An Information-Theoretic Derivation of Min-Cut-Based Clustering

Anil Raj and Chris H. Wiggins

Abstract—Min-cut clustering, based on minimizing one of two heuristic cost functions proposed by Shi and Malik nearly a decade ago, has spawned tremendous research, both analytic and algorithmic, in the graph partitioning and image segmentation communities over the last decade. It is, however, unclear if these heuristics can be derived from a more general principle, facilitating generalization to new problem settings. Motivated by an existing graph partitioning framework, we derive relationships between optimizing relevance information, as defined in the Information Bottleneck method, and the regularized cut in a K-partitioned graph. For fast-mixing graphs, we show that the cost functions introduced by Shi and Malik can be well approximated as the rate of loss of predictive information about the location of random walkers on the graph. For graphs drawn from a generative model designed to describe community structure, the optimal information-theoretic partition and the optimal min-cut partition are shown to be the same with high probability.

Index Terms—Graphs, clustering, information theory, min-cut, Information Bottleneck, graph diffusion.

1 Introduction

MIN-CUT-BASED graph partitioning has been used successfully to find clusters in networks, with applications in image segmentation as well as clustering biological and sociological networks. The central idea is to develop fast and efficient algorithms that optimally cut the edges between graph nodes, resulting in a separation of graph nodes into clusters. Particularly since Shi and Malik successfully showed [1] that the *average* cut and the *normalized* cut (defined below) were useful heuristics to be optimized, there has been tremendous research in constructing the best normalized-cut-based cost function in the image segmentation community.

Additionally, several insightful works have focused on providing an interpretation and a justification for min-cut-based clustering, within the framework of graph diffusion. Meila and Shi [2] showed rigorous connections between normalized min-cut-based clustering and the lumpability of the Markov chains underlying the corresponding discrete diffusion operator. More recently, Lafon and Lee [3] and Nadler et al. [4] showed the close relationship between the problem of spectral clustering and that of learning locality-preserving embeddings of data, using diffusion maps.

The Information Bottleneck (IB) method [5], [6] is a clustering technique, based on rate distortion theory [7], that has been successfully applied in a wide variety of contexts, including clustering word documents and gene expression profiles [8]. The IB method is also capable of learning clusters in graphs and has been successfully used for synthetic and natural networks [9]. In the hard clustering case, given the diffusive probability distribution over a graph, IB optimally assigns probability distributions associated with nodes, into distinct groups. These assignment rules define a separation of the graph nodes into clusters.

Here, we illustrate how minimizing the two cut-based heuristics introduced by Shi and Malik can be well approximated by the rate of loss of *relevance information*, defined in the IB method applied to clustering graphs. To establish these relations, we must first define the graphs to be partitioned; we assume hard clustering and the cluster cardinality to be K. We show, numerically, that maximizing mutual information and minimizing *regularized* cut amount to the same partition with high probability, for more modular 32-node graphs, where *modularity* is defined by the probability of intercluster edge connections in the Stochastic Block Model for graphs (see Section 5). We also show that the optimization goal of maximizing relevance information is equivalent to minimizing the regularized cut for 16-node graphs.[1]

2 The Min-Cut Problem

Following [10], for an undirected, unweighted graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with n nodes and m edges, represented[2] by its adjacency matrix $\mathbf{A} := \{A_{xy} = 1 \iff x \sim y\}$, we define for two not necessarily disjoint sets of nodes $\mathbf{V}_+, \mathbf{V}_- \subseteq \mathbf{V}$, the association:

$$W(\mathbf{V}_+, \mathbf{V}_-) = \sum_{x \in \mathbf{V}_+, y \in \mathbf{V}_-} A_{xy}. \tag{2.1}$$

1. We chose 16-node graphs so the network and its partitions could be parsed visually with ease.

2. We use the shorthand $x \sim y$ to mean x is adjacent to y.

- *A. Raj is with the Department of Applied Physics and Applied Mathematics, Columbia University, 200 S. W. Mudd Building, MC 4701, 500 W. 120th Street, New York, NY 10027. E-mail: ar2384@columbia.edu.*
- *C.H. Wiggins is with the Department of Applied Physics and Applied Mathematics, Columbia University, 200 S. W. Mudd Building, MC 4701, 500 W. 120th Street, New York, NY 10027, and the Center for Computational Biology and Bioinformatics, Columbia University, 200 S. W. Mudd Building, MC 4701, 500 W. 120th Street, New York, NY 10027. E-mail: chris.wiggins@columbia.edu.*

Manuscript received 26 Nov. 2008; revised 9 Apr. 2009; accepted 18 May 2009; published online 8 June 2009.
Recommended for acceptance by K. Murphy.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-11-0815.
Digital Object Identifier no. 10.1109/TPAMI.2009.124.

We define a bisection of $\mathbf{V}$ into $\mathbf{V}_\pm$ if $\mathbf{V}_+ \cup \mathbf{V}_- = \mathbf{V}$ and $\mathbf{V}_+ \cap \mathbf{V}_- = \emptyset$. For a bisection of $\mathbf{V}$ into $\mathbf{V}_+$ and $\mathbf{V}_-$, the "cut" is defined as $c(\mathbf{V}_+, \mathbf{V}_-) = W(\mathbf{V}_+, \mathbf{V}_-)$. We also quantify the size of a set $\mathbf{V}_+ \subseteq \mathbf{V}$ in terms of the number of nodes in the set $\mathbf{V}_+$ or the number of edges with at least one node in the set $\mathbf{V}_+$:

$$\begin{aligned} \omega(\mathbf{V}_+) &= \sum_{x \in \mathbf{V}_+} 1, \\ \Omega(\mathbf{V}_+) &= \sum_{x \in \mathbf{V}_+} d_x, \end{aligned} \tag{2.2}$$

where d_x is the degree of node x.

