# Graph Classification Using Signal-Subgraphs: Applications in Statistical Connectomics

Joshua T. Vogelstein, William R. Gray, *Member, IEEE*,
R. Jacob Vogelstein, *Member, IEEE*, and Carey E. Priebe, *Senior Member, IEEE*

**Abstract**—This manuscript considers the following "graph classification" question: Given a collection of graphs and associated classes, how can one predict the class of a newly observed graph? To address this question, we propose a statistical model for graph/class pairs. This model naturally leads to a set of estimators to identify the class-conditional signal, or "signal-subgraph," defined as the collection of edges that are probabilistically different between the classes. The estimators admit classifiers which are asymptotically optimal and efficient, but which differ by their assumption about the "coherency" of the signal-subgraph (coherency is the extent to which the signal-edges "stick together" around a common subset of vertices). Via simulation, the best estimator is shown to be not just a function of the coherency of the model, but also the number of training samples. These estimators are employed to address a contemporary neuroscience question: Can we classify "connectomes" (brain-graphs) according to sex? The answer is yes, and significantly better than all benchmark algorithms considered. Synthetic data analysis demonstrates that even when the model is correct, given the relatively small number of training samples, the estimated signal-subgraph should be taken with a grain of salt. We conclude by discussing several possible extensions.

**Index Terms**—Statistical inference, graph theory, network theory, structural pattern recognition, connectome, classification

## 1 Introduction

Graphs are emerging as a prevalent form of data representation in fields ranging from optical character recognition and chemistry [1] to neuroscience [2]. While statistical inference techniques for vector-valued data are widespread, statistical tools for the analysis of graph-valued data are relatively rare [1]. In this work, we consider the task of *labeled graph classification*: Given a collection of labeled graphs and their corresponding classes, can we accurately infer the class for a new graph? Note that we assume throughout that each vertex has a unique label, and that all graphs have the same number of vertices with the same vertex labels. The methods developed herein, however, can straightforwardly be relaxed to more general settings.

We propose and analyze a joint graph/class model—sufficiently simple to characterize its asymptotic properties, and sufficiently rich to afford useful empirical applications. This model admits a class-conditional signal encoded in a subset of edges, the *signal-subgraph*. Finding the signal-subgraph amounts to providing an understanding of the differences between the two graph classes. Moreover, borrowing a term from the compressive sensing literature [3], [4], we are interested in learning to what extent this signal is *coherent*, that is, to what extent are the signal-subgraph edges incident to a relatively small set of vertices. In other words, if the signal is sparse in the edges, then the signal-subgraph is incoherent; if it is also sparse in the vertices, then the signal-subgraph is coherent (we formally define these notions below).

This graph-model based approach is qualitatively different from most previous approaches, which utilize only unique vertex labels or graph structure. In the former case, simply representing the adjacency matrix with a vector and applying standard machine learning techniques ignores graph structure (for instance, it is not clear how to implement a coherent signal-subgraph estimator in this representation). In the latter case, computing a set of graph invariants (such as clustering coefficient), and then classifying using only these invariants ignores vertex labels [1], [5], [6].

While some of the above approaches consider attributed vertices or edges, we are unable to find any that utilize both *unique* vertex labels and graph structure. The field of *connectomics* (the study of brain-graphs), however, is ripe with many examples of brain-graphs with vertex labels. In invertebrate brain-graphs, for example, often each neuron is named such that one can compare neurons across individuals of the same species [7]. In vertebrate neurobiology, while neurons are rarely named, "neuron types" [8] and neuroanatomical regions [9] are named. Moreover, a widely held view is that many psychiatric issues are fundamentally "connectopathies" [10], [11]. For prognostic and diagnostic purposes, merely being able to differentiate groups of brain-graphs from one another is sufficient. However, for treatment, it is desirable to know which vertices and/or edges are malfunctioning such that therapy can be targeted to those locations. This is the motivating application for our work.

- *J.T. Vogelstein is with the Department of Mathematics and Statistics, Duke University, PO Box 90251, 214 Old Chemistry, Research Drive, Durham, NC 27708-0251. E-mail: jovo@stat.duke.edu, joshuav@jhu.edu.*
- *W.R. Gray and R.J. Vogelstein are with the Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723. E-mail: {William.Gray, jacob.vogelstein}@jhuapl.edu.*
- *C.E. Priebe is with the Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218-2682. E-mail: cep@jhu.edu.*

*Manuscript received 8 Aug. 2011; revised 1 July 2012; accepted 16 Oct. 2012; published online 26 Oct. 2012.*
*Recommended for acceptance by N. Lawrence.*
*For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-08-0537.*
*Digital Object Identifier no. 10.1109/TPAMI.2012.235.*

