AFRL-RI-RS-TR-2017-149

FUSION AND INFERENCE FROM MULTIPLE AND MASSIVE DISPARATE DISTRIBUTED DYNAMIC DATA SETS

JOHNS HOPKINS UNIVERSITY

JULY 2017

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th ABW, Wright-Patterson AFB Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2017-149 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
NANCY ROBERTS
Work Unit Manager

/ S /
MICHAEL J. WESSING
Deputy Chief, Information Intelligence
Systems and Analysis Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
JULY 2017	FINAL TECHNICAL REPORT	SEP 2012 – MAR 2017

4. TITLE AND SUBTITLE

FUSION AND INFERENCE FROM MULTIPLE AND MASSIVE DISPARATE DISTRIBUTED DYNAMIC DATA SETS

5a. CONTRACT NUMBER
FA8750-12-2-0303

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
62702E

6. AUTHOR(S)

Carey E. Priebe

5d. PROJECT NUMBER
XDAT

5e. TASK NUMBER
A0

5f. WORK UNIT NUMBER
01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Johns Hopkins University
Research Projects Administration
1101 E. 33rd Street, B001
Baltimore, MD 21218

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIEA
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2017-149

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. PA# 88ABW-2017-3410
Date Cleared: 17 Jul 17

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We have developed the first principled methodology for two-sample graph testing; designed a provably almost-surely perfect vertex clustering algorithm for block model graphs; proved analogues of classical limit theorems for the adjacency and Laplacian embedding's for random graphs, which have led, in turn, to significantly improved algorithms for latent position estimation; established the accuracy of and efficiently implemented a fast, successfully scalable program for an approximate solution to the NP-hard problem of matching graphs; developed efficient methods for vertex nomination in graphs; determined precisely how to mitigate information loss across shuffled networks. This has led to dozens of papers published in top journals. Moreover, we have employed these theoretically-justified techniques on a suite of applications, conducting end-to-end analyses of real data from domains as varied as neuroscience, speech and language processing, threat detection, and social networks.

15. SUBJECT TERMS
Graph Embedding, Laplacian Embedding, Euclidean Space

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**NANCY ROBERTS**
U	U	U	UU	19	19b. TELEPHONE NUMBER (*Include area code*) **NA**

Table of Contents

1 SUMMARY

Our team has, in the course of this DARPA XDATA project, developed the first principled methodology for two-sample graph testing; designed a provably almost-surely perfect vertex clustering algorithm for block model graphs; proved analogues of classical limit theorems for the adjacency and Laplacian embeddings for random graphs, which have led, in turn, to significantly improved algorithms for latent position estimation; established the accuracy of and efficiently implemented a fast, successfully scaleable program for an approximate solution to the NP-hard problem of matching graphs; developed efficient methods for vertex nomination in graphs; and determined precisely how to mitigate information loss across shuffled networks. This has lead to dozens of papers published in top journals. Moreover, we have employed these theoretically-justified techniques on a suite of applications, conducting end-to-end analyses of real data from domains as varied as neuroscience, speech and language processing, threat detection, and social networks.

The grant period was 9/10/2012 - 3/9/2017. In addition to PI Priebe, Drs. Minh Tang, Avanti Athreya, Donniell Fishkind, Vince Lyzinski, and Nam Lee worked on the theory and methods development; Drs. Randal Burns and Alex Szalay worked on the computational aspects; and via a subcontract to Harvard Dr. Edo Airoldi assisted with the software development. Graduate students involved include Daniel Sussman (PhD dissertation defended December 2013), Sancar Adali (PhD dissertation defended March 2014), Lee Chen (PhD dissertation defended March 2015), Heng Wang (PhD dissertation defended December 2015), Jordan Yoder (PhD dissertation defended March 2016), Jason Matterer, Runze Tang, Heather Patsolic, Joshua Cape, and Mingyue Gao. In addition to PI Priebe, Sancar Adali (2013), Runze Tang (2014), Jason Matterer (2015), and Keith Levin (2016 and 2017) spearheaded our participation in the summer camps and hackathons.

2 INTRODUCTION

Graph embedding – namely, the representation of a graph in a suitably low-dimensional Euclidean space – allows the full arsenal of statistical and machine learning methodology for multivariate Euclidean data to be deployed for graph inference. We have developed and used graph embeddings to design provably accurate statistical methods and scaleable, implementable algorithms for successful large-graph inference in the context of various random graph models, ranging from the fairly simple stochastic blockmodel to the very general latent position random graph.