Shi and Malik [1] defined a pair of regularized cuts, for a bisection of $\mathbf{V}$ into $\mathbf{V}_+$ and $\mathbf{V}_-$; the *average cut* was defined as

$$\mathcal{A} = \frac{W(\mathbf{V}_+, \mathbf{V}_-)}{\omega(\mathbf{V}_+)} + \frac{W(\mathbf{V}_+, \mathbf{V}_-)}{\omega(\mathbf{V}_-)} \tag{2.3}$$

and the *normalized cut* as

$$\mathcal{N} = \frac{W(\mathbf{V}_+, \mathbf{V}_-)}{\Omega(\mathbf{V}_+)} + \frac{W(\mathbf{V}_+, \mathbf{V}_-)}{\Omega(\mathbf{V}_-)}. \tag{2.4}$$

This definition can be generalized, for a K-partition of $\mathbf{V}$ into $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_K$ [10], to

$$\mathcal{A} = \sum_j \frac{W(\mathbf{V}_j, \bar{\mathbf{V}}_j)}{\omega(\mathbf{V}_j)}, \tag{2.5}$$

$$\mathcal{N} = \sum_j \frac{W(\mathbf{V}_j, \bar{\mathbf{V}}_j)}{\Omega(\mathbf{V}_j)}, \tag{2.6}$$

where $\bar{\mathbf{V}}_j = \mathbf{V} \setminus \mathbf{V}_j$.

For the graph $\mathcal{G}$, we can define the graph Laplacian $\mathbf{\Delta} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is a diagonal matrix of vertex degrees. For a bisection of $\mathbf{V}$, we also define the partition indicator vector $\mathbf{h}$ [11]

$$h_x = \begin{cases} +1, & \forall x \in \mathbf{V}_+, \\ -1, & \forall x \in \mathbf{V}_-. \end{cases} \tag{2.7}$$

Specifying two "prior" probability distributions over the set of nodes $\mathbf{V}$: 1) $p(x) \propto 1$ and 2) $p(x) \propto d_x$, we then define the *average* of $\mathbf{h}$ to be

$$\begin{aligned} \bar{\mathbf{h}} &= \frac{\sum_{x \in \mathbf{V}} h_x}{n}, \\ \langle \mathbf{h} \rangle &= \frac{\sum_{x \in \mathbf{V}} d_x h_x}{2m}. \end{aligned} \tag{2.8}$$

The cut, as defined by Fiedler [11], and the regularized cuts, as defined by Shi and Malik [1], can then be written in terms of $\mathbf{h}$ as (see Appendix A)

$$\begin{aligned} c &= \frac{1}{4} \mathbf{h}^\mathbf{T} \mathbf{\Delta} \mathbf{h}, \\ \mathcal{A} &= \frac{1}{n} \frac{\mathbf{h}^\mathbf{T} \mathbf{\Delta} \mathbf{h}}{1 - \bar{\mathbf{h}}^2}, \\ \mathcal{N} &= \frac{1}{2m} \frac{\mathbf{h}^\mathbf{T} \mathbf{\Delta} \mathbf{h}}{1 - \langle \mathbf{h} \rangle^2}. \end{aligned} \tag{2.9}$$

More generally, for a K-partition, we define the partition indicator matrix $\mathbf{Q}$ as

$$Q_{zx} \equiv p(z \mid x) = 1, \quad \forall x \in z, \tag{2.10}$$

where $z \in \{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_K\}$, and define $\mathbf{P}$ as a diagonal matrix of the "prior" probability distribution over the nodes. The regularized cut can then be generalized as

$$\mathcal{C} = \sum_j \frac{[\mathbf{Q} \mathbf{\Delta} \mathbf{Q}^\mathbf{T}]_{jj}}{[\mathbf{Q} \mathbf{P} \mathbf{Q}^\mathbf{T}]_{jj}}, \tag{2.11}$$

where, for $p(x) \propto 1$, $\mathcal{C} = \mathcal{A}$ and, for $p(x) \propto d_x$, $\mathcal{C} = \mathcal{N}$.

Inferring the optimal $\mathbf{h}$ (or $\mathbf{Q}$), however, has been shown to be an NP-hard combinatorial optimization problem [12].

3 Information Bottleneck

Rate distortion theory, which provides the foundations for lossy data compression, formulates clustering in terms of a compression problem; it determines the code with minimum average length such that information can be transmitted without exceeding some specified distortion. Here, the model complexity, or *rate*, is measured by the mutual information between the data and their representative codewords (average number of bits used to store a data point). Simpler models correspond to smaller rates but typically suffer from relatively high *distortion*. The distortion measure, which can be identified with loss functions, usually depends on the problem; in the simplest of cases, it is the variance of the difference between an example and its cluster representative.

The Information Bottleneck method [6] proposes the use of mutual information as a natural distortion measure. In this method, the data are compressed into clusters while maximizing the amount of information that the "cluster representation" preserves about some specified *relevance* variable.[3] For example, in clustering word documents, one could use the "topic" of a document as the relevance variable.

For a graph $\mathcal{G}$, let X be a random variable over graph nodes, Y be the relevance variable, and Z be the random variable over clusters. Graph partitioning using the IB method [9] learns a probabilistic cluster assignment function $p(z \mid x)$ which gives the probability that a given node x belongs to cluster z. The optimal $p(z \mid x)$ minimizes the mutual information between X and Z, while minimizing the loss of predictive information between Z and Y. This complexity-fidelity trade-off can be expressed in terms of a functional to be minimized

$$\mathcal{F}[p(z \mid x)] = -I[Y; Z] + TI[X; Z], \tag{3.1}$$

where the temperature T parameterizes the relative importance of precision over complexity. As $T \to 0$, we reach the "hard clustering" limit where each node is assigned with unit probability to one cluster (i.e., $p(z \mid x) \in \{0, 1\}$).

Graph clustering, as formulated in terms of the IB method, requires a joint distribution $p(y, x)$ to be defined on the graph; we use the distribution given by continuous-time graph diffusion[4] as it naturally captures topological information about the network [9]. The relevance variable Y then ranges over the nodes of the graph and is defined as the node at which a random walker ends at time t if the random walker starts at node x at time 0. For continuous-time diffusion, the conditional distribution $p(y \mid x)$ is given as

3. See [8] for a detailed discussion on the relevance variable.
4. See [13] for a detailed discussion on graph diffusion and mixing.

$$G^t_{yx} = p(y \mid x) = \left[e^{-t\mathbf{\Delta P}^{-1}}\right]_{yx}, \tag{3.2}$$

where $\mathbf{\Delta}$ is the positive semidefinite graph Laplacian and $\mathbf{P}$ is a diagonal matrix of the prior distribution over the graph nodes, as described earlier. Note that the diagonal matrix $\mathbf{P}$ can be any prior distribution over the graph nodes. The characteristic diffusion time scale τ of the system is given by the inverse of the smallest nonzero eigenvalue of the diffusion operator exponent $\mathbf{\Delta P}^{-1}$ and characterizes the slowest decaying mode in the system. A more common convention for matrix operations in graph theory is to use left eigenvectors. However, since our analysis is within a probabilistic framework, we use right eigenvectors throughout this paper to conform to conventions used in probability theory (particularly with regard to conditional and marginal distributions).