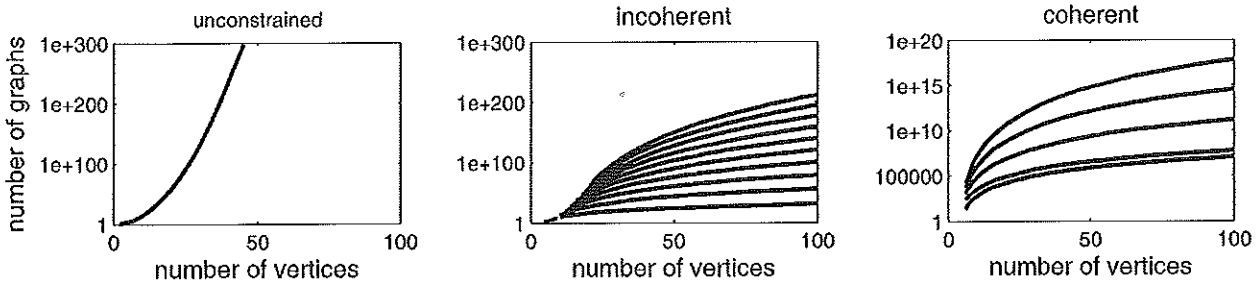


Fig. 1. Exhaustive searches for the signal-subgraph, even given severe constraints, are computationally intractable even for small graphs. The three panels illustrate the number of unique simple subgraphs as a function of the number of vertices $V$ for the three different constraint types considered: unconstrained, edge constrained, and both edge and vertex constrained (coherent). Note the ordinates are all log scale. On the left is the unconstrained scenario, that is, all possible subgraphs for a given number of vertices. In the middle panel, each line shows the number of subgraphs with a fixed number of signal-edges, $s$, ranging from 10 to 100, incrementing by 10 with each line. The right panel shows the number of subgraphs for various fixed $s$ and only a single signal-vertex, that is, all edges are incident to one vertex.

$$h_* = \arg\min_{h\in\mathcal{H}} \mathbb{E}_F[\ell_h(\mathbb{G}, Y)] = \arg\max_{y\in\mathcal{Y}} F_{\mathbb{G}|Y=y} F_{Y=y}. \quad (5)$$

Given the proposed model, (5) can be further factorized using the above four assumptions:

$$h_*(G) = \arg\max_{y\in\mathcal{Y}} \prod_{u,v\in\mathcal{S}} \text{Bern}(A_{uv}; p_{uv|y})\pi_y. \quad (6)$$

Unfortunately, Bayes optimal classifiers are typically unavailable. In such settings, it is therefore desirable to induce a classifier estimate from a set of *training data*. Formally, let $\mathcal{T}_n = \{(\mathbb{G}_i, Y_i)\}_{i\in[n]}$ denote the training corpus, where each graph-class pair is sampled exchangeably from the true but unknown distribution: $(\mathbb{G}_i, Y_i) \overset{exch}{\sim} F_{\mathbb{G},Y}$. Given such a training corpus and an unclassified graph $G$, an induced classifier predicts the true (but unknown) class of $G$, $\widehat{h} : \mathcal{G} \times (\mathcal{G} \times \mathcal{Y})^n \to \mathcal{Y}$. When a model $\mathcal{F}_{\mathbb{G},Y}$ is specified, a beloved approach is ~~to use~~ a *Bayes plugin classifier*. Due to the above simplifying assumptions, the Bayes plugin classifier for this model is defined as follows: First, ~~estimate the model~~ parameters $\widehat{\boldsymbol{\theta}} = \{\widehat{\mathcal{S}}, \widehat{\boldsymbol{p}}, \widehat{\boldsymbol{\pi}}\}$. Second, plug those estimates into the above equation. The result is a Bayes plugin graph classifier:

$$\widehat{h}(G; \mathcal{T}_n) \triangleq \arg\max_{y\in\mathcal{Y}} \prod_{u,v\in\widehat{\mathcal{S}}} \widehat{p}_{uv|y}^{a_{uv}} (1 - \widehat{p}_{uv|y})^{(1-a_{uv})} \widehat{\pi}_y, \quad (7)$$

where the Bernoulli probability is explicit. To implement such a classifier estimate, we specify estimators for $\mathcal{S}$, $\boldsymbol{\pi}$, and $\boldsymbol{p}$.

## 2.4 Estimators

### 2.4.1 Desiderata

We desire a sequence of estimators, $\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2, \ldots$, that satisfy the following five desiderata, listed in no particular order:

1. Consistent. An estimator is consistent (in some specified sense) if its sequence converges in the limit to the true value: $\lim_{n\to\infty} \widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}$.
2. Robust. An estimator is robust if the resulting estimate is relatively insensitive to small model misspecifications. Because the space of models is massive (uncountably infinite), it is intractable to consider all misspecifications, so we consider only a few of them, as described below.
3. Quadratic complexity. Computational time complexity should be no more than quadratic in the number of vertices.
4. Interpretable. We desire that the parameters are interpretable with respect to a subset of vertices and/or edges.
5. Finite sample/empirical performance. At the end of the day, we are concerned with having a classifier that works to solve our applied problems.

### 2.4.2 Signal-Subgraph Estimators

Naively, one might consider a search over all possible signal-subgraphs by plugging each one in to the classifier and selecting the best performing option. This strategy is intractable because the number of signal-subgraphs scales superexponentially with the number of vertices (see Fig. 1, left panel). Specifically, the number of possible edges in a simple graph with $V$ vertices is $d_V = \binom{V}{2}$, so the number of unique possible signal-subgraphs is $2^{\binom{V}{2}}$. Searching over all of them is sufficiently computationally taxing as to motivate the search for other alternatives.