For the stochastic blockmodel, with a finite number of blocks of stochastically equivalent vertices, [1] and [2] show that clustering the embedded points using k-means accurately partitions the vertices into the correct blocks, even when the embedding dimension is misspecified or the number of blocks is unknown. Furthermore, [3] gives a significant improvement over these results, by exhibiting an almost-surely perfect clustering in which, in the limit, no vertices whatsoever are

misclassified. For the more general random dot product graph model, [4] shows that the latent positions are consistently estimated by the embedding, which then allows for accurate learning in a supervised vertex classification framework. [5] strengthen these results to more general latent position models, establishing a surprising and powerful universal consistency result for vertex classification of general latent position graphs, and also exhibiting an efficient embedding of vertices which were not observed in the original graph. [6] and [7] provide distributional results, akin to a central limit theorem, for both the adjacency and Laplacian spectral embedding, respectively; the former leads to a nontrivially superior algorithm for the estimation of block memberships [8] and the latter resolves, through an elegant comparison of Chernoff information, a long-standing open question of the relative merits of the adjacency and Laplacian representations. Moreover, graph embedding plays a central role in foundational work of [9, 10] on two-sample graph comparison: these works give the first – and to date, only – theoretically-justified valid and consistent hypothesis tests for the semiparametric problem of determining whether two random dot product graphs have the same latent positions and the nonparametric problem of determining whether two random dot product graphs have the same underlying distributions. This, then, yields a systematic framework for determining statistical similarity across graphs, which in turn underpins yet another provably consistent – and again, among the first of its kind – algorithm for the decomposition of random graphs with hierarchical structure [11]. Moreover, concentration results of [9] have implications for optimal error tolerance in numerical eigendecomposition algorithms, allowing users to streamline computation time [12].

For another perspective on network similarity, [13, 14] demonstrates the accuracy and scaleability of an approximate solution to the NP-hard problem of matching graphs, and this is currently being leveraged to create a consistent algorithm for simultaneous large-graph decomposition and model selection for repeated-motif hierarchical stochastic block models. In addition, [15] details precisely how to exploit graph matching to mitigate information loss in a network in which vertex labels are shuffled.

For the critical inference task of vertex nomination, [16] introduces a suite of principled vertex nomination algorithms – the canonical, maximum likelihood and spectral vertex nomination schemes – and demonstrates the algorithms' effectiveness on both synthetic and real data, including on a application to a real data set from DARPA's MEMEX program. In [17] the consistency of the maximum likelihood vertex nomination scheme is established, a scalable restricted version of the algorithm is introduced, and the algorithms are adapted to incorporate general vertex features. A scalable version of the canonical vertex nomination scheme and a novel semi-supervised Gaussian Mixture Model (GMM) based spectral vertex nomination scheme are under development (manuscript in preparation). The algorithms are developed to scale to very large graphs and demonstrate excellent performance with few training examples.

Overall, our library of techniques for graph inference, graph comparison, vertex nomination, and information recovery have been implemented in the user-friendly software packages `http://igraph.org` and `http://flashx.io`. Our theoretical foundation and computational execution together form a comprehensive, widely-applicable paradigm for statistical graph inference.

3 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 Spectral Embedding of Graphs

Vertex clustering in a stochastic blockmodel graph has wide applicability and has been the subject of extensive research. In [3], "Perfect Clustering for Stochastic Blockmodel Graphs via Adjacency Spectral Embedding," we provide a short proof that the adjacency spectral embedding can be used to obtain perfect clustering for the stochastic blockmodel and the degree-corrected stochastic blockmodel. We also show an analogous result for the more general random dot product graph model.

[6] provides a limit theorem for scaled eigenvectors of random dot product graphs. We prove a central limit theorem for the components of the largest eigenvectors of the adjacency matrix of a finite-dimensional random dot product graph whose true latent positions are unknown. In particular, we follow the methodology outlined in [1] to construct consistent estimates for the latent positions, and we show that the appropriately scaled differences between the estimated and true latent positions converge to a mixture of Gaussian random variables. As a corollary, we obtain a central limit theorem for the first eigenvector of the adjacency matrix of an Erdös-Renyi random graph.

This work is followed up with the new and important paper "Limit theorems for eigenvectors of the normalized Laplacian for random graphs" [7], wherein we prove and compare and contrast central limit theorems for adjacency spectral embedding vs. Laplacian spectral embedding.

The twin papers [9] and [10] present theory and methods for two-sample hypothesis testing for random dot product graphs. Two-sample hypothesis testing for random graphs arises naturally in neuroscience, social networks, and machine learning. In [9] we consider a semiparametric problem of two-sample hypothesis testing for a class of latent position random graphs. We formulate a notion of consistency in this context and propose a valid test for the hypothesis that two finite-dimensional random dot product graphs on a common vertex set have the same generating latent positions or have generating latent positions that are scaled or diagonal transformations of one another. Our test statistic is a function of a spectral decomposition of the adjacency matrix for each graph and our test procedure is consistent across a broad range of alternatives. We apply our test procedure to real biological data: in a test-retest data set of neural connectome graphs, we are able to distinguish between scans from different subjects; and in the *C.elegans* connectome, we are able to distinguish between chemical and electrical networks. The latter example is a concrete demonstration that our test can have power even for small sample sizes. We conclude by discussing the relationship between our test procedure and generalized likelihood ratio tests. In [10] we consider the problem of testing whether two finite-dimensional random dot product graphs have generating latent positions that are independently drawn from the same distribution, or distributions that are related via scaling or projection. We propose a test statistic that is a kernel-based function of the adjacency spectral embedding for each graph. We obtain a limiting distribution for our test statistic under the null hypothesis and we show that our test procedure is

consistent across a broad range of alternatives.