To calculate the joint distribution $p(y, x)$ from the conditional $\mathbf{G}^t$, we must specify an initial or prior distribution;[5] we use the two different priors $p(x)$, used in (2.8), to calculate $\bar{\mathbf{h}}$ and $\langle \mathbf{h} \rangle$: 1) $p(x) \propto 1$ and 2) $p(x) \propto d_x$. Throughout this paper, time dependence needs to be considered only when the conditional distribution $p(y \mid x)$ is replaced by the diffusion Green's function $\mathbf{G}$; thus, time dependence will be explicitly denoted only once $\mathbf{G}$ is invoked.

4 Rate of Information Loss in Graph Diffusion

Here, we analyze the rate of loss of predictive information between the relevance variable Y and the cluster variable Z, during diffusion on a graph $\mathcal{G}$, after the graph nodes have been hard-partitioned into K clusters.

4.1 Well-Mixed Limit of Graph Diffusion

For a given partition $\mathbf{Q}$ of the graph, defined in (2.10), we approximate the mutual information $I[Y; Z]$ when diffusion on the graph reaches its well-mixed limit. We introduce the linear *dependence* $\eta(y, z)$ such that

$$p(y, z) = p(y)p(z)(1 + \eta). \tag{4.1}$$

This implies $\langle \eta \rangle_y = \langle \eta \rangle_z = 0$ and $\langle \langle \eta^2 \rangle_z \rangle_y = \langle \eta \rangle$, where $\langle\rangle$ denotes expectation over the joint distribution and $\langle \rangle_y$ and $\langle \rangle_z$ denote expectation over the corresponding marginals.

In the well-mixed limit, we have $|\eta| \ll 1$. The predictive information (expressed in nats) can then be approximated as:

$$\begin{aligned} I[Y; Z] &= \left\langle \ln \frac{p(z, y)}{p(z)p(y)} \right\rangle \\ &= \langle \langle (1 + \eta) \ln (1 + \eta) \rangle_y \rangle_z \\ &\approx \left\langle \left\langle (1 + \eta)\left(\eta - \frac{1}{2}\eta^2\right) \right\rangle_y \right\rangle_z \\ &\approx \left\langle \left\langle \eta + \frac{1}{2}\eta^2 \right\rangle_y \right\rangle_z \\ &= \frac{1}{2} \langle \langle \eta^2 \rangle_y \rangle_z \\ &= \frac{1}{2} \sum_{y,z} p(y)p(z) \left(\frac{p(z, y)}{p(z)p(y)} - 1 \right)^2 \\ &= \frac{1}{2} \left(\sum_{y,z} \frac{p(y, z)^2}{p(y)p(z)} - 1 \right) \equiv \iota. \end{aligned} \tag{4.2, 4.3}$$

Here, we define ι as a first-order approximation to $I[Y; Z]$ in the well-mixed limit of graph diffusion. This quadratic approximation for $I[Y; Z]$ is known as the χ^2-approximation.

Note that the joint and marginal distributions can also be related by the exponential dependence $\theta(y, z)$ defined by

$$p(y, z) = p(y)p(z)e^{\theta}.$$

Under this definition, the domain of the dependence is unbounded (i.e., $\theta \in \mathbb{R}$) and the mutual information is easily expressed as $I[Y; Z] = \langle \theta \rangle$. We also have

$$\sum_{i=1}^{\infty} \frac{\langle \theta^i \rangle_y}{i!} = \sum_{i=1}^{\infty} \frac{\langle \theta^i \rangle_z}{i!} = 0.$$

However, in the well-mixed limit $|\theta| \ll 1$, to first nontrivial order, $\theta \approx \eta$ and the expression for $I[Y; Z]$ in terms of θ has the same form as (4.2).

We also have

$$\begin{aligned} \eta(y, z) &= \frac{p(z \mid y)}{p(z)} - 1 \\ &\leq \frac{1}{p(z)} - 1 \\ &\leq \max_z \left(\frac{1}{p(z)} \right) - 1 \\ \eta(y, z) &\leq \max_y \left(\frac{1}{p(y)} \right) - 1 \\ \Rightarrow \eta(y, z) &\leq \min\left(\max_z \left(\frac{1}{p(z)} \right), \quad \max_y \left(\frac{1}{p(y)} \right) \right) - 1. \end{aligned} \tag{4.4}$$

Thus, $\eta(y, z)$ is bounded from below by -1 (by definition) and from above as shown in (4.4). However, $\theta(y, z)$ is unbounded and negatively divergent for short times. Since η is much better behaved than θ for short times, for the sake of simplicity we choose to use the linear dependence instead of the exponential dependence.

4.1.1 Well-Mixed K-Partitioned Graph

As in the IB method, the Markov condition $Z - X - Y$ allows us to make several simplifications for the conditional distributions and associated information-theoretic measures. For a K-partition $\mathbf{Q}$ of the graph, we have

$$\begin{aligned} p(y, z) &= \sum_x p(x, y, z) \\ &= \sum_x p(z \mid y, x)p(y \mid x)p(x) \\ &= \sum_x p(z \mid x)p(y \mid x)p(x) \equiv \mathbf{QPG}^{t\mathbf{T}}. \end{aligned} \tag{4.5}$$

5. Strictly speaking, any diagonal matrix $\mathbf{P}$ that we specify determines the steady-state distribution. Since we are modeling the distribution of random walkers at statistical equilibrium, we always use this distribution as our initial or prior distribution.

$$\begin{aligned} p(y,z)^2 &= \left(\sum_x p(z \mid x)p(y \mid x)p(x)\right)^2 \\ &= \sum_{x,x'=1}^{n} p(z \mid x)p(y \mid x)p(x)p(z \mid x')p(y \mid x')p(x') \\ &= \sum_{x,x'=1}^{n} Q_{zx}G^t_{yx}P_xQ_{zx'}G^t_{yx'}P_{x'}. \end{aligned} \tag{4.6}$$

$$\begin{aligned} p(z) &= \sum_x p(z \mid x)p(x) \\ &= \sum_x Q_{zx}P_x. \end{aligned} \tag{4.7}$$