Before proceeding, recall that each edge is independent; thus, one can evaluate each edge separately (although treating edges independently is not necessarily advisable, consider the Stein estimator [14]). Formally, consider a hypothesis test for each edge. The simple null hypothesis is that the class-conditional edge distributions are the same, so $H_0 : F_{uv|0} = F_{uv|1}$. The composite alternative hypothesis is that they differ, $H_A : F_{uv|0} \neq F_{uv|1}$. Given such hypothesis tests, one can construct test statistics $T_{uv}^{(n)} : \mathcal{T}_n \to \mathbb{R}_+$. We reject the null in favor of the alternative whenever the value of the test statistic is greater than some critical-value: $T_{uv}^{(n)}(\mathcal{T}_n) > c$. We can therefore construct a *significance matrix* $\boldsymbol{T} \triangleq T_{uv}^{(n)}$, which is the sufficient statistic for the signal-subgraph estimators. Example test statistics include Fisher's and chi-squared, which will be discussed further below. Whichever test statistic one uses, the sufficient statistics are captured in a $2 \times |\mathcal{Y}|$ contingency table, indicating the number of times edge $u, v$ was observed in each class. For example, the two-class contingency table for each edge is given by

| | Class 0 | Class 1 | Total |
|---|---|---|---|
| Edge | $n_{uv\|0}$ | $n_{uv\|1}$ | $n_{uv}$ |
| No Edge | $n_0 - n_{uv\|0}$ | $n_1 - n_{uv\|1}$ | $n - n_{uv}$ |
| Total | $n_0$ | $n_1$ | $n$. |