Our spectral embedding of graphs work culminates with the masterpiece "Community Detection and Classification in Hierarchical Stochastic Blockmodels," [11], wherein we propose a robust, scalable, integrated methodology for community detection and community comparison in graphs. In our procedure, we first embed a graph into an appropriate Euclidean space to obtain a low-dimensional representation, and then cluster the vertices into communities. We next employ non-parametric graph inference techniques to identify structural similarity among these communities. These two steps are then applied recursively on the communities, allowing us to detect more fine-grained structure. We describe a hierarchical stochastic blockmodel—namely, a stochastic blockmodel with a natural hierarchical structure—and establish conditions under which our algorithm yields consistent estimates of model parameters and motifs, which we define to be stochastically similar groups of subgraphs. Finally, we demonstrate the effectiveness of our algorithm in both simulated and real data. Specifically, we address the problem of locating similar subcommunities in a partially reconstructed Drosophila connectome and in the social network Friendster.

3.1.1 Vertex Classification

In [4], "Consistent Latent Position Estimation and Vertex Classification for Random Dot Product Graphs," we show that, using the eigen-decomposition of the adjacency matrix, we can consistently estimate latent positions for random dot product graphs provided that the latent positions are independent and identically distributed. If class labels are observed for a number of vertices tending to infinity, then we show that the remaining vertices can be classified with error converging to Bayes optimal using the k-nearest-neighbors classification rule. We evaluate the proposed methods on simulated data and a graph derived from Wikipedia.

In [5], "Universally consistent vertex classification for latent positions graphs," we show that, using the eigen-decomposition of the adjacency matrix, we can consistently estimate feature maps for latent position graphs with a positive definite link function, again provided that the latent positions are independent and identically distributed. We then consider the exploitation task of vertex classification where the link function belongs to the class of universal kernels and class labels are observed for a number of vertices tending to infinity and that the remaining vertices are to be classified. We show that minimization of the empirical risk for some convex surrogate of 0-1 loss over a class of linear classifiers with increasing complexities yields a universally consistent classifier, that is, a classification rule with error converging to Bayes optimal for any distribution.

For random graphs distributed according to stochastic blockmodels, a special case of latent position graphs, adjacency spectral embedding followed by appropriate vertex classification is asymptotically Bayes optimal; but this approach requires knowledge of and critically depends on the model dimension. In "Robust Vertex Classification" [18], we propose a sparse representation vertex classifier which does not require information about the model dimension. This classifier represents a test vertex as a sparse combination of the vertices in the training set and uses the recovered coefficients to classify the test vertex. We prove consistency of our proposed classifier for stochastic

blockmodels, and demonstrate that the sparse representation classifier can predict vertex labels with higher accuracy than adjacency spectral embedding approaches via both simulation studies and real data experiments. Our results demonstrate the robustness and effectiveness of our proposed vertex classifier when the model dimension is unknown.

3.1.2 Errorful

Manifold learning and dimensionality reduction techniques are ubiquitous in science and engineering, but can be computationally expensive procedures when applied to large data sets or when similarities are expensive to compute. To date, little work has been done to investigate the trade-off between computational resources and the quality of learned representations. In [19] and [20] we consider statistical inference on errorfully observed graphs, we present both theoretical and experimental explorations of this question.

We studied this problem by formulating a quantity/quality tradeoff for a simple class of random graphs model, namely the stochastic blockmodel. We then consider a simple but optimal vertex classifier and we derive the optimal quantity/quality operating point for subsequent graph inference in the face of this trade-off. The optimal operating points for the quantity/quality trade-off are surprising and illustrate the issue that methods for intermediate tasks should be chosen to maximize performance for the ultimate inference task. Finally, we investigate the quantity/quality tradeoff for errorful observations of the *C. elegans* connectome graph.

In [20] we consider Laplacian eigenmaps embeddings based on a kernel matrix, and explore how the embeddings behave when this kernel matrix is corrupted by occlusion and noise. Our main theoretical result shows that under modest noise and occlusion assumptions, we can (with high probability) recover a good approximation to the Laplacian eigenmaps embedding based on the uncorrupted kernel matrix. Our results also show how regularization can aid this approximation. Experimentally, we explore the effects of noise and occlusion on Laplacian eigenmaps embeddings of two real-world data sets, one from speech processing and one from neuroscience, as well as a synthetic data set.

3.1.3 Empirical Bayes

We employ our new adjacency spectral embedding theory – in particular, a random dot product latent position graph formulation of the stochastic blockmodel informs a mixture of normal distributions for the adjacency spectral embedding – in [8], "Empirical Bayes Estimation for the Stochastic Blockmodel," to provide an empirical Bayes methodology for estimation of block memberships of vertices in a random graph drawn from the stochastic blockmodel, and demonstrate its practical utility. The posterior inference is conducted using a Metropolis-within-Gibbs algorithm. The theory and methods are illustrated through Monte Carlo simulation studies, both within the stochastic blockmodel and beyond, and experimental results on a Wikipedia data set are presented.