Graph diffusion being a Markov process, we have $\sum_{y=1}^{n} G^t_{x'y}G^t_{yx} = G^{2t}_{x'x}$. Using this and Bayes rule $G^t_{yx}P_x = G^t_{xy}P_y$, we have

$$\begin{aligned} \iota &= \frac{1}{2}\left(\sum_{y,z}\frac{\sum_{x,x'=1}^{n} Q_{zx}G^t_{yx}P_xQ_{zx'}G^t_{yx'}P_{x'}}{(\sum_{x''}Q_{zx''}P_{x''})P_y} - 1\right) \\ &= \frac{1}{2}\left(\sum_{y,z}\frac{\sum_{x,x'=1}^{n} Q_{zx}Q_{zx'}P_yG^t_{x'y}G^t_{yx}P_x}{(\sum_{x''}Q_{zx''}P_{x''})P_y} - 1\right) \\ &= \frac{1}{2}\left(\sum_{z=1}^{K}\frac{\sum_{x,x'=1}^{n} Q_{zx}Q_{zx'}\left(\sum_{y=1}^{n}G^t_{x'y}G^t_{yx}\right)P_x}{(\sum_{x''}Q_{zx''}P_{x''})} - 1\right) \\ &= \frac{1}{2}\left(\sum_{z=1}^{K}\frac{\sum_{x,x'=1}^{n} Q_{zx}Q_{zx'}G^{2t}_{x'x}P_x}{(\sum_{x''}Q_{zx''}P_{x''})} - 1\right). \end{aligned} \tag{4.8}$$

In the hard clustering case, $\sum_x Q_{zx}P_x = p(z) = [\mathbf{QPQ^T}]_{zz}$, and we have

$$\iota = \frac{1}{2}\left(\sum_{z=1}^{K}\frac{[\mathbf{Q}(\mathbf{G}^{2t}\mathbf{P})\mathbf{Q}^{\mathbf{T}}]_{zz}}{[\mathbf{QPQ}^{\mathbf{T}}]_{zz}} - 1\right). \tag{4.9}$$

4.1.2 Well-Mixed 2-Partitioned Graph

We can rewrite ι as

$$\begin{aligned} \iota &= \frac{1}{2}\langle\langle\eta^2\rangle_y\rangle_z \\ &= \frac{1}{2}\left\langle\left\langle\frac{(p(z \mid y) - p(z))^2}{p(z)^2}\right\rangle_z\right\rangle_y. \end{aligned} \tag{4.10}$$

For a bisection $\mathbf{h}$ of the graph, $z \in \{+1, -1\}$, and we have

$$p(z \mid x) = \frac{1}{2}(1 \pm h_x) \equiv \frac{1}{2}(1 + zh_x). \tag{4.11}$$

$$\begin{aligned} p(z \mid y) &= \frac{1}{p(y)}\sum_x p(z,y,x) \\ &= \frac{1}{p(y)}\sum_x p(z \mid x)p(y \mid x)p(x) \\ &= \frac{1}{2}\sum_x (1 + zh_x)p(x \mid y) \\ &= \frac{1}{2}(1 + z\langle\mathbf{h} \mid y\rangle). \end{aligned} \tag{4.12}$$

$$\begin{aligned} p(z) &= \sum_x p(z,x) = \sum_x p(z \mid x)p(x) \\ &= \frac{1}{2}\sum_x (1 + zh_x)p(x) \\ &= \frac{1}{2}(1 + z\langle\mathbf{h}\rangle). \end{aligned} \tag{4.13}$$

$$\begin{aligned} p(z \mid y) - p(z) &= \frac{1}{2}(1 + z\langle\mathbf{h} \mid y\rangle) - \frac{1}{2}(1 + z\langle\mathbf{h}\rangle) \\ &= \frac{1}{2}z(\langle\mathbf{h} \mid y\rangle - \langle\mathbf{h}\rangle). \end{aligned} \tag{4.14}$$

We then have

$$\begin{aligned} \left\langle\frac{(p(z \mid y) - p(z))^2}{p(z)^2}\right\rangle_z &= \sum_{z=-1,1}\frac{\frac{1}{4}(\langle\mathbf{h} \mid y\rangle - \langle\mathbf{h}\rangle)^2}{\frac{1}{2}(1 + z\langle\mathbf{h}\rangle)} \\ &= \frac{(\langle\mathbf{h} \mid y\rangle - \langle\mathbf{h}\rangle)^2}{2}\sum_{z=-1,1}\frac{1}{1 + z\langle\mathbf{h}\rangle} \\ &= \frac{(\langle\mathbf{h} \mid y\rangle - \langle\mathbf{h}\rangle)^2}{1 - \langle\mathbf{h}\rangle^2}. \end{aligned} \tag{4.15}$$

The mutual information $I[Y;Z]$ can then be approximated as

$$\begin{aligned} \iota &= \frac{1}{2}\frac{\langle(\langle\mathbf{h} \mid y\rangle - \langle\mathbf{h}\rangle)^2\rangle_y}{1 - \langle\mathbf{h}\rangle^2} \\ &= \frac{1}{2}\frac{\sigma^2_y(\langle\mathbf{h} \mid y\rangle)}{1 - \langle\mathbf{h}\rangle^2}. \end{aligned} \tag{4.16}$$

Using Bayes rule $p(x \mid y)p(y) = p(y \mid x)p(x)$, we have

$$\langle\mathbf{h} \mid y\rangle = \sum_x h_x p(x \mid y) = \sum_x \frac{h_x p(y \mid x)p(x)}{p(y)}. \tag{4.17}$$

$$\begin{aligned} \langle\langle\mathbf{h} \mid y\rangle^2\rangle_y &= \sum_{y=1}^{n} p(y)\sum_{x,x'=1}^{n} h_xh_{x'}\frac{p(y \mid x)p(x)p(x' \mid y)}{p(y)} \\ &= \sum_{y=1}^{n}\sum_{x,x'=1}^{n} h_xh_{x'}p(x' \mid y)p(y \mid x)p(x). \end{aligned} \tag{4.18}$$

Again, graph diffusion being a Markov process,

$$\begin{aligned} \langle\langle\mathbf{h} \mid y\rangle^2\rangle_y &= \sum_{x,x'=1}^{n} h_xh_{x'}p^{2t}(x' \mid x)p(x) \\ &= \langle h_xh_{x'}\rangle_{2t}. \end{aligned} \tag{4.19}$$