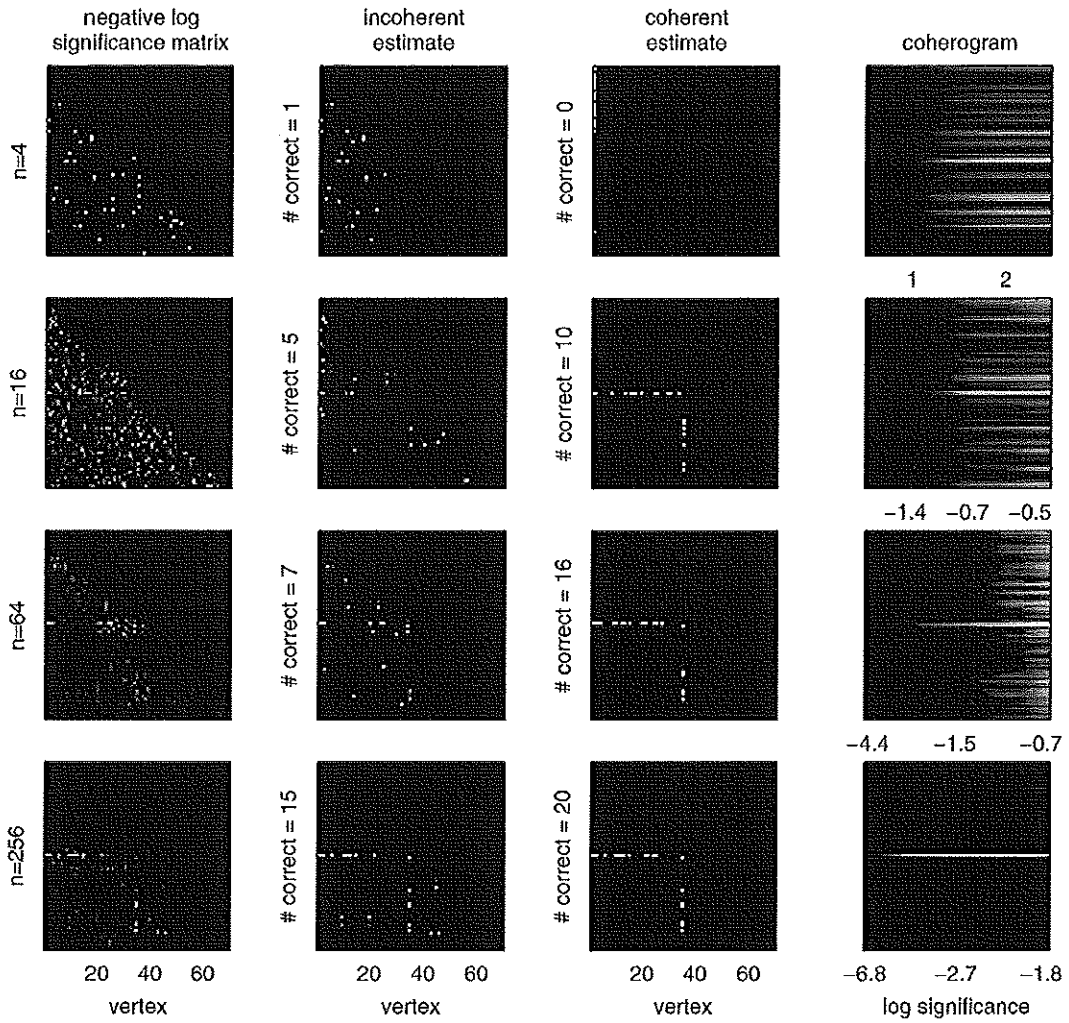


Fig. 2. An example of the coherent signal-subgraph estimate's improved accuracy over the incoherent signal-subgraph estimate for a particular homogeneous two-class model specified by: $\mathcal{M}_{70}(1, 20; 0.5, 0.1, 0.3)$. Each row shows the same columns but for increasing the number of graph/class samples. The columns show: (far left) the negative log-significant matrix, computed using Fisher's exact test (lighter means more significant; each panel is scaled independently of the others because only relative significance matters here); (middle left) the incoherent estimate of the signal-subgraph; (middle right) the coherent estimate of the signal-subgraph; (far right) the coherogram. As the number of training samples increases (lower rows), both the incoherent and coherent estimates converge to the truth (the ordinate labels of the middle panels indicate the number of edges correctly identified). For these examples, the coherent estimator tends to find more true edges. The coherogram visually depicts the coherency of the signal; it is also converging to the truth—the signal-subgraph here contains a single signal-vertex.

$$\widehat{p}_{uv|y} = \begin{cases} \eta_n, & \text{if } \max_i a_{uv}^{(i)} = 0, \\ 1 - \eta_n, & \text{if } \min_i a_{uv}^{(i)} = 1, \\ \widehat{p}_{uv|y}^{MLE}, & \text{otherwise,} \end{cases} \tag{11}$$

where we let $\eta_n = 1/(10n)$.

### 2.4.4 Prior Estimators

The priors are the simplest. The prior probabilities are Bernoulli, and we are concerned only with the case where $|\mathcal{Y}| \ll n$, so the maximum likelihood estimators suffice:

$$\widehat{\pi}_y = \frac{n_y}{n}, \tag{12}$$

where $n_y = \sum_{i \in [n]} \mathbb{I}\{y_i = y\}$.

### 2.4.5 Hyperparameter Selection

The signal-subgraph estimators require specifying the number of signal-edges $s$, as well as the number of signal-vertices $m$ for the coherent classifier. In both cases, the number of possible values of finite. In particular, $s \in [d_V]$ and $m \in [V]$. Thus, to select the best hyperparameters we implement cross-validation procedures (see Section 2.5.2 for details), iterating over $(s, m) \in \vec{s} \times \vec{m} \subseteq [d_V] \times [V]$. Note that when $m = V$, the coherent signal subgraph estimator reduces to the incoherent signal subgraph estimator. For all simulated data, we compare hyperparameter performance via a training and held-out set. For the real data application, we decided to use a leave-one-out cross-validation procedure due to the small sample size.

### 2.4.6 All Together

Putting the above pieces together, Algorithm 3 provides pseudocode for implementing our signal-subgraph classifiers. MATLAB code is available from the first author's website, http://jovo.me.

sample efficiency whenever the model does not induce too much bias, as will be shown below.

### 3.3 Bayes Plugin Classifier

Lemma 3.3. *The Bayes plug-in classifier, using the signal-subgraph, likelihood, and prior estimators described above, is consistent under the model defined by (2).*

Proof. A Bayes plugin classifier is a consistent classifier whenever the estimates that are plugged in are consistent [13]. Because the likelihood, prior, and signal-subgraph estimates are all consistent, the Bayes plugin classifier is also consistent. □

Note that naive Bayes classifiers often exhibit impressive finite sample performance due to their winning the bias-variance tradeoff relative to other classifiers [19]. In other words, even when edges are highly dependent, because marginal probability estimates are more efficient than joint probability estimates, an independent edge-based classifier will often outperform a classifier based on dependences.

## 4 Simulated Experiments

### 4.1 Simulation Details

To better assess the finite sample properties of the signal-subgraph estimators, we conduct a number of simulated experiments. Consider the following *homogeneous* model: each simple graph has $V = 70$ vertices. Class 0 graphs are Erdos-Renyi with probability $p$ for each edge; that is, $f_{uv|0} = p \; \forall \; (u, v) \in \mathcal{E}$. Class 1 graphs are a mixture of two Erdos-Renyi models: all edges in the *signal-subgraph* have probability $q$, and all others have probability $p$, so that $f_{uv|1} = q \; \forall (u, v) \in \mathcal{S}$, and $f_{uv|1} = p \; \forall (u, v) \in \mathcal{E} \backslash \mathcal{S}$. The signal-subgraph is constrained to have $m$ signal-vertices and $s$ signal-edges. Let the class-prior probabilities be given by $F_{Y=0} = \pi$ and $F_{Y=1} = 1 - \pi$. Thus, the model is characterized by $F_0 = \mathcal{M}_V(m, s; \pi, p, q)$, where $V$ is a constant, $m$ and $s$ are hyperparameters, and $\pi$, $p$, and $q$ are parameters.

### 4.2 A Simple Demonstration

To provide some insight with respect to the finite sample performance of the incoherent and coherent signal-subgraph estimators for this model, we run the following simulated experiments, with results depicted in Fig. 2. In each row we sample from $\mathcal{M}_{70}(1, 20; 0.5, 0.1, 0.3)$ (note that we are actually conditioning on the class-conditional sample size). Given these $n$ samples, we compute the significance matrix (first column), which contains the sufficient statistics for both estimators. The incoherent estimator simply chooses the $s$ most significant edges as the signal-subgraph (second column). The coherent estimator jointly estimates both the $m$ signal-vertices and the $s$ signal-edges incident to at least one of those vertices (third column). The coherogram shows the "coherency" of the data (fourth column).