3.2 (Seeded) Graph Matching

Given two graphs on the same number of vertices, the graph matching problem considered in [21] is to find a bijection between the two vertex sets which minimizes the number of adjacency disagreements between the two graphs. The seeded graph matching problem is the graph matching problem with an additional constraint that the bijection assigns some particular vertices of one vertex set to respective particular vertices of the other vertex set. Solving the (seeded) graph matching problem will enable methodologies for many graph inference tasks, but the problem is NP-hard. We modify the state-of-the-art approximate graph matching algorithm to make it a fast approximate seeded graph matching algorithm. We demonstrate the effectiveness of our algorithm - and the potential for dramatic performance improvement from incorporating just a few seeds - via simulation and real data experiments.

In the 2014 *Journal of Machine Learning Research* paper "Seeded graph matching for correlated Erdős-Rényi graphs" [22] we present theoretical and practical results on the consistency of graph matching for estimating a latent alignment function between the vertex sets of two graphs, as well as subsequent algorithmic implications when the latent alignment is partially observed. In the correlated Erdős-Rényi graph setting, we prove that graph matching provides a strongly consistent estimate of the latent alignment in the presence of even modest correlation. We then investigate a tractable, restricted-focus version of graph matching, which is only concerned with adjacency involving vertices in a partial observation of the latent alignment; we prove that a logarithmic number of vertices whose alignment is known is sufficient for this restricted-focus version of graph matching to yield a strongly consistent estimate of the latent alignment of the remaining vertices. We show how Frank-Wolfe methodology for approximate graph matching, when there is a partially observed latent alignment, inherently incorporates this restricted focus graph matching. Lastly, we illustrate the relationship between seeded graph matching and restricted-focus graph matching by means of an illuminating example from human connectomics.

We follow this up with "Graph Matching: Relax at Your Own Risk" [14], wherein we prove that an indefinite relaxation (when solved exactly) almost always discovers the optimal permutation, while a common convex relaxation almost always fails to discover the optimal permutation. These theoretical results suggest that initializing the indefinite algorithm with the convex optimum might yield improved practical performance. Indeed, experimental results illuminate and corroborate these theoretical findings, demonstrating that excellent results are achieved in both benchmark and real data problems by amalgamating the two approaches.

In [13] "Spectral Clustering for Divide-and-Conquer Graph Matching" we present a parallelized bijective graph matching algorithm that leverages seeds and is designed to match very large graphs. Our algorithm combines spectral graph embedding with existing state-of-the-art seeded graph matching procedures. We justify our approach by proving that modestly correlated, large stochastic block model random graphs are correctly matched utilizing very few seeds through our divide-and-conquer procedure. We also demonstrate the effectiveness of our approach in matching very large graphs in simulated and real data examples, showing up to a factor of 8 improvement in runtime with minimal sacrifice in accuracy.

3.3 Vertex Nomination (and Classification)

In our first major vertex nomination effort [16], "Vertex Nomination Schemes for Membership Prediction," we suppose that a graph is realized from a stochastic block model where one of the blocks is of interest, but many or all of the vertices' block labels are unobserved. The task is to order the vertices with unobserved block labels into a "nomination list" such that, with high probability, vertices from the interesting block are concentrated near the list's beginning. We propose several vertex nomination schemes. Our basic - but principled - setting and development yields a best nomination scheme (which is a Bayes-Optimal analogue), and also a likelihood maximization nomination scheme that is practical to implement when there are a thousand vertices, and which is empirically near-optimal when the number of vertices is small enough to allow comparison to the best nomination scheme. We then illustrate the robustness of the likelihood maximization nomination scheme to the modeling challenges inherent in real data, using examples which include a DARPA MEMEX social network involving human trafficking, the Enron Graph, a worm brain connectome and a political blog network.

In [17] we consider a graph in which a few vertices are deemed interesting a priori. The vertex nomination task is to order the remaining vertices into a nomination list such that there is a concentration of interesting vertices at the top of the list. Previous work has yielded several approaches to this problem, with theoretical results in the setting where the graph is drawn from a stochastic block model (SBM), including a vertex nomination analogue of the Bayes optimal classifier. In this follow-on effort, we prove that maximum likelihood (ML)-based vertex nomination is consistent, in the sense that the performance of the ML-based scheme asymptotically matches that of the Bayes optimal scheme. We prove theorems of this form both when model parameters are known and unknown. Additionally, we introduce and prove consistency of a related, more scalable restricted-focus ML vertex nomination scheme. Finally, we incorporate vertex and edge features into ML-based vertex nomination and briefly explore the empirical effectiveness of this approach.

We present "A Comparison of Graph Embedding Methods for Vertex Nomination" in [23]. Given an attributed graph representation of data, vertex nomination works to find the group of vertices which are of interest, e.g., those vertices whose attributes are different from others', or the connection among those vertices are more frequent. We present an algorithm to estimate the power of nominating these interesting vertices. This algorithm is based on Wilcoxon rank sum test. It requires to embed graph vertices into a low dimensional space. Two graph embedding methods, adjacency spectral embedding and multidimensional scaling composed with canonical correlation analysis are employed. We investigate a case where two graphs are available for modeling the same objects in different spaces, and show the effects of data fusion on vertex nomination power.