Time dependence is explicitly denoted here to highlight the fact that diffusion on the graph is till time $2t$. Substituting $\langle\mathbf{h} \mid y\rangle$ in (4.16), we get

$$\begin{aligned} \sigma^2(\langle\mathbf{h} \mid y\rangle) &= \langle\langle\mathbf{h} \mid y\rangle^2\rangle_y - \langle\mathbf{h}\rangle^2 \\ &= \langle h_xh_{x'}\rangle_{2t} - \langle\mathbf{h}\rangle^2, \end{aligned} \tag{4.20}$$

$$\iota = \frac{1}{2}\frac{\langle h_xh_{x'}\rangle_{2t} - \langle\mathbf{h}\rangle^2}{1 - \langle\mathbf{h}\rangle^2}. \tag{4.21}$$

4.2 Fast-Mixing Graphs

When diffusion on a graph reaches its well-mixed limit in short times, we have $\mathbf{G}^{2t} \approx \mathbf{I} - 2t\mathbf{\Delta}\mathbf{P}^{-1}$. Thus, for a K-partition of a graph,

$$\begin{aligned}\mathbf{Q}(\mathbf{G}^{2t}\mathbf{P})\mathbf{Q}^{\mathbf{T}} &\approx \mathbf{Q}(\mathbf{P} - 2t\mathbf{\Delta})\mathbf{Q}^{\mathbf{T}} \\ &= \mathbf{Q}\mathbf{P}\mathbf{Q}^{\mathbf{T}} - 2t\mathbf{Q}\mathbf{\Delta}\mathbf{Q}^{\mathbf{T}}.\end{aligned} \tag{4.22}$$

For bisections, the short-time approximation of $\langle h_x h_{x'} \rangle_{2t}$ can be written as

$$\begin{aligned}\langle h_x h_{x'} \rangle_{2t} &= \sum_{x,x'=1}^{n} h_{x'} p^{2t}(x', x) h_x \\ &= \mathbf{h}^{\mathbf{T}}\mathbf{G}^{2t}\mathbf{P}\mathbf{h} \\ &\approx \mathbf{h}^{\mathbf{T}}(I - 2t\mathbf{\Delta}\mathbf{P}^{-1})\mathbf{P}\mathbf{h} \\ &= \mathbf{h}^{\mathbf{T}}\mathbf{P}\mathbf{h} - 2t\mathbf{h}^{\mathbf{T}}\mathbf{\Delta}\mathbf{h} \\ &= 1 - 2t\mathbf{h}^{\mathbf{T}}\mathbf{\Delta}\mathbf{h}.\end{aligned} \tag{4.23}$$

Note that this approximation to $\langle h_x h_{x'} \rangle_{2t}$ makes no assumption about the choice of prior distribution $\mathbf{P}$ on the nodes of the graph. Furthermore, if the discrete-time diffusion operator is used instead, $\langle h_x h_{x'} \rangle_{2t}$ does not approximate to $\mathbf{h}^{\mathbf{T}}\mathbf{\Delta}\mathbf{h}$ in such a simple manner (see Appendix B).

For fast-mixing graphs, the long-time and short-time approximations for $I[Y;Z]$ and $\langle h_x h_{x'} \rangle_{2t}$, respectively, hold simultaneously

$$\begin{aligned}I[Y;Z](t) \approx \iota(t) &\approx \left(\frac{1}{2} - t\frac{\mathbf{h}^T\mathbf{\Delta}\mathbf{h}}{1 - \langle \mathbf{h} \rangle^2}\right) \\ \Rightarrow \frac{dI[Y;Z]}{dt} &\approx \frac{d\iota}{dt} \propto \begin{cases} \mathcal{A}, & p(x) \propto 1, \\ \mathcal{N}, & p(x) \propto d_x. \end{cases}\end{aligned} \tag{4.24}$$

We have shown analytically that, for fast-mixing graphs, the heuristics introduced by Shi and Malik are proportional to the rate of loss of relevance information. The error incurred in the approximations $I[Y;Z] \approx \iota$ and $\langle h_x h_{x'} \rangle_{2t} \approx 1 - 2t\mathbf{h}^{\mathbf{T}}\mathbf{\Delta}\mathbf{h}$ can be defined as

$$\mathcal{E}_0(t) = \left| \frac{\langle h_x h_{x'} \rangle_{2t} - (1 - 2t\mathbf{h}^{\mathbf{T}}\mathbf{\Delta}\mathbf{h})}{\langle h_x h_{x'} \rangle_{2t}} \right|, \tag{4.25}$$

$$\mathcal{E}_1(t) = \left| \frac{I[Y;Z](t) - \iota(t)}{I[Y;Z](t)} \right|. \tag{4.26}$$

5 Numerical Experiments

The validity of the two approximations can be seen in a typical plot of $\mathcal{E}_1(t)$ and $\mathcal{E}_0(t)$ as a function of normalized diffusion time $\tilde{t} = t/\tau$, for the two different choices of prior distributions over the nodes. $\mathcal{E}_1$, as seen in Fig. 1, is often found to be nonmonotonic and sometimes exhibits oscillations. This suggests defining $\mathcal{E}_\infty$, a modified monotonic "$\mathcal{E}_1$":

$$\mathcal{E}_\infty(t) \equiv \max_{t' \geq t} \mathcal{E}_1(t'). \tag{5.1}$$

$\mathcal{E}_\infty(t)$ is the maximum of $\mathcal{E}_1$ over all time greater than or equal to t. We do not need to define a monotonic form for $\mathcal{E}_0$ since this error is always found to be monotonically increasing in time.