From this figure, one might notice a few tendencies. First, both the incoherent and coherent signal-subgraphs seem to converge to the true signal-subgraph. Second, while both estimators perform poorly with $n < 16$, the coherent estimator converges more quickly than the incoherent estimator. Third, the coherogram sharpens with additional samples, showing after only approximately 50 samples that this model is strongly coherent.

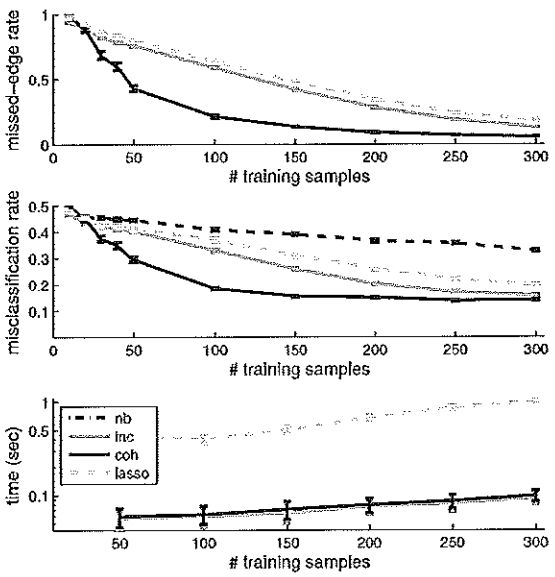


Fig. 3. Performance statistics as a function of sample size demonstrate that the coherent signal-subgraph estimator outperforms the incoherent signal-subgraph estimator in terms of both the signal-subgraph identification and classification for nearly all $n$, using the same model as in Fig. 2: $\mathcal{M}_{70}(1, 20; 0.5, 0.1, 0.3)$. Moreover, even the incoherent classifier outperforms the $\ell_1$-penalized logistic regression (lasso) on all our metrics. The top panel shows the missed-edge rate for each estimator as a function of the number of training samples, $n$. The middle panel shows the corresponding misclassification rate for the estimators, as well as the naive Bayes plugin classifier. Performance of all estimators improves (nearly) monotonically with $n$ for both criteria. The bottom panel shows total training and testing time for each classifier. Clearly, the lasso is about 10 times slower than the others. Error bars show the standard error of the mean here and elsewhere unless otherwise noted (averaged over 100 trials; each trial used 100 samples for held-out data). Error bars on the lower panel show the interquartile range. Note that for most values of $n$, we have $L_{\hat{*}} > \widehat{L}_{nb} > \widehat{L}_{lasso} > \widehat{L}_{inc} > \widehat{L}_{coh} > L_*$. Legend: "inc": incoherent; "coh": coherent; "nb": naive Bayes, "lasso": lasso.

### 4.3 Quantitative Comparisons

To better characterize the relative performance of the two signal-subgraph estimators, Fig. 3 shows their performance as a function of the number of training samples, $n$, for the $\mathcal{M}_{70}(1, 20; 0.5, 0.1, 0.3)$ model. The top panel shows the mean and standard error of the missed-edge rate—the fraction of edges incorrectly identified—averaged over 200 trials. For essentially all $n$, the coherent estimator (black solid line) performs better than the incoherent estimator (gray solid line). We also compare the performance of an $\ell_1$-penalized logistic regression classifier ("lasso" hereafter [20]). As expected, the missed edge rate for the lasso (gray dashed line) and the incoherent classifier are about the same. The improvement in signal-edge detection of the coherent signal-subgraph estimator over the incoherent's and lasso's performance translates directly to improved classification performance (middle panel), where the plugin classifier using the coherent signal-subgraph estimator has a better misclassification rate than either the incoherent signal-subgraph classifier or the lasso for essentially all $n$. Note that the incoherent classifier also admits better