3.4 Scan Statistics

In [24], "Locality statistics for anomaly detection in time series of graphs," we consider the problem of detecting change-points in a dynamic network or a time series of graphs, an increasingly

important task in many applications of the emerging discipline of graph signal processing. We formulate change-point detection as a hypothesis testing problem in terms of a generative latent position model, focusing on the special case of the Stochastic Block Model time series. We analyze two classes of scan statistics, based on distinct underlying locality statistics presented in the literature. Our main contribution is the derivation of the limiting distributions and power characteristics of the competing scan statistics. Performance is compared theoretically, on synthetic data, and on the Enron email corpus. We demonstrate that both statistics are admissible in one simple setting, while one of the statistics is inadmissible in a second setting.

[25] considers scan statistics and demonstrates the successful detection of known but also more hidden events of neuronal activities as well as motor response associated signatures in three-dimensional time-series of functional calcium intensity data. We validate our discoveries using persistence analysis by varying a set of in-put parameters.

[26] considers the canonical problem in graph mining of the detection of dense communities. This problem is exacerbated for a graph with a large order and size – the number of vertices and edges – as many community detection algorithms scale poorly. In this work we propose a novel framework for detecting active communities that consist of the most active vertices in massive graphs. The framework is applicable to graphs having billions of vertices and hundreds of billions of edges. Our framework utilizes a parallelizable trimming algorithm based on a locality statistic to filter out inactive vertices, and then clusters the remaining active vertices via spectral decomposition on their similarity matrix. We demonstrate the validity of our method with synthetic Stochastic Block Model graphs, using Adjusted Rand Index as the performance metric. We further demonstrate its practicality and efficiency on a most recent real-world Hyperlink Web graph consisting of over 3.5 billion vertices and 128 billion edges.

3.5 Joint Optimization of Fidelity and Commensurability

In various data settings, it is necessary to compare observations from disparate data sources. In "Fidelity-Commensurability Tradeoff in Joint Embedding of Disparate Dissimilarities" [27], we assume the data is in the dissimilarity representation and investigate a joint embedding method that results in a commensurate representation of disparate dissimilarities. We further assume that there are matched observations from different conditions which can be considered to be highly similar, for the sake of inference. The joint embedding results in the joint optimization of fidelity (preservation of within-condition dissimilarities) and commensurability (preservation of between-condition dissimilarities between matched observations). We show that the tradeoff between these two criteria can be made explicit using weighted raw stress as the objective function for multidimensional scaling. In our investigations, we use a weight parameter to control the tradeoff, and choose match detection as the inference task. Our results show weights that are optimal (with respect to the inference task) are different than equal weights for commensurability and fidelity and the proposed weighted embedding scheme provides significant improvements in statistical power.

The Joint Optimization of Fidelity and Commensurability (JOFC) manifold matching methodology

embeds an omnibus dissimilarity matrix consisting of multiple dissimilarities on the same set of objects. One approach to this embedding optimizes the preservation of fidelity to each individual dissimilarity matrix together with commensurability of each given observation across modalities via iterative majorization of a raw stress error criterion by successive Guttman transforms. In "Fast Embedding for JOFC Using the Raw Stress Criterion" [28] we exploit the special structure inherent to JOFC to exactly and efficiently compute the successive Guttman transforms, and as a result we are able to greatly speed up the JOFC procedure for both in-sample and out-of-sample embedding. We demonstrate the scalability of our implementation on both real and simulated data examples.

In [29], "Seeded Graph Matching Via Joint Optimization of Fidelity and Commensurability," we present a novel approximate graph matching algorithm that incorporates seeded data into the graph matching paradigm. Our Joint Optimization of Fidelity and Commensurability (JOFC) algorithm embeds two graphs into a common Euclidean space where the matching inference task can be performed. Through real and simulated data examples, we demonstrate the versatility of our algorithm in matching graphs with various characteristics–weightedness, directedness, loopiness, many-to-one and many-to-many matchings, and soft seedings.

3.6 The Incommensurability Phenomenon

In [30], "On the Incommensurability Phenomenon," we suppose that two large, multi-dimensional data sets are each noisy measurements of the same underlying random process, and principle components analysis is performed separately on the data sets to reduce their dimensionality. In some circumstances it may happen that the two lower-dimensional data sets have an inordinately large Procrustean fitting-error between them. We quantify this "incommensurability phenomenon." In particular, under specified conditions, the square Procrustean fitting-error of the two normalized lower-dimensional data sets is (asymptotically) a convex combination (via a correlation parameter) of the Hausdorff distance between the projection subspaces and the maximum possible value of the square Procrustean fitting-error for normalized data. We show how this gives rise to the incommensurability phenomenon, and we employ illustrative simulations as well as a real data experiment to explore how the incommensurability phenomenon may have an appreciable impact.