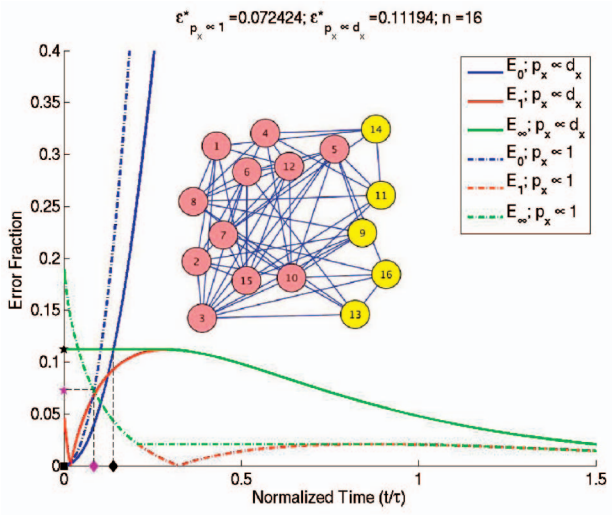

Fig. 1. $\mathcal{E}_1$ and $\mathcal{E}_0$ versus normalized diffusion time for two choices of priors over the graph nodes. $\mathcal{E}_1$ (red) typically tends to have a nonmonotonic behavior which motivates defining a monotonic $\mathcal{E}_\infty$ (green). ★—$\mathcal{E}^*$, ■—$\tilde{t}^*_-$, and ◆—$\tilde{t}^*_+$. Black—$p_x \propto d_x$ and magenta—$p_x \propto 1$.

By fast-mixing graphs, we mean graphs which become well-mixed in short times, i.e., graphs for which both the long-time and short-time approximations hold simultaneously within a certain range of time $\tilde{t}^*_- \leq \tilde{t} \leq \tilde{t}^*_+$, as illustrated in Fig. 1, where we define

$$\mathcal{E}(t) = \max(\mathcal{E}_\infty(t), \mathcal{E}_0(t)), \tag{5.2}$$

$$\mathcal{E}^* = \min_t \mathcal{E}(t), \tag{5.3}$$

$$\tilde{t}^*_- = \min\left(\arg\min_{\tilde{t}} \mathcal{E}(\tilde{t})\right), \tag{5.4}$$

$$\tilde{t}^*_+ = \max\left(\arg\min_{\tilde{t}} \mathcal{E}(\tilde{t})\right). \tag{5.5}$$

$\mathcal{E}(t)$ is the larger of the modified long-time and short-time errors, $\mathcal{E}_\infty$ and $\mathcal{E}_0$, at time t. $\mathcal{E}^*$ is the minimum of $\mathcal{E}(t)$ over all time. For some graphs, the plot of $\mathcal{E}(t)$ at its minimum might exhibit a plateau instead of a single point, as in Fig. 1 (for prior proportional to degree). $\tilde{t}^*_-$ and $\tilde{t}^*_+$ denote the left and right limits of this plateau. Note that the use of $\mathcal{E}_\infty$ instead of $\mathcal{E}_1$ overestimates the value of $\mathcal{E}^*$; the $\mathcal{E}^*$ calculated is an upper bound.

Graphs were drawn randomly from a Stochastic Block Model (SBM) distribution [14], with block cardinality 2, to analyze the distribution of $\mathcal{E}^*$, $\tilde{t}^*_-$, and $\tilde{t}^*_+$. As is commonly done in community detection [15], for a graph of n nodes, the average degree per node is fixed at $n/4$ for graphs drawn from the SBM distribution: Two nodes are connected with probability p_+ if they belong to the same block, but with probability $p_- < p_+$, if they belong to different blocks. The two probabilities are, thus, constrained by the relation

$$p_+\left(\frac{n}{2} - 1\right) + p_-\left(\frac{n}{2}\right) = \frac{n}{4} \tag{5.6}$$

leaving only one free parameter p_- that tunes the "modularity" of graphs in the distribution. Starting with a graph drawn from a distribution specified by a p_- value

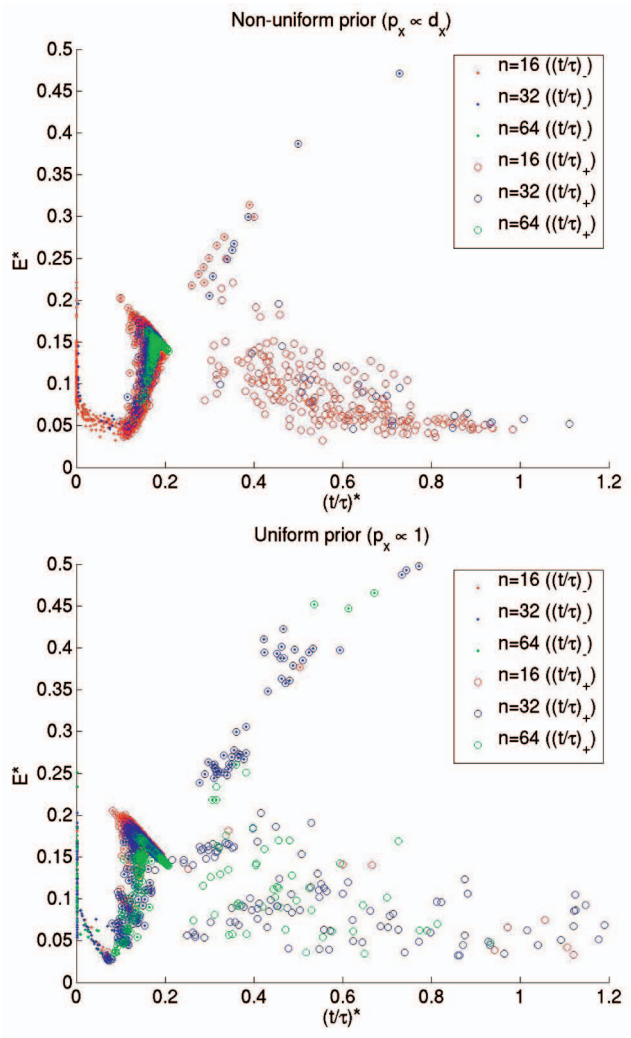

Fig. 2. $\mathcal{E}^*$ versus $\tilde{t}^*$ for graphs of different sizes and different prior distributions over the graph nodes. In the above plot, $\tilde{t}^*_-$ and $\tilde{t}^*_+$ are represented by $\cdot$ and $\circ$, respectively.

and specifying an initial cluster assignment as given by the SBM distribution, we make local moves—adding or deleting an edge in the graph and/or reassigning a node's cluster label—and search exhaustively over this move-set for local minima of $\mathcal{E}^*$. Fig. 2 compares the values of $\mathcal{E}^*$ and $\{\tilde{t}^*_-, \tilde{t}^*_+\}$ for graphs obtained in this systematic search, starting with a graph drawn from a distribution with $p_- = 0.02$ and $n = \{16, 32, 64\}$. We note that the scatter plots for graphs of different sizes collapse on one another when $\mathcal{E}^*$ is plotted against normalized time, confirming the Fiedler value $1/\tau$ to be an appropriate characteristic diffusion time scale, as used in [9]. A plot of $\mathcal{E}^*$ against actual diffusion time shows that the scatter plots of graphs of different sizes no longer collapse (see the supplemental figures, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.124).