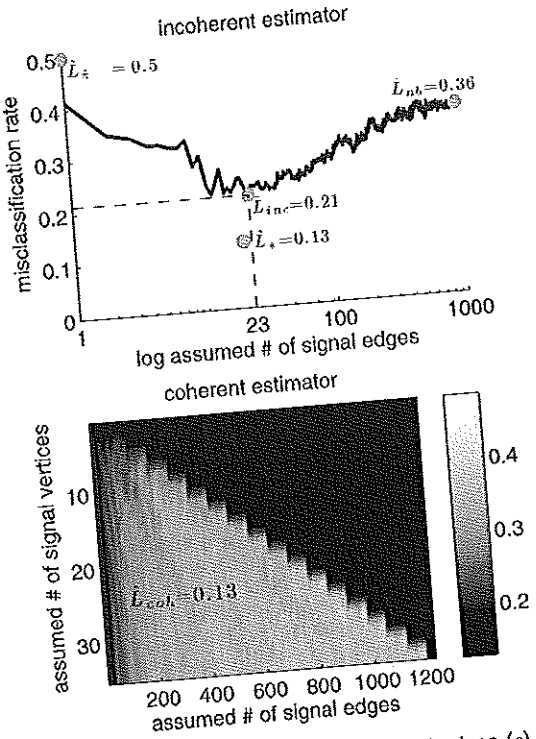


Fig. 5. When constraints on the number of signal-edges ($s$) or signal-vertices ($m$) are unknown, a search over these hyperparameters can yield estimates $\widehat{s}$ and $\widehat{m}$. Both panels depict held-out cross-validation error as a function of varying these parameters for the model $\mathcal{M}_{70}(1, 20; 0.5, 0.1, 0.3)$ (the same as in Figs. 2 and 3), using 200 training samples and 500 test samples, with $m = 1$ and $s = 20$. The top panel depicts the misclassification rate of the incoherent estimator as a function of the number of estimated signal-edges on a log scale, with the best performing classifier achieving $\widehat{L}_{inc} = 0.21$. Note that in this simulation, $s = 20 < \widehat{s}_{inc} = 23$. This "conservatism" is typical and appropriate in many model selection situations. The bottom panel shows $\widehat{L}_{coh}$ as a function of both $m'$ and $s'$. For this simulation, $\widehat{m}_{coh} = 1$ and $\widehat{s}_{coh} = 24$, further corroborating the conservative stance on model selection. Note that $L_{\boldsymbol{x}} > \widehat{L}_{nb} > \widehat{L}_{inc} > \widehat{L}_{coh} \geq L_*$, as one would hope for this coherent simulation. Incidentally, the coherent classifier achieved Bayes error here, $L_* = 0.13$.

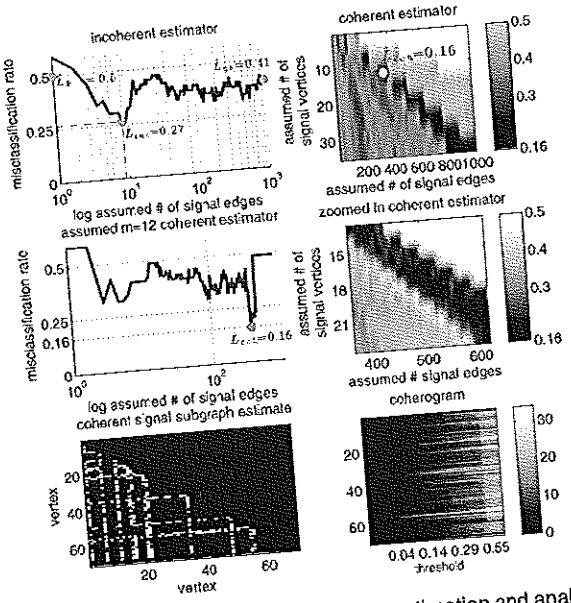


Fig. 6. MR connectome sex signal-subgraph estimation and analysis. By cross-validating over hyperparameters and models, we estimate that the "best" incoherent signal-subgraph (for this inference task on these data) has $\widehat{s}_{inc} = 10$ and yields a misclassification rate of $\widehat{L}_{inc} = 0.27$, whereas the best coherent signal-subgraph has $\widehat{m}_{coh} = 12$ and $\widehat{s}_{coh} = 360$, achieving $\widehat{L}_{coh} = 0.16$. The top two panels depict the same information as Fig. 5. The middle two depict misclassification rate (left) for different choices of $m' = 12$ as a function of $s'$ and (right) a zoomed-in depiction of the top right panel. The bottom left panel shows the estimated signal-subgraph, and the bottom right shows the coherogram. Together, these bottom panels suggest that the signal-subgraph for these data is at least somewhat coherent.

pair achieved $\widehat{L}_{coh} = 0.13$ (which is equal to the Bayes error) with $\widehat{m}_{coh} = 1$ and $\widehat{s}_{coh} = 24$, suggesting that $n$ was sufficiently large to correctly find the true signal-vertex, and further corroborating the "better safe than sorry" attitude to selecting the signal-edges.

## 5 MR Connectome Sex Classification

A connectome is a brain-graph [23]. MR connectomes utilize multimodal magnetic resonance (MR) imaging to determine both the vertex and edge set for each individual [2]. This section investigates the utility of the classifiers developed above on data collected for the Baltimore Longitudinal Study of Aging, as described previously [24]. Briefly, 49 subjects (25 males, 24 females) underwent a diffusion-weighted MRI protocol. The Magnetic Resonance Connectome Automated Pipeline (MRCAP) was used to convert each subject's raw multimodal MR data into a connectome [25] (each connectome is a simple graph with 70 vertices and up to $\binom{70}{2} = 2{,}415$ edges). Lacking strong priors on either the number of signal-edges or signal-vertices in the signal-subgraph (or even whether a signal-subgraph exists), we searched over a large space of hyperparameters using leave-one-out cross-validated misclassification performance as our metric of success (Fig. 6). The naive Bayes classifier—which assumes the signal-subgraph is the whole edge set, $\widehat{\mathcal{S}}_{nb} = \mathcal{E}$—performs marginally better than chance: $\widehat{L}_{nb} = 0.41$ ($p$-value $\approx 0.05$ assessed by a permutation test). With a relatively small number of incoherent edges—$\widehat{s}_{inc} = 10$—the incoherent classifier (top left panel) achieves $\widehat{L}_{inc} = 0.27$, significantly better than chance ($p$-value $< 0.0007$), but not significantly better than the naive Bayes classifier (using McNemar's test). The coherent classifier achieved a minimum of $\widehat{L}_{coh} = 0.16$ (top right and middle panels), significantly better than both chance and the naive Bayes classifier ($p$-values $< 10^{-5}$ and $< 0.004$, respectively). This improved performance upon using the coherent classifier suggests that the signal-subgraph is at least approximately coherent. Using $\widehat{m}_{coh} = 12$ and $\widehat{s}_{coh} = 360$ from the best performing coherent classifier, we can estimate the signal-subgraph (bottom left). The coherogram suggests that, indeed, the signal is somewhat, but not entirely coherent (bottom right).