3.7 Semi-Supervised Clustering Methodology

In [31], "A Model-based Semi-Supervised Clustering Methodology," we consider an extension of model-based clustering to the semi-supervised case, where some of the data are pre-labeled. We provide a derivation of the Bayesian Information Criterion (BIC) approximation to the Bayes factor in this setting. We then use the BIC to the select number of clusters and the variables useful for clustering. We demonstrate the efficacy of this adaptation of the model-based clustering paradigm through two simulation examples and a fly larvae behavioral dataset in which lines of neurons are clustered into behavioral groups.

3.8 Robust Hypothesis Testing

In [32], "Robust Hypothesis Testing via Lq-Likelihood," we introduce a robust hypothesis testing procedure: the Lq-likelihood-ratio-type test (LqRT). By deriving the asymptotic distribution of this test statistic, we demonstrate its robustness both analytically and numerically, and investigate the properties of both its influence function and its breakdown point. A proposed method to select the tuning parameter offers a good efficiency/robustness trade-off, compared with the traditional likelihood ratio test (LRT) and other robust tests. A simulation and real data analysis provides further evidence of the advantages of the proposed LqRT method. In particular, for the special case of testing the location parameter in the presence of gross error contamination, the LqRT dominates the Wilcoxon-Mann-Whitney test and the sign test at various levels of contamination.

3.9 Model Selection

In [33] "A model selection approach for clustering a multinomial sequence with non-negative factorization," we consider a problem of clustering a sequence of multinomial observations by way of a model selection criterion. We propose a form of a penalty term for the model selection procedure. Our approach subsumes both the conventional Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) but also extends the conventional criteria in a way that it can be applicable also to a sequence of sparse multinomial observations, where even within a same cluster, the number of multinomial trials may be different for different observations. In addition, as a preliminary estimation step to maximum likelihood estimation, and more generally, to maximum Lq estimation, we propose to use reduced rank projection in combination with non-negative factorization. We motivate our approach by showing that our model selection criterion and preliminary estimation step yield consistent estimates under simplifying assumptions. We also illustrate our approach through numerical experiments using real and simulated data.

3.10 Sparse Representation Classification

In [34], "Sparse Representation Classification Beyond L1 Minimization and the Subspace Assumption," we consider the sparse representation classifier (SRC) proposed in Wright et al. [35], which has recently gained much attention from the machine learning community. It makes use of L1 minimization, and is known to work well for data satisfying a subspace assumption. We use the notion of class dominance as well as a principal angle condition to investigate and validate the classification performance of SRC, without relying on L1 minimization and the subspace assumption. We prove that SRC can still work well using faster subset regression methods such as orthogonal matching pursuit and marginal regression, and its applicability is not limited to data satisfying the subspace assumption. We illustrate our theorems via various real data sets including face images, text features, and network data.

3.11 Manifold Matching

In [36], "Manifold Matching using Shortest-Path Distance and Joint Neighborhood Selection," we propose a nonlinear manifold matching algorithm to match multiple data sets using shortest-path distance and joint neighborhood selection. Based on the correspondence information, a neighborhood graph is jointly constructed; then the shortest-path distance within each data set is computed from the joint neighborhood graph, followed by embedding into and matching in a common low-dimensional Euclidean space. Our approach exhibits superior and robust performance for matching data from disparate sources, compared to algorithms that do not use shortest-path distance or joint neighborhood selection.

3.12 A Joint Graph Inference Case Study

We present a novel and illustrative joint graph inference case study in [37] which incorporates many of our theories and methods. We investigate joint graph inference for the chemical and electrical connectomes of the Caenorhabditis elegans roundworm. The C.elegans connectomes consist of 253 nonisolated neurons with known functional attributes, and there are two types of synaptic connectomes, resulting in a pair of graphs. We formulate our joint graph inference from the perspectives of seeded graph matching and joint vertex classification. Our results suggest that connectomic inference should proceed in the joint space of the two connectomes, which has significant neuroscientific implications.

3.13 Science

In [38], "Discovery of Brainwide Neural-Behavioral Maps via Multiscale Unsupervised Structure Learning," with our collaborators at the Howard Hughes Medical Institute's Janelia Research Campus, we present a seminal neuroscientific development. A single nervous system can generate many distinct motor patterns. Identifying which neurons and circuits control which behaviors has been a laborious piecemeal process, usually for one observer-defined behavior at a time. We present a fundamentally different approach to neuron-behavior mapping. We optogenetically activated 1,054 identified neuron lines in Drosophila larva and tracked the behavioral responses from 37,780 animals. Applying multiscale unsupervised structure learning methods to the behavioral data identified 29 discrete statistically distinguishable and observer-unbiased behavioral phenotypes. Mapping the neural lines to the behavior(s) they evoke provides a behavioral reference atlas for neuron subsets covering a large fraction of larval neurons. This atlas is a starting point for connectivity- and activity-mapping studies to further investigate the mechanisms by which neurons mediate diverse behaviors.