Having shown analytically that, for fast-mixing graphs, the regularized min-cut is approximately the rate of loss of relevance information, it would be instructive to compare the actual partitions that optimize these goals. Graphs of size $n = 32$ were drawn from the SBM distribution with $p_- = \{0.1, 0.12, 0.14, 0.16\}$. Starting with an equal-sized partition specified by the model itself, we performed

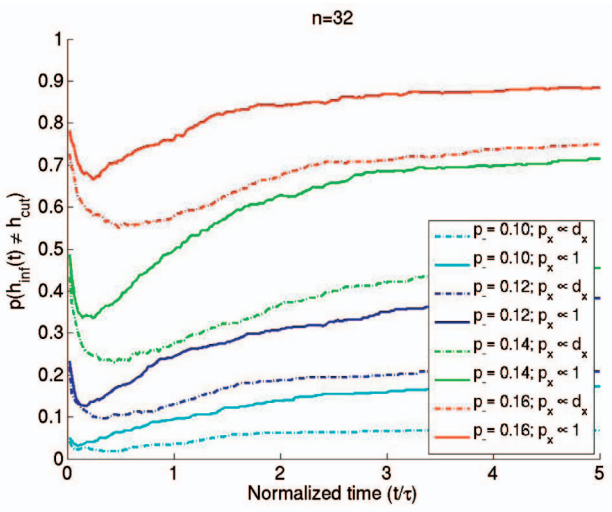

Fig. 3. $p(\mathbf{h}_{\inf}(t) \neq \mathbf{h}_{\text{cut}})$ versus normalized diffusion time, averaged over 500 graphs drawn from a distribution parameterized by a given p_- value, is plotted for different graph distributions.

iterative coordinate descent to search (independently) for the partition that minimized the regularized cut ($\mathbf{h}_{\text{cut}}$) and one that minimized the relevance information ($\mathbf{h}_{\inf}(t)$), i.e., we reassigned each node's cluster label and searched for the reassignment that gave the new lowest value for the cost function being optimized. Plots comparing the partitions $\mathbf{h}_{\inf}(t)$ and $\mathbf{h}_{\text{cut}}$, learned by optimizing the two goals (averaged over 500 graphs drawn from each distribution), are shown in Fig. 3.

6 Concluding Remarks

We have shown that the normalized cut and average cut, introduced by Shi and Malik as useful heuristics to be minimized when partitioning graphs, are well approximated by the rate of loss of predictive information for fast-mixing graphs. Deriving these cut-based cost functions from rate distortion theory gives them a more principled setting, makes them interpretable, and facilitates generalization to appropriate cut-based cost functions in new problem settings. We have also shown (see Fig. 2) that the inverse Fiedler value is an appropriate normalization for diffusion time, justifying its use in [9] to capture long-time behaviors on the network.

Absent from this manuscript is a discussion of how to not overpartition a graph, i.e., a criterion for selecting K. It is hoped that, by showing how these heuristics can be derived from a more general problem setting, lessons learned by investigating stablilty, cross-validation, or other approaches may benefit those using min-cut-based approaches as well. Furthermore, a derivation of some rigorous bounds on the magnitude of the approximation errors, under some conditions, and analysis of algorithms used in rate distortion theory and min-cut minimization are highly promising avenues for research.

Appendix A

Normalized and Average Cut

Using the definition of $\mathbf{\Delta}$, for any general vector $\mathbf{f}$ over the graph nodes, we have, for symmetric $\mathbf{A}$,

$$\begin{aligned}\mathbf{f^T}\Delta\mathbf{f} &= \mathbf{f^T}\mathbf{D}\mathbf{f} - \mathbf{f^T}\mathbf{A}\mathbf{f}\\ &= \sum_x d_x f_x^2 - \sum_{x,y=1}^{n} f_x f_y A_{xy}\\ &= \sum_x \left(\sum_{y=1}^{n} A_{xy}\right) f_x^2 - \sum_{x,y=1}^{n} f_x f_y A_{xy}\\ &= \frac{1}{2}\left(\sum_{x,y=1}^{n} f_x^2 A_{xy} - 2\sum_{x,y=1}^{n} f_x f_y A_{xy} + \sum_{x,y=1}^{n} f_y^2 A_{xy}\right)\\ &= \frac{1}{2}\sum_{x,y=1}^{n} A_{xy}(f_x - f_y)^2.\end{aligned} \tag{A.1}$$

When $\mathbf{f} = \mathbf{h}$, with $h_x \in \{-1, 1\}$, we have

$$\begin{aligned}\mathbf{h^T}\Delta\mathbf{h} &= \frac{1}{2}\sum_{h_x \times h_y = -1} 4A_{xy}\\ &= 4 \times c.\end{aligned} \tag{A.2}$$

The factor $\frac{1}{2}$ disappears because summation over all nodes counts each adjacent pair of nodes twice.

Using the definitions of $\mathcal{A}$ and $\mathcal{N}$, we have

$$\begin{aligned}\mathcal{A} &= c \times \left(\frac{1}{\sum_{h_x=+1} 1} + \frac{1}{\sum_{h_x=-1} 1}\right)\\ &= c \times \left(\frac{1}{\sum_x \left(\frac{1+h_x}{2}\right)} + \frac{1}{\sum_x \left(\frac{1-h_x}{2}\right)}\right)\\ &= 2c \times \left(\frac{\sum_x (1 - h_x + 1 + h_x)}{\sum_x (1 + h_x)\sum_x (1 - h_x)}\right)\\ &= 2c \times \left(\frac{2n}{(n + \sum_x h_x)(n - \sum_x h_x)}\right)\\ &= 2c \times \left(\frac{2}{n(1 + \bar{\mathbf{h}})(1 - \bar{\mathbf{h}})}\right)\\ &= \frac{4}{n}\frac{c}{1 - \bar{\mathbf{h}}^2},\end{aligned} \tag{A.3}$$

$$\begin{aligned}\mathcal{N} &= c \times \left(\frac{1}{\sum_{h_x=+1} d_x} + \frac{1}{\sum_{h_x=-1} d_x}\right)\\ &= c \times \left(\frac{1}{\sum_x d_x \left(\frac{1+h_x}{2}\right)} + \frac{1}{\sum_x d_x \left(\frac{1-h_x}{2}\right)}\right)\\ &= 2c \times \left(\frac{\sum_x d_x (1 - h_x + 1 + h_x)}{\sum_x (d_x(1 + h_x))\sum_x (d_x(1 - h_x))}\right)\\ &= 2c \times \left(\frac{4m}{(2m + \sum_x h_x d_x)(2m - \sum_x h_x d_x)}\right)\\ &= 2c \times \left(\frac{1}{m(1 + \langle \mathbf{h} \rangle)(1 - \langle \mathbf{h} \rangle)}\right)\\ &= \frac{2}{m}\frac{c}{1 - \langle \mathbf{h} \rangle^2}.\end{aligned} \tag{A.4}$$