We next compare the performance of our classifiers on this MR connectome sex classification dataset to several other classifiers. First, a standard parametric classifier: lasso. We chose the regularization parameter via a 10-fold cross-validation. Second, a nonparametric (distribution free) classifier: $k_n$-nearest neighbor ($k$NN), which operates directly on graphs [26]. This $k$NN classifier uses the Frobenius norm distance metric. We tried all $k \in [n]$ and simply report the best performance. The universal consistency of this $k$NN classifier is useful in assessing the

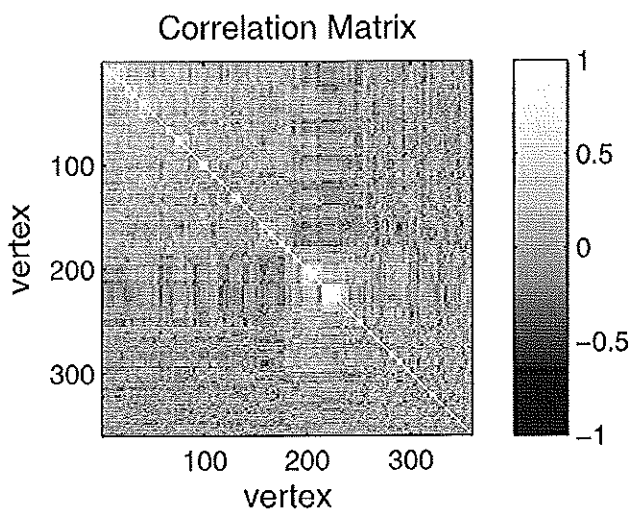


Fig. 8. The correlation matrix between all the edges in the coherent signal-subgraph estimate. Edges are organized by co-clustering to highlight any similarities. Although most edges are uncorrelated, several groups of edges cluster, indicative of the fact that the edges are not independent ($p$-value of $\approx 0$ using a two-sample Kolmogorov-Smirnov test comparing the real and synthetic correlation matrices).

statistical investigation. Moreover, it presents two approaches for estimating the signal-subgraph: The first using only vertex label information, the second also utilizing graph structure. The resulting estimators satisfy the five above mentioned desiderata (Section 2.4.1): (model) consistency, robust to model misspecifications, quadratic complexity, interpretability in terms of the vertices and edges, and state-of-the-art finite sample/empirical performance. Third, simulated data analysis indicates that neither approach dominates the other; rather, the best approach is a function of both the model and the amount of training data. And while the lasso classifier has similar error properties as our incoherent classifier, lasso's computational time is about an order of magnitude longer.

Fourth, these classifiers are applied to an MR connectome sex classification dataset; the coherent classifier performs significantly better than a variety of benchmark classifiers. More specifically, the coherent classifier outperformed a pair of classifiers that use only vertex labels (the naive Bayes and Lasso classifiers) as well as a classifier that only uses structural information (the invariant-$k$NN classifier). Only the signal subgraph classifier and graph-$k$NN classifier use *both* vertex labels and graph structure. Since the graph-$k$NN classifier, however, is universally consistent, it has high variance and therefore takes much longer than the coherent classifier to converge to a good estimate.

Fifth, synthetic data analysis suggests that while we can use the signal-subgraph estimators to improve classification performance, we should not expect that all the edges in the estimated signal-subgraph will be the true signal-edges, even when the model is correct. Moreover, we might expect a drastic improvement in classification performance with only a few additional data samples. Finally, model checking suggests that the independent edge assumption does not fit the data well.

### 6.2 Related Work

Our signal-subgraph classifiers represent somewhat of a departure from previous work. Most graph classification algorithms come from the "structural pattern recognition" school of thought, lacking an explicit statistical model and associated provable properties (see [1, Section 2.1 and references therein] for an excellent review). On the other hand, most work on "statistical pattern recognition" begins by assuming the data to be classified are euclidean vectors [31]. Our work is a unification of the two. Moreover, because the sufficient statistics are essentially encoded in a matrix, our work can be related to recent developments in matrix decompositions. For example, sparse and low-rank matrix decompositions are close in spirit to our coherent signal subgraph estimators [32], [33], [34]. Note, however, that our coherent estimator is robust to signal-vertices having a subset of its edges highly nonsignificant, that is, the coherent signal-subgraph estimator can be thought of as a sparse and *locally* low-rank decomposition.