4 RESULTS AND DISCUSSION

We have developed a multitude of theory and methods for spectral graph embedding and graph matching, and for subsequent inference based on these principled transformations. In particular, the important and timely vertex nomination application is now on firm theoretical and methodological footing.

Graph embedding – namely, the representation of a graph on n vertices as n points in a suitably low-dimensional Euclidean space – allows the full arsenal of statistical and machine learning methodology for multivariate Euclidean data to be deployed for graph inference. We have recently shown that neither of the two main embedding methods – Laplacian spectral embedding vs. adjacency spectral embedding – dominates the other for subsequent inference, indicating that optimal graph inference will always involve challenging empirical modeling and implementation issues. Our project has propelled the mathematical development, statistical design, and computational implementation of graph embeddings to provide provably accurate subsequent statistical inference via scaleable, implementable algorithms for successful large-graph inference in the context of various random graph models, ranging from fairly simple stochastic blockmodels to general latent position random graphs … and beyond, to real data.

5 CONCLUSIONS

Graph inference is a burgeoning field in the applied and theoretical statistics communities, as well as throughout the wider world of science, engineering, business, etc. This DARPA XDATA project has provided DoD, and the wider statistical inference community, the wherewithal to proceed apace in the analysis of complex, graph-valued data, positively impacting important programs reliant on the processing of such data.

6 REFERENCES

[1] Sussman, D. L., Tang, M., Fishkind, D. E., and Priebe, C. E., "A consistent adjacency spectral embedding for stochastic blockmodel graphs," *Journal of the American Statistical Association*, **107**(499):1119–1128, Dec. 2012, 1108.2228v3.

[2] Fishkind, D. E., Sussman, D. L., Tang, M., Vogelstein, J. T., and Priebe, C. E., "Consistent Adjacency-Spectral Partitioning for the Stochastic Block Model When the Model Parameters Are Unknown," *SIAM Journal on Matrix Analysis and Applications*, **34**(1):23–39, 2013.

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

[4] Sussman, D. L., Tang, M., and Priebe, C. E., "Consistent Latent Position Estimation and Vertex Classification for Random Dot Product Graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**(1):48–57, Jan. 2014.

[5] Tang, M., Sussman, D. L., and Priebe, C. E., "Universally Consistent Vertex Classification for Latent Positions Graphs," *Annals of Statistics*, **41**(3):1406–1430, June 2013.

[6] Athreya, A., Priebe, C. E., Tang, M., Lyzinski, V., Marchette, D. J., and Sussman, D. L., "A limit theorem for scaled eigenvectors of random dot product graphs," *Sankhya A. Mathematical Statistics and Probability*, **78**(1):1–18, 2016.

[7] Tang, M. and Priebe, C. E., "Limit theorems for eigenvectors of the normalized Laplacian for random graphs," *arXiv.org*, July 2016, 1607.08601v1.

[8] Suwan, S., Lee, D. S., Tang, R., Sussman, D. L., Tang, M., and Priebe, C. E., "Empirical Bayes estimation for the stochastic blockmodel," *Electronic Journal of Statistics*, **10**(1):761–782, 2016.

[9] Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., and Priebe, C. E., "A semiparametric two-sample hypothesis testing problem for random dot product graphs," *Journal of Computational and Graphical Statistics*, accepted for publication 2016, 1403.7249v3.

[10] Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., and Priebe, C. E., "A nonparametric two-sample hypothesis testing problem for random dot product graphs," *Bernoulli Journal*, accepted for publication 2015, 1409.2344v2.

[11] Lyzinski, V., Tang, M., Athreya, A., Park, Y., and Priebe, C. E., "Community Detection and Classification in Hierarchical Stochastic Blockmodels," *IEEE Transactions on Network Science and Engineering*, accepted for publication 2016, 1503.02115v5.

[12] Athreya, A., Tang, M., Lyzinski, V., Park, Y., Lewis, B., Kane, M., and Priebe, C., "Numerical tolerance for spectral decompositions of random dot product graphs," *arXiv.org*, Aug. 2016, 1608.00451v1.

[13] Lyzinski, V., Sussman, D. L., Fishkind, D. E., Pao, H., Chen, L., Vogelstein, J. T., Park, Y., and Priebe, C. E., "Spectral clustering for divide-and-conquer graph matching," *Parallel Computing*, **47**:70–87, Aug. 2015.

[14] Lyzinski, V., Fishkind, D. E., Fiori, M., Vogelstein, J. T., Priebe, C. E., and Sapiro, G., "Graph Matching: Relax at Your Own Risk.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(1):60–73, Jan. 2016.

[15] Lyzinski, V., "Information Recovery in Shuffled Graphs via Graph Matching," *arXiv.org*, May 2016, 1605.02315v1.

[16] Fishkind, D. E., Lyzinski, V., Pao, H., Chen, L., and Priebe, C. E., "Vertex Nomination Schemes for Membership Prediction," *Annals of Applied Statistics*, **9**(3):1510–1532, Sept. 2015.

[17] Lyzinski, V., Levin, K., Fishkind, D. E., and Priebe, C. E., "On the Consistency of the Likelihood Maximization Vertex Nomination Scheme: Bridging the Gap Between Maximum Likelihood Estimation and Graph Matching," *Journal of Machine Learning Research*, **17**(179):1–34, 2016.