APPENDIX B
DISCRETE-TIME DIFFUSION

For discrete-time diffusion, the conditional distribution $p(y \mid x)$ is given as

$$\tilde{G}^s_{yx} = p(y \mid x) = [(\mathbf{A}\mathbf{D}^{-1})^s]_{yx}, \tag{B.1}$$

where $\mathbf{D}$ is the diagonal matrix of node degrees, $\mathbf{A}$ is the adjacency matrix, and s is the number of time steps. For any s, substituting $\mathbf{A} = \mathbf{D} - \Delta$ and expanding the binomial gives

$$\begin{aligned}\tilde{\mathbf{G}}^{2s} &= (\mathbf{I} - \Delta\mathbf{D}^{-1})^{2s}\\ &= \mathbf{I} - 2s\Delta\mathbf{D}^{-1} + \sum_{j=2}^{2s}(-1)^j \binom{2s}{j}(\Delta\mathbf{D}^{-1})^j.\end{aligned} \tag{B.2}$$

Thus, for $p(x) \propto d_x$, the expression for $\langle h_x h_{x'} \rangle_{2s}$ becomes

$$\begin{aligned}\langle h_x h_{x'} \rangle_{2s} &= 1 - \frac{2s}{m}\mathbf{h^T}\Delta\mathbf{h}\\ &\quad + \sum_{j=2}^{2s}\frac{(-1)^j}{m}\binom{2s}{j}\mathbf{h^T}\Delta\left(\mathbf{D}^{-1}\Delta\right)^j\mathbf{h}.\end{aligned} \tag{B.3}$$

From the above equation, we see that, even when $s = 1$, unlike in the continuous-time diffusion case, $\langle h_x h_{x'} \rangle_{2s}$ does not approximate as simply to the cut and ι does not approximate to the normalized or average cut.

ACKNOWLEDGMENTS

The work of Chris H. Wiggins was supported by the National Institutes of Health (NIH) under grants NIH 1U54CA121852-01A1 and NIH 5PN2EY016586-03, and by the US National Science Foundation under grant IIS-0705580.

REFERENCES

[1] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, Aug. 2000.
[2] M. Meila and J. Shi, "A Random Walks View of Spectral Segmentation," *Proc. Int'l Conf. AI and Statistics*, 2001.
[3] S. Lafon and A.B. Lee, "Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning, and Data Set Parameterization," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1393-1403, Sept. 2006.
[4] B. Nadler, S. Lafon, R.R. Coifman, and I.G. Kevrekidis, "Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators," *Advances in Neural Information Processing Systems 18*, pp. 955-962, MIT Press, 2005.
[5] N. Tishby, F.C. Pereira, and W. Bialek, "The Information Bottleneck Method," *arXiv preprint physics*, http://arxiv.org/abs/physics/0004057, Jan. 2000.
[6] N. Tishby and N. Slonim, "Data Clustering by Markovian Relaxation and the Information Bottleneck Method," *Advances in Neural Information Processing Systems*, MIT Press, Jan. 2001.
[7] C. Shannon, "A Mathematical Theory of Communication," *ACM SIGMOBILE Mobile Computing and Comm. Rev.*, vol. 5, no. 1, pp. 3-55, Jan. 2001.
[8] N. Slonim, "The Information Bottleneck: Theory and Applications," PhD dissertation, The Hebrew Univ. of Jerusalem, 2002.
[9] E. Ziv, M. Middendorf, and C.H. Wiggins, "Information-Theoretic Approach to Network Modularity," *Physical Rev. E*, vol. 71, p. 046117, Jan. 2005.
[10] U. von Luxburg, "A Tutorial on Spectral Clustering," arXiv, vol. cs.DS, http://arxiv.org/abs/0711.0189v1, Nov. 2007.
[11] M. Fiedler, "Algebraic Connectivity of Graphs," *Czechoslovak Math. J.*, vol. 23, no. 98, pp. 298-305, 1973.

[12] D. Wagner and F. Wagner, "Between Min Cut and Graph Bisection," *Proc. 18th Int'l Symp. Math. Foundations of Computer Science,* pp. 744-750. 1993.
[13] F.R. Chung, *Spectral Graph Theory.* Am. Math. Soc., 1997.
[14] P.W. Holland and S. Leinhardt, "Local Structure in Social Networks," *Sociological Methodology,* pp. 1-45, Jan. 1976, http://www.cs.cmu.edu/Web/Groups/NIPX/00papers-pu-on-web/TishbySlonim.ps.gz.
[15] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing Community Structure Identification," *J. Statistical Mechanics: Theory and Experiment,* Jan. 2005.

Anil Raj received the bachelor of technology degree in aerospace engineering from the Indian Institute of Technology Madras, Chennai, in 2005, the MS degree in applied physics and the MPhil degree in applied mathematics from Columbia University, New York, in 2006 and 2009, respectively. He is currently working toward the PhD degree under the supervision of Dr. Chris Wiggins. His research interests include information theory, graph theory, statistical learning theory, and machine learning and their application to problems in pattern recognition, data mining, and computational biology.

Chris H. Wiggins received the PhD degree in physics from Princeton University in 1998. Between 1998 and 2001, he was a US National Science Foundation Mathematical Sciences Foundation postdoctoral research fellow at the Courant Institute at New York University. Since 2001, he has been a member of the Department of Applied Physics and Applied Mathematics of Columbia University, New York, where he is currently an associate professor. He is also affiliated with the Center for Computational Biology and Bioinformatics and the NanoMedicine Center for Mechanical Biology at Columbia University. His research interests include applications of machine learning, statistical inference, and information theory for the inference, analysis, and organization of biological networks.