### 6.3 Future Work

Collectively, the above analyses suggest a number of possible next steps. First, collect more data. Second, relax various assumptions, including the 1) independent edge assumption by considering conditionally independent edges [35], [36], [37], 2) binary edge and two-class assumptions, and 3) uniquely labeled and identical vertices in each graph assumptions. Specifically, extension to situations for which none of the vertices are labeled [38], [39], only some subset of vertices are labeled [40], [41], [42], or data are otherwise errorfully observed [43] are all avenues of future investigation. Third, transform a number of conjectures that have arisen due to these results into theorems. For instance, perhaps the misclassification rate is a monotonic function of the missed-edge rate. Fourth, (Bayesian) model-averaging to combine estimated signal-subgraphs instead of picking one might improve performance (perhaps at the cost of computational resources and interpretability).

We hope the proposed approaches will yield many applications. To that end, all the data and code used in this work is available from the author's website, http://jovo.me.

## Acknowledgments

This work was partially supported by the Research Program in Applied Neuroscience. The authors would like to thank Michael Trosset for a helpful suggestion.

## References

[1] H. Bunke and K. Riesen, "Towards the Unification of Structural and Statistical Pattern Recognition," *Pattern Recognition Letters*, vol. 33, no. 7, http://linkinghub.elsevier.com/retrieve/pii/S0167865511001309 http://www.sciencedirect.com/science/article/pii/S0167865511001309, pp. 811-825, May 2011.

[2] P. Hagmann, L. Cammoun, X. Gigandet, S. Gerhard, P. Ellen Grant, V. Wedeen, R. Meuli, J.-P. Thiran, C.J. Honey, and O. Sporns, "MR Connectomics: Principles and Challenges," *J. Neuroscience Methods*, vol. 194, no. 1, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=20096730, pp. 34-45, 2010.

[3] D.L.D. Donoho, M. Elad, and V.N.V. Temlyakov, "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise," *IEEE Trans. Information Theory*, vol. 52, no. 1, pp. 6-18, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1564423, Jan. 2006.

**Joshua T. Vogelstein** received the BS degree from the Department of Biomedical Engineering at Washington University in St. Louis, Missouri, in 2002, the MS degree from the Department of Applied Mathematics and Statistics (AMS) at Johns Hopkins University (JHU), Baltimore, Maryland, in 2009, and the PhD degree from the Department of Neuroscience at Johns Hopkins School of Medicine, Baltimore, Maryland, in 2009. He was a postdoctoral fellow in AMS@JHU from 2009 until 2011, at which time he was appointed an assistant research scientist in the Department of Applied Mathematics and Statistics at JHU, with a joint appointment in the Human Language Technology Center of Excellence, and as a member of the Institute for Data Intensive Science and Engineering. He is currently a visiting assistant research professor at Duke University jointly in the Departments of Mathematics and Statistics. His research interests primarily include computational statistics, focusing on ultrahigh-dimensional and non-euclidean neuroscience data. His research has been featured in a number of prominent scientific and engineering journals including the *Annals of Applied Statistics*, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *Nature Neuroscience*, *SIAM Journal of Matrix Analysis and Applications*, and *Science Translational Medicine*.

**William R. Gray** received the bachelor's degree in electrical engineering from Vanderbilt University in 2003, and the MS degree in electrical engineering from the University of Southern California, Los Angeles, in 2005. Currently, he is working toward the PhD degree in electrical engineering at Johns Hopkins University, Baltimore, Maryland, where he is conducting research in the areas of connectivity, signal and image processing, and machine learning. He is also a member of the technical staff at the Johns Hopkins University Applied Physics Laboratory, where he manages projects in the Biomedicine and Undersea Warfare business areas. He is a member of the IEEE, Eta Kappa Nu, and Tau Beta Pi.

**R. Jacob Vogelstein** received the ScB degree in neuroengineering from Brown University, Providence, Rhode Island, and the PhD degree in biomedical engineering from the Johns Hopkins University School of Medicine, Baltimore, Maryland. He currently serves as the program manager for applied neuroscience at the Johns Hopkins University Applied Physics Laboratory and has an appointment as an assistant research professor at the JHU Whiting School of Engineering's Department of Electrical and Computer Engineering. He has worked on neuroscience technologies for over a decade, focusing primarily on neuromorphic systems and closed-loop brainmachine interfaces. His research has been featured in a number of prominent scientific and engineering journals, including the *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, the *IEEE Transactions on Biomedical Circuits and Systems*, and the *IEEE Transactions on Neural Networks*. He is a member of the IEEE.

**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, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland. At Johns Hopkins, he holds joint appointments in the Department of Computer Science, the Department of Electrical and Computer Engineering, the Center for Imaging Science, the Human Language Technology Center of Excellence, and the Whitaker Biomedical Engineering Institute. He is a past president of the Interface Foundation of North America—Computing Science and Statistics, a past chair of the American Statistical Association Section on Statistical Computing, a past vice president of the International Association for Statistical Computing, and is on the editorial boards of the *Journal of Computational and Graphical Statistics*, *Computational Statistics and Data Analysis*, and *Computational Statistics*. 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.