[18] Chen, L., Shen, C., Vogelstein, J., and Priebe, C. E., "Robust Vertex Classification.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(3):1–1, July 2015.

[19] Priebe, C. E., Sussman, D. L., Tang, M., and Vogelstein, J. T., "Statistical Inference on Errorfully Observed Graphs," *Journal of Computational and Graphical Statistics*, **24**(4):930–953, 2015.

[20] Levin, K. and Lyzinski, V., "Laplacian Eigenmaps From Sparse, Noisy Similarity Measurements.," *IEEE Transactions on Signal Processing*, **65**(8):1988–2003, 2017.

[21] Fishkind, D. E., Adali, S., and Priebe, C. E., "Seeded graph matching," *arXiv.org*, 2012. [22]

Lyzinski, V., Fishkind, D. E., and Priebe, C. E., "Seeded Graph Matching for Correlated Erdos-Renyi Graphs," *Journal of Machine Learning Research*, **15**:3513–3540, Nov. 2014.

[23] Sun, M., Tang, M., and Priebe, C., "A comparison of graph embedding methods for vertex nomination," in *2012 11th International Conference on Machine Learning and Applications (ICMLA)*, Boca Raton, FL, 2012.

[24] Wang, H., Tang, M., Park, Y., and Priebe, C. E., "Locality Statistics for Anomaly Detection in Time Series of Graphs," *IEEE Transactions on Signal Processing*, **62**(3):703–717, Feb. 2014.

[25] Park, Y., Wang, H., Nöbauer, T., Vaziri, A., and Priebe, C. E., "Anomaly detection on whole-brain functional imaging of neuronal activity using graph scan statistics," in *21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-2015), Workshop on Outlier Definition, Detection, and Description (ODDx3)*, Sydney, Australia, August 2015.

[26] Wang, H., Zheng, D., Burns, R., and Priebe, C. E., "Active community detection in massive graphs," in *SDM-Networks 2015: The Second SDM Workshop on Mining Networks and Graphs: A Big Data Analytic Challenge*, Vancouver, BC, Canada, April 2015.

[27] Adali, S. and Priebe, C. E., "Fidelity-Commensurability Tradeoff in Joint Embedding of Disparate Dissimilarities," *Journal of Classification*, **33**(3):485–506, Oct. 2016.

[28] Lyzinski, V., Park, Y., Priebe, C. E., and Trosset, M. W., "Fast Embedding for JOFC Using the Raw Stress Criterion," *arXiv.org*, Feb. 2015, 1502.03391v3.

[29] Lyzinski, V., Adali, S., Vogelstein, J. T., Park, Y., and Priebe, C. E., "Seeded Graph Matching Via Joint Optimization of Fidelity and Commensurability," *arXiv.org*, Jan. 2014, 1401.3813v1.

[30] Fishkind, D. E., Shen, C., Park, Y., and Priebe, C. E., "On the Incommensurability Phenomenon," *Journal of Classification*, **33**(2):185–209, July 2016.

[31] Yoder, J. and Priebe, C. E., "A Model-based Semi-Supervised Clustering Methodology," *arXiv.org*, Dec. 2014, 1412.4841v2.

[32] Qin, Y. and Priebe, C. E., "Robust Hypothesis Testing via Lq-Likelihood," *Statistica Sinica*, accepted for publication 2016, 1310.7278v3.

[33] Lee, N., Tang, R., Priebe, C., and Rosen, M., "A model selection approach for clustering a multinomial sequence with non-negative factorization.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(12):2345–2358, Jan. 2016.

[34] Shen, C., Chen, L., and Priebe, C. E., "Sparse Representation Classification Beyond L1 Minimization and the Subspace Assumption," *arXiv.org*, Feb. 2015, 1502.01368v2.

[35] Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y., "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**(2):210–227, Feb 2009.

[36] Shen, C., Vogelstein, J. T., and Priebe, C. E., "Manifold Matching using Shortest-Path Distance and Joint Neighborhood Selection," *arXiv.org*, Dec. 2014, 1412.4098v3.

[37] Chen, L., Vogelstein, J. T., Lyzinski, V., and Priebe, C. E., "A joint graph inference case study: the C. elegans chemical and electrical connectomes," *Worm*, **5**(2):e1142041, Mar. 2016.

[38] Vogelstein, J. T., Park, Y., Ohyama, T., Kerr, R. A., Truman, J. W., Priebe, C. E., and Zlatic, M., "Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning.," *Science*, **344**(6182):386–392, Apr. 2014.

7 LIST OF ACRONYMS

Acronyms

AIC Akaike Information Criterion
BIC Bayesian Information Criterion
GMM Gaussian Mixture Model
JOFC Joint Optimization of Fidelity and Commensurability
LqRT Lq-Likelihood-Ratio-Type Test
LRT Likelihood Ratio Test
ML Maximum Likelihood
SBM Stochastic Block Model
SRC Sparse Representation Classifier