ASE,GW). That is, LSE finds the affinity $\{$Left, Right$\}$ structure and ASE finds the core–periphery $\{$Gray, White$\}$ structure. The two-truths structure in our connectome dataset illustrated in Fig. 5 leads to fundamentally different but equally meaningful LSE vs. ASE spectral clustering performance. This is our two-truths phenomenon in spectral graph clustering.

Conclusion

The results presented herein demonstrate that practical spectral graph clustering exhibits a two-truths phenomenon with respect to Laplacian vs. adjacency spectral embedding. This phenomenon can be understood theoretically from the perspective of affinity vs. core–periphery stochastic block models and via consideration of the two a priori projections of a four-block two-truths SBM onto the two-block SBM. For connectomics, this phenomenon manifests itself via LSE better capturing the left hemisphere/right hemisphere affinity structure and ASE better capturing the gray matter/white matter core–periphery structure and suggests that a connectivity-based parcellation based on spectral clustering should consider both LSE and ASE, as the two spectral embedding approaches facilitate the identification of different and complementary connectivity-based clustering truths.

ACKNOWLEDGMENTS. The authors thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, United Kingdom, for support and hospitality during the program Theoretical Foundations for Statistical Network Analysis (Engineering and Physical Sciences Research Council Grant EP/K032208/1), where a portion of the work on this paper was undertaken, and the University of Haifa, where these ideas were conceived in June 2014. This work is partially supported by Defense Advanced Research Projects Agency (XDATA, GRAPHS, SIMPLEX, D3M), Johns Hopkins University Human Language Technology Center of Excellence, and the Acheson J. Duncan Fund for the Advancement of Research in Statistics.

1. Peel L, Larremore DB, Clauset A (2017) The ground truth about metadata and community detection in networks. *Sci Adv* 3:e1602548.
2. von Luxburg U (2007) A tutorial on spectral clustering. *Stat Comput* 17:395–416.
3. Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic blockmodel. *Ann Stat* 39:1878–1915.
4. Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: First steps. *Soc Networks* 5:109–137.
5. Olhede SC, Wolfe PJ (2014) Network histograms and universality of blockmodel approximation. *Proc Natl Acad Sci USA* 111:14722–14727.
6. Athreya A, et al. (2016) A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A* 78:1–18.
7. Tang M, Priebe CE (2018) Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *Ann Stat* 46:2360–2415.
8. Rubin-Delanchy P, Priebe CE, Tang M, Cape J (2018) The generalised random dot product graph. Available at https://arxiv.org/abs/1709.05506. Preprint, posted July 29, 2018.
9. Chernoff H (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann Math Stat* 23:493–507.
10. Chernoff H (1956) Large sample theory: Parametric case. *Ann Math Stat* 27:1–22.
11. Cape J, Tang M, Priebe CE, On spectral embedding performance and elucidating network structure in stochastic block model graphs. *Network Science*, in press.
12. Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2:193–218.
13. Danon L, Díaz-Guilera A, Duch J, Arena A (2005) Comparing community structure identification. *J Stat Mech Theory Exp* 2005:P09008.
14. Meilă M (2007) Comparing clusterings–an information based distance. *J Multivar Anal* 98:873–195.
15. Jaccard P (1912) The distribution of the flora in the alpine zone. *New Phytol* 11:37–50.
16. Jackson JE (2004) *A User's Guide to Principal Components* (Wiley, Hoboken, NJ).
17. Chatterjee S (2015) Matrix estimation by universal singular value thresholding. *Ann Stat* 43:177–214.
18. Zhu M, Ghodsi A (2006) Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Comput Stat Data Anal* 51:918–930.
19. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Autom Control* 19:716–723.
20. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464.
21. Rissanen J (1978) Modeling by shortest data description. *Automatica* 14:465–471.
22. Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis and density estimation. *J Am Stat Assoc* 97:611–631.
23. Kiar G, et al. (2018) A high-throughput pipeline identifies robust connectomes but troublesome variability. Available at https://www.biorxiv.org/node/94401. Preprint, posted April 24, 2